

# Approximate state-space Gaussian processes via spectral transformation

Toni Karvonen and Simo Särkkä

This article has been published in *Proceedings of 26th IEEE International Workshop on Machine Learning for Signal Processing*, held in Salerno, Italy, on September 13–16, 2016. When citing this article, use the following reference:

T. Karvonen and S. Särkkä, “Approximate state-space Gaussian processes via spectral transformation,” in *26th IEEE International Workshop on Machine Learning for Signal Processing*, 2016.

DOI: 10.1109/MLSP.2016.7738812

**Copyright.** ©2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

# APPROXIMATE STATE-SPACE GAUSSIAN PROCESSES VIA SPECTRAL TRANSFORMATION

Toni Karvonen\*

Aalto University  
Department of Electrical Engineering  
and Automation  
Espoo, Finland

Simo Särkkä†

Aalto University  
Department of Electrical Engineering  
and Automation  
Espoo, Finland

## ABSTRACT

State-space representations of Gaussian process regression use Kalman filtering and smoothing theory to downscale the computational complexity of the regression in the number of data points from cubic to linear. As their exact implementation requires the covariance function to possess rational spectral density, rational approximations to the spectral density must be often used. In this article we introduce new spectral transformation based methods for this purpose: a spectral composition method and a spectral preconditioning method. We study convergence of the approximations theoretically and run numerical experiments to attest their accuracy for different densities, in particular the fractional Matérn.

**Index Terms**— Gaussian process regression, state-space approximation, fractional Matérn, composite approximation, spectral preconditioning

## 1. INTRODUCTION

In machine learning, the problem of estimating the value  $f(t)$  of an unknown function  $f$  from a finite number of possibly noisy function value observations is ubiquitous. Gaussian process (GP) regression [1] approaches the problem by postulating a non-parametric Gaussian process prior over the function  $f$ . What this means is that, for every  $t$ ,  $f(t)$  is taken as a zero-mean (arbitrary mean is easily accommodated if necessary) Gaussian random variable such that each finite collection  $f(t_1), \dots, f(t_n)$  of random variables is jointly Gaussian. A zero-mean Gaussian process is completely specified by its covariance function  $k(t, t') = \mathbb{E}[f(t)f(t')]$ , and the whole model, with noisy function evaluations  $\mathbf{y} = (y_1, \dots, y_N)^\top$  (the data) at  $t_1, \dots, t_N$ , can be written as

$$\begin{aligned} f &\sim \text{GP}(0, k(t, t')), \\ y_i &= f(t_i) + \varepsilon_i, \end{aligned} \quad (1)$$

\*Thanks to Aalto ELEC Doctoral School for funding.

†Thanks to Academy of Finland for funding.

where  $\varepsilon_i$  are independent zero-mean Gaussian random variables with variances  $\sigma_i^2$ . The posterior  $p(f(t) | \mathbf{y})$  is then Gaussian with explicitly computable mean  $\mu(t)$  and variance  $V(t)$ :

$$\begin{aligned} \mu(t) &= \mathbf{k}^\top(t)(\mathbf{K} + \mathbf{\Sigma})^{-1}\mathbf{y}, \\ V(t) &= k(t, t) - \mathbf{k}^\top(t)(\mathbf{K} + \mathbf{\Sigma})^{-1}\mathbf{k}(t), \end{aligned} \quad (2)$$

where  $\mathbf{k}_i(t) = k(t, t_i)$ ,  $[\mathbf{K}]_{ij} = k(t_i, t_j)$ , and  $\mathbf{\Sigma}$  is a diagonal matrix containing the noise variances  $\sigma_i^2$ .

Now, as can be seen from (2), a naive implementation of GP regression with  $N$  data points requires inverting an  $N \times N$  matrix  $\mathbf{K} + \mathbf{\Sigma}$  and therefore suffers from cubic time-complexity  $\mathcal{O}(N^3)$ , which is usually computationally prohibitive for other than small sets of data. To counter this problem, state-space representations of Gaussian processes were introduced in [2–4]. In state-space methodology one constructs a continuous-time linear state-space model of the form

$$\begin{aligned} \frac{d\mathbf{x}(t)}{dt} &= \mathbf{A}\mathbf{x}(t) + \mathbf{L}w(t), \\ y_i &= \mathbf{H}\mathbf{x}(t_i) + \varepsilon_i, \end{aligned} \quad (3)$$

with  $\mathbf{A}$ ,  $\mathbf{L}$  and  $\mathbf{H}$  suitable matrices,  $w(t)$  a white noise process, and  $\varepsilon_i$  as in (1). The building blocks of this model are chosen such that the model becomes equivalent to the Gaussian process regression problem at hand—typically the function  $f(t)$  is augmented as a component of the state  $\mathbf{x}(t)$ . Because the posterior state distribution for (3) can be computed with linear time-complexity  $\mathcal{O}(N)$  using classical Kalman filtering and smoothing theory [5], the problem is transformed into computationally feasible form.

Of course, the problem is then, given a Gaussian process regression problem (1), to actually transform it to the state-space form. As demonstrated in [4], this is possible exactly if and only if the covariance function  $k$  has a spectral density  $S(\omega)$  that is a rational function of  $\omega^2$  with order of the numerator smaller than that of the denominator. Consider, for instance,

the Matérn class of covariance functions

$$k_M(t, t') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}|t-t'|}{\ell} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu}|t-t'|}{\ell} \right),$$

where  $\sigma$ ,  $\nu$ , and  $\ell$  are certain positive parameters and  $K_\nu$  is the modified Bessel function of the second kind. If  $\nu = 1/2 + n$  for some  $n \in \mathbb{N}$ , this covariance function has a spectral density of the desired form [1, Chapter 4]:

$$S_M(\omega) = \sigma^2 \frac{2\sqrt{\pi}\Gamma(\nu + 1/2)}{\Gamma(\nu)} \lambda^{2\nu} (\lambda^2 + \omega^2)^{-(\nu+1/2)},$$

with  $\lambda = \sqrt{2\nu}/\ell$ . However, for all other values of  $\nu$  the spectral density is not a rational function. Unfortunately, perhaps the most frequently used covariance function, the squared exponential (SE)

$$k_{SE}(t, t') = \sigma^2 \exp \left( -\frac{(t-t')^2}{2\ell^2} \right),$$

has a spectral density not of the desired rational form.

Different rational approximations can be used to these non-rational spectral densities [2, 4, 6]. Särkkä and Piché [7] provided theoretical framework for using such approximations and, in the case of the SE, numerically experimented with their accuracy and convergence.

There are three criteria for a successful approximation: (i) the approximation must be valid spectral density, amounting to positivity; (ii) it must be rational function of  $\omega^2$ ; and (iii) it should converge to the true density as well as corresponding true posterior mean and variance functions (2). In contrast to the results in [7], straightforward use of Taylor and Padé approximants does not usually produce approximants with these properties for covariance functions other than the SE. Some new methods are therefore needed.

In this article we provide several novel rational approximations to spectral densities using composite Taylor and Padé approximations and spectral preconditioning. These approximations are based on spectral transformations. We experimentally investigate their performance for covariance functions other than the SE. Particular attention is given to Matérn covariance functions with non-rational spectral densities.

## 2. MAIN RESULTS

This section presents different methods of creating rational approximations to spectral densities using spectral compositions and preconditioning. Numerical experiments and practical comparison of the algorithms will be carried out later in Section 3.

### 2.1. Composition of spectral functions

Suppose we have a spectral density  $S$  that can be expressed as a composition of a function  $f$  and a continuous ‘‘basis’’ spectral

density  $B$ :

$$S(\omega) = B(f(\omega)). \quad (4)$$

Given that  $B$  and  $f$  admit convergent approximations  $B_n$  and  $f_m$ , we can use composition of these approximations to produce a sequence  $S_{n,m}$  of rational approximations to  $S$  as

$$S_{n,m}(\omega) = B_n(f_m(\omega)). \quad (5)$$

By Wiener–Khinchin theorem [1, Chapter 4], this leads to a sequence of approximations  $k_{n,m}$  to the corresponding covariance function  $k$ . In the following we show that, under some non-restrictive assumptions, the approach of using approximations (5) is theoretically sound.

Let us therefore consider sequences of positive functions  $B_n \rightarrow B$  and  $f_m \rightarrow f$  such that the composite functions  $B_n(f_m(\omega))$  are bounded by an integrable function and convergence of  $B_n$  to  $B$  is uniform. These standing assumptions are encapsulated in the following.

**Assumption 2.1.** *The sequence  $f_m$  converges to  $f$  point-wise and the sequence of positive functions  $B_n$  converges to continuous  $B$  uniformly. This is to say that for every  $\varepsilon > 0$  there is  $N > 0$  such that  $|B_n(\omega) - B(\omega)| < \varepsilon$  for all  $n > N$  and  $\omega \in \mathbb{R}$ .*

**Assumption 2.2.** *There exists a function  $\bar{S}$  such that  $\int_{\mathbb{R}} \bar{S}(\omega) d\omega < \infty$  and  $B_n(f_m(\omega)) \leq \bar{S}(\omega)$  for all  $n, m$  and  $\omega$ .*

We now consider the composite function  $B_n(f_m(\omega))$  and prove that it converges to  $B(f(\omega)) = S(\omega)$  when  $n, m \rightarrow \infty$ . Note that we are working with the double limit  $\lim_{n,m \rightarrow \infty} B_n(f_m(\omega))$ , defined, if existing, as the number  $a$  for which there is for every  $\varepsilon > 0$  an integer  $N$  such that  $|B_n(f_m(\omega)) - a| < \varepsilon$  for all  $n, m > N$ . This is distinct from the iterated limits (see [8, Section 1.5])

$$\lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} B_n(f_m(\omega)) \quad \text{and} \quad \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} B_n(f_m(\omega)).$$

**Theorem 2.3.** *The sequence of composite approximations  $B_n(f_m(\omega))$  converges to  $B(f(\omega))$  point-wise as  $m, n \rightarrow \infty$ .*

*Proof.* By the triangle inequality,

$$\begin{aligned} & |B_n(f_m(\omega)) - B(f(\omega))| \\ & \leq |B_n(f_m(\omega)) - B(f_m(\omega))| + |B(f_m(\omega)) - B(f(\omega))|, \end{aligned}$$

and the point-wise convergence follows from Assumption 2.1. Specifically, the first term on the right-hand side vanishes by the uniform convergence assumption of  $B_n$  to  $B$  and the second by continuity of  $B$  and the point-wise convergence of  $f_m$  to  $f$ .  $\square$

Combined with Theorems 2.1 and 2.2 of [7], Assumption 2.2 and Theorem 2.3 then yield the following result on the convergence of corresponding covariance function  $k_{n,m}$  as well as posterior mean and variance (2).

**Theorem 2.4.** *The sequence of covariance function approximations  $k_{n,m}$  is uniformly bounded and converges uniformly to  $k$ , the covariance function corresponding to the spectral density  $S$ . Furthermore, if the noise covariance  $\Sigma$  is positive-definite, then the posterior mean and variance functions  $\mu_{n,m}$  and  $V_{n,m}$ , corresponding to  $k_{n,m}$ , converge uniformly to  $\mu$  and  $V$ , those of  $k$ .*

## 2.2. Composite approximation methods

We use two different basis spectral densities that yield positive densities in the powers of  $\omega^2$ :

$$\begin{aligned} B_{\text{SE}}(\omega) &= \exp(-\omega^2), \\ B_{\text{M}(1/2)}(\omega) &= \frac{1}{1 + \omega^2}, \end{aligned} \quad (6)$$

the second one being basically the Matérn density with  $\nu = 1/2$ . The problem is then to construct a sequence of approximations to  $f$  and ensure that this chosen sequence actually converges point-wise. Both our basis densities have one-sided inverses  $B_+^{-1}$ , so the function  $f$  can be recovered as

$$f(\omega) = B_+^{-1}(S(\omega)).$$

This means that

$$\begin{aligned} f_{\text{SE}}(\omega) &= B_{\text{SE}}^{-1}(S(\omega)) = \sqrt{-\ln S(\omega)}, \\ f_{\text{M}(1/2)}(\omega) &= B_{\text{M}(1/2)}^{-1}(S(\omega)) = \sqrt{1/S(\omega) - 1}, \end{aligned} \quad (7)$$

can then be used to form the truncated Taylor series approximations  $f_{[M]}$  at the origin:

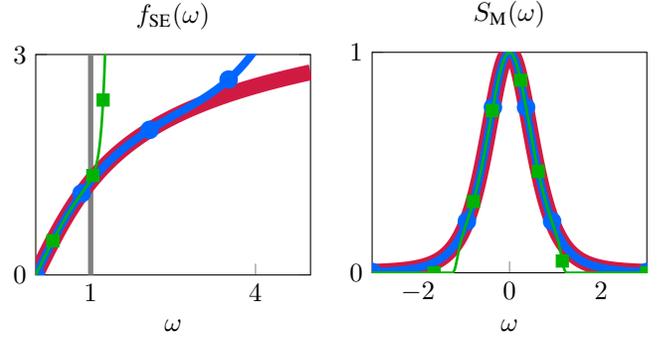
$$f_{[M]}(\omega) = \sum_{n=0}^M \frac{f^{(n)}(0)\omega^n}{n!}. \quad (8)$$

Convergence of these Taylor series in the set of analyticity of  $S$  is guaranteed by the following theorem.

**Theorem 2.5.** *Consider the spectral decomposition (4) of  $S$  with basis densities (6). If  $S$  is analytic in a set  $\Omega = \{|\omega| \leq \gamma\}$  and  $0 < S(\omega) \leq 1$  for all  $\omega \in \Omega$ , then the Taylor series of functions (7) at the origin converge in  $\Omega$ .*

*Proof.* Because  $S$  is analytic and positive in the set  $\Omega$ , so are  $\ln S(\omega)$  and  $1/S(\omega)$ . Furthermore, as  $S(\omega) \leq 1$  in  $\Omega$ , we have that  $-\ln S(\omega) \geq 0$  and  $1/S(\omega) \geq 1$  in  $\Omega$ , implying that  $f_{\text{SE}}$  and  $f_{\text{M}(1/2)}$  are analytic in this set. Therefore their Taylor series converge in  $\Omega$ .  $\square$

The requirement that  $S(\omega) \leq 1$  is not a restriction because the approximation to  $f$  can be formed on the basis of normalized density  $S(\omega)/\max_{\omega \in \mathbb{R}} S(\omega) \leq 1$  and after this multiplied multiplied by  $\max_{\omega \in \mathbb{R}} S(\omega)$ .



**Fig. 1:** Example of composite approximations to normalized Matérn spectral density (■)  $S_{\text{M}}(\omega) = 1/(1 + \omega^2)^\alpha$  with  $\alpha = 7/3$  and  $\ell = \sqrt{2\nu}$ . Here  $B_{\text{SE}}$  is used and  $f_{\text{SE}}$  is approximated with Taylor series of order 17 (■—■) and Padé approximant  $[12/5]$  (●—●). Convergence radius of Taylor series is marked on the figure for  $f_{\text{SE}}$ .

Unfortunately, the convergence radius of a Taylor series of  $S$  is often quite small. For example, a Taylor series for Matérn density with  $\ell = \sqrt{2\nu}$ ,

$$S_{\text{M}}(\omega) \propto (1 + \omega^2)^{-(\nu+1/2)}, \quad (9)$$

converges only in  $\{|\omega| \leq 1\}$ .

Often a more fruitful approach is to match the  $L + M$  first derivatives of the composition in (4) at the origin and form the Padé approximant [9]

$$f_{[L/M]}(x) = \frac{b_0 + b_1x + \dots + b_Lx^L}{1 + a_1x + \dots + a_Mx^M}$$

out of the derivatives similarly to what was done for the SE covariance function in [7]. This approximant usually has a larger radius of convergence than the Taylor one. Fig. 1 illustrates the difference in the convergence radii of Taylor and Padé approximants for (9) with  $\nu = 7/3 - 1/2$ .

As  $B_{\text{M}(1/2)}$  is already in the rational form, it does not require any approximations. The other basis density  $B_{\text{SE}}$  can be approximated by the methods in [7] that convergence uniformly, satisfying hence the assumptions in Section 2.1.

## 2.3. Direct approximation of covariance functions

Sometimes, due to the construction of the model, the spectral density  $S$  is not directly available. This makes it difficult to form the Taylor series or Padé expansions which need the derivatives of the spectral density at origin. Fortunately, by computing moments of the covariance function, the derivatives can be evaluated through Fourier transform:

$$\left. \frac{d^n S(\omega)}{d\omega^n} \right|_{\omega=0} = (-i)^n \int_{\mathbb{R}} \tau^n k(\tau) d\tau,$$

where we have assumed that the covariance function is stationary:  $k(t, t') \triangleq k(t - t')$  (with a slight abuse of notation).

In practice, we can evaluate the moments via numerical integration and then form the Taylor and Padé approximants using those numerical values. This allows us to form rational approximations of the spectral density without the need to explicitly evaluate the spectral density or its derivatives.

## 2.4. Spectral preconditioning

The accuracy of the Taylor series expansion of the Matérn spectral density is limited by the exponent  $\alpha = \nu + 1/2$  in

$$S_M(\omega) \propto (\lambda^2 + \omega^2)^{-\alpha},$$

because it defines how many terms we can include to the expansion while still retaining a valid spectral density [10]. To cope with this problem we introduce spectral preconditioning, where we write the spectral density  $S$  as

$$S(\omega) = S_{p,n}(\omega)S(\omega)/S_{p,n}(\omega) \quad (10)$$

with  $S_{p,n}$  some function, and approximate this by functions  $S_n^{\text{pre}}$  in some way.

For fractional Matérn spectral density this procedure works particularly well when  $S_{p,n}$  is selected as

$$S_{p,n}(\omega) = (\lambda^2 + \omega^2)^n$$

and  $S_n^{\text{pre}}$  is constructed as

$$S_n^{\text{pre}}(\omega) = S_{p,n}(\omega)/P_n(\omega) = (\lambda^2 + \omega^2)^n/P_n(\omega), \quad (11)$$

where

$$S_{p,n}(\omega)S(\omega) = (\lambda^2 + \omega^2)^{\alpha+n} \approx P_n(\omega) = \sum_{k=0}^{\lceil \alpha \rceil + n} a_{k,n} \omega^{2k}.$$

Here

$$a_{k,n} = \frac{1}{k!} \lambda^{2(\alpha+n)-2k} \prod_{m=0}^{k-1} (\alpha + n - m) \quad (12)$$

are Taylor series coefficients of  $(\lambda^2 + \omega^2)^{\alpha+n}$  and  $P_n$  is therefore Taylor series truncated just before the first negative coefficient. In practice, although not indicated by the following theorem, this approximation seems to be a significant improvement over the conventional Taylor series approximation in terms of the radius of convergence.

**Theorem 2.6.** *The preconditioned approximation  $S_n^{\text{pre}}(\omega)$  in (11) converges to  $S(\omega)$  exponentially for every  $|\omega| \leq |\lambda|$ .*

*Proof.* Write  $(\lambda^2 + \omega^2)^{\alpha+n} = P_n(\omega) + R_n(\omega)$ . The quotient (10) can then be written as

$$\frac{(\lambda^2 + \omega^2)^n}{P_n(\omega)} = \frac{1}{(\lambda^2 + \omega^2)^\alpha \left(1 - \frac{R_n(\omega^2)}{(\lambda^2 + \omega^2)^{\alpha+n}}\right)}. \quad (13)$$

We need to show that  $\lim_{n \rightarrow \infty} R_n(\omega)/(\lambda^2 + \omega^2)^{\alpha+n} = 0$ . Denote  $n_\alpha = \lceil \alpha \rceil + n$  and  $(\alpha)_k = \alpha(\alpha-1) \times \dots \times (\alpha-k)$ . Now, by Taylor's theorem,  $R_n(\omega)$  has the Lagrange form

$$\begin{aligned} R_n(\omega) &= \frac{(\alpha+n)_{n_\alpha}}{(n_\alpha+1)!} (\lambda^2 + c^2)^{\alpha+n-n_\alpha-1} \omega^{2(n_\alpha+1)} \\ &= \frac{(\alpha+n)_{n_\alpha}}{(n_\alpha+1)!} (\lambda^2 + c^2)^{\alpha+n} \left(\frac{\omega^2}{\lambda^2 + c^2}\right)^{n_\alpha+1} \end{aligned}$$

for some  $0 < c < |\omega| \leq |\lambda|$ . Hence

$$\begin{aligned} R_n(\omega)/(\lambda^2 + \omega^2)^{\alpha+n} &= \frac{(\alpha+n)_{n_\alpha}}{(n_\alpha+1)!} \left(\frac{\lambda^2 + c^2}{\lambda^2 + \omega^2}\right)^{\alpha+n} \left(\frac{\omega^2}{\lambda^2 + c^2}\right)^{n_\alpha+1} \\ &= \frac{(\alpha+n)_{n_\alpha}}{(n_\alpha+1)!} \left(\frac{\lambda^2 + c^2}{\lambda^2 + \omega^2}\right)^{\alpha+n} \left(\frac{\omega^2}{\lambda^2 + c^2}\right)^{n_\alpha+1}. \end{aligned}$$

The second and third term vanish exponentially as  $n \rightarrow \infty$  so it needs to be shown that the first term remains bounded. Now,

$$\begin{aligned} \left| \frac{(\alpha+n)_{n_\alpha}}{(n_\alpha+1)!} \right| &= \frac{(\alpha+n)(\alpha+n-1) \times \dots \times |\alpha - \lceil \alpha \rceil|}{(\lceil \alpha \rceil + n)!} \\ &\leq \frac{(\lceil \alpha \rceil + n)!}{(\lceil \alpha \rceil + n)!} \\ &= 1. \end{aligned}$$

□

## 2.5. Other possible methods

Experiments of this article are limited to composite approximations with the basis densities (6) and spectral preconditioning, but there is a great number of different approximation methods to be tried. For example, instead of inner transforms (4), outer transforms of the form  $S(\omega) = h(B(\omega))$  or combined transforms  $S(\omega) = h(B(f(\omega)))$  could be used, with convergence properties like those established in Section 2.1.

One could also use an orthogonal polynomial series [11, Chapter 3] to do the approximation. Based on our limited experiments, not presented in this article, with approximating Matérn density with Laguerre polynomials, this works well for some values of  $\nu$  while failing for others.

## 3. EXPERIMENTAL RESULTS

In this section we test the accuracy of approximations introduced in the preceding section for covariance functions lacking rational spectral densities. The covariance functions investigated are fractional Matérn and a product of SE and fractional Matérn. The experiments are analogous to those in [7]: we use the same data (see Fig. 2 of that article) and measure maximum absolute error of the true and approximated posterior mean and covariance functions.

The experiments were carried out by using Kalman filtering on the state-space form and the approximation order, as a function of which the maximum errors are plotted, is the denominator order in  $\omega^2$ .

### 3.1. Fractional Matérn covariance

We test the approximations to Matérn covariance function with parameters  $\sigma = \ell = 1$  and  $\nu = 7/3 - 1/2$ . The relevant spectral density is therefore  $S_M(\omega) \propto (\sqrt{11/3} + \omega^2)^{-7/3}$ . We increase the approximation order  $n$  in  $\omega^2$  and cover the following approximations (see Fig. 2):

- Composite approximation with basis density  $B_{M(1/2)}$  using Padé approximants  $[2 + n/1 + n]$  to  $f_{M(1/2)}$ . Due to heavy computational cost of computing high-order derivatives, the approximation is only up to order 20.
- ◆ Spectral preconditioning of Eq. (11).
- ▲ Composite approximation with basis density  $B_{SE}$  that is approximated with Padé approximants  $[2n/4n]$  from [7, Theorem 4.1] and  $f_{SE}$  that is approximated with Padé  $[2/1]$  approximant.
- ✱ Composite approximation with basis density  $S_{SE}$  that is approximated with truncated Taylor series of order  $n$  and  $f_{SE}$  that is approximated with Padé  $[3/2]$  approximant.

For comparison, we also display the two simple approximations of rounding the Matérn exponent below and above:  $-\lceil \nu + 1/2 \rceil$  (■) and  $-\lfloor \nu + 1/2 \rfloor$  (◆).

The selection of Padé approximants for  $f$  is a delicate matter. Most importantly, these cannot have denominator order  $M$  exceeding that of numerator  $L$  as this would result to  $f_{L/M}(\omega)$  vanishing when  $\omega \rightarrow \infty$ , a consequence of which would be that  $B(f_{L/M}(\omega)) \rightarrow 1$  as  $\omega \rightarrow \infty$ . Too large a difference in  $L$  and  $M$  easily results to too fast convergence of  $B(f_{L/M}(\omega))$ . We have therefore used the smallest possible difference of  $M$  and  $L$  (note that the Padé approximant orders are given in terms of order of  $\omega^2$ ).

Because Taylor series approximations to  $f$  have limited convergence radius as illustrated in Fig. 1, they have been omitted from these experiments as incapable of yielding exact enough approximations.

### 3.2. Product covariance

Our another test case is the product covariance function

$$k(t, t') = k_{SE}(t, t')k_M(t, t') \quad (14)$$

of SE and fractional Matérn with  $\sigma = \ell = 1$  (for the both covariance functions) and  $\nu = 7/3 - 1/2$ . This lacks closed-form spectral density so the needed derivatives have to be calculated with the moment approach of Section 2.3. The following composite approximations were used (see Fig. 3):

- ✱ Basis density  $B_{M(1/2)}$  and Taylor expansions of increasing order of  $f_{M(1/2)}$ .
- ◆ Basis density  $B_{M(1/2)}$  and Padé approximants  $[n + 1/n - 1]$  of  $f_{M(1/2)}$ .

### Fractional Matérn covariance

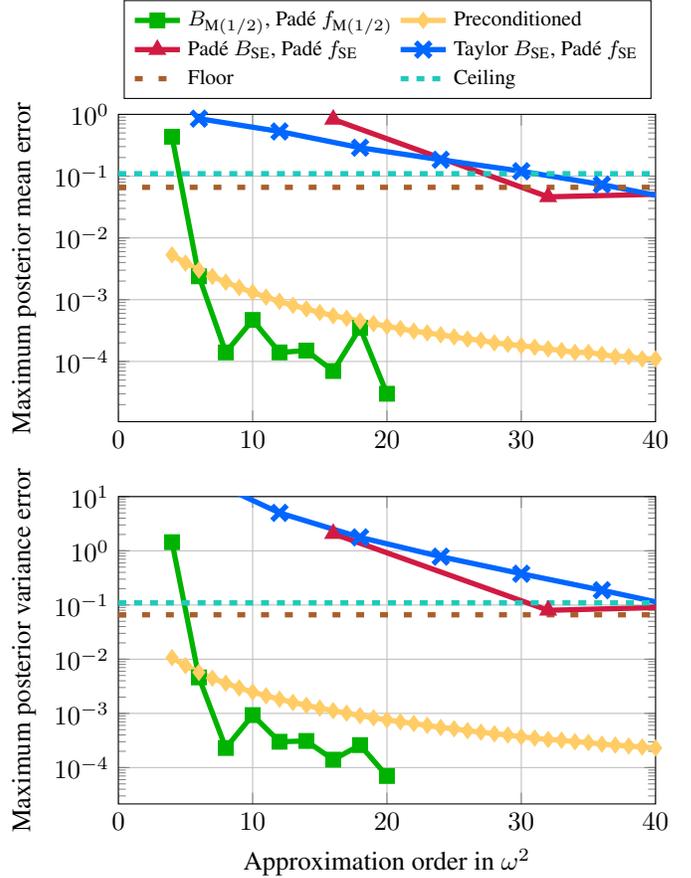


Fig. 2: Absolute maximum error in posterior mean and variance approximations to fractional Matérn covariance.

### 3.3. Comparison of accuracy

Evidently composite methods based on the basis density  $B_{SE}$  do not work well. They involve approximations to two different functions and neither of these approximations can be made very accurate without the overall order of the composite approximation becoming large, slowing down the implementation and causing numerical errors. Padé approximations, combined with the basis density  $B_{M(1/2)}$ , work well, although their estimates do not completely consistently improve when approximation order is increased. The preconditioned approximation for fractional Matérn works well and is consistent but requires somewhat higher approximation order than  $B_{M(1/2)}$  with Padé. However, forming it is simple as the Taylor series coefficients (12) are easily calculated. No approximant has an exponential convergence rate (for preconditioned approximant exponential convergence is only guaranteed inside the radius of convergence of the Taylor series).

The Taylor series based approximation to the product covariance does not work nearly as well as the Padé one and has

less consistent behaviour. It may be that this Taylor expansion does not converge everywhere. In this case, high-order approximants are probably hindered by reliable computation of high-order moments being difficult and more demanding.

### 3.4. On time-complexity

Naive GP regression is of time-complexity  $\mathcal{O}(N^3)$  whereas converting the problem to state-space form reduces this to  $\mathcal{O}(N)$ . Taking into account also  $M$ , the approximation order in  $\omega^2$ , corresponding to state-space dimension, the complete time-complexity of our methods is  $\mathcal{O}(M^3N)$ . Comparison of computational time needed by state-space methodology and naive GP regression has been carried out in, for example, [2]. Comparisons for the methods presented will appear in a future article where they will also be compared to other state-of-the-art fast GP algorithms such as [12, 13].

## 4. CONCLUSION AND DISCUSSION

We developed novel rational approximations to spectral densities based on composite Taylor and Padé approximations and spectral preconditioning and experimentally investigated their performance in approximating Matérn covariance functions with non-rational spectral densities. The methods seem to lead to promising covariance function approximations.

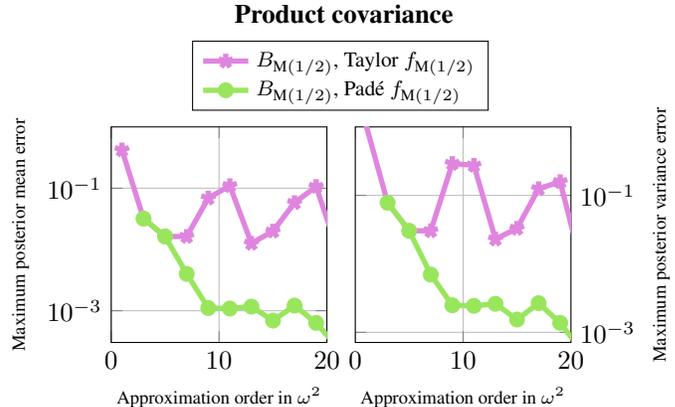
Further research is possible by using approximations briefly outlined in Section 2.5, particularly towards the use of orthogonal polynomial series expansions. Methodology of this article can also be applied to other basis densities besides the two considered and different preconditioners. The proof of Theorem 2.3 indicates that basis densities with small Lipschitz constants should have faster convergence rates. Also, statement of Theorem 2.6 could probably be improved.

## 5. ACKNOWLEDGEMENT

We thank Robert Piché for discussion regarding other potential rational approximations as well as reading and commenting the manuscript. We also thank the reviewers for some comments.

## 6. REFERENCES

- [1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
- [2] J. Hartikainen and S. Särkkä, “Kalman filtering and smoothing solutions to temporal Gaussian process regression models,” in *IEEE International Workshop on Machine Learning for Signal Processing*, 2010.
- [3] S. Särkkä and J. Hartikainen, “Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression,” in *JMLR Workshop and Conference Proceedings Volume 22*, 2012, pp. 993–1001.



**Fig. 3:** Absolute maximum error in posterior mean and variance approximations to product covariance (14).

- [4] S. Särkkä, A. Solin, and J. Hartikainen, “Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing,” *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 51–61, 2013.
- [5] S. Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.
- [6] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Gordon and Breach, 1963.
- [7] S. Särkkä and R. Piché, “On convergence and accuracy of state-space approximations of squared exponential covariance functions,” in *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.
- [8] P. Walker, *Examples and Theorems in Analysis*, Springer, 2004.
- [9] G. A. Baker and P. Graves-Morris, *Padé Approximants, Part I: Basic Theory*, Addison-Wesley, 1981.
- [10] L. Roininen, S. Lasanen, M. Orispää, and S. Särkkä, “Sparse approximations of fractional Matérn fields,” 2014, Preprint, arXiv:1502.06800.
- [11] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1989.
- [12] J. Quiñero-Candela and C. E. Rasmussen, “A unifying view of sparse approximate Gaussian process regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.
- [13] F. Lindgren, H. Rue, and J. Lindström, “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 4, pp. 423–498, 2011.