2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

ON CONVERGENCE AND ACCURACY OF STATE-SPACE APPROXIMATIONS OF SQUARED EXPONENTIAL COVARIANCE FUNCTIONS

*Simo Särkkä**

Aalto University
02150 Espoo, Finland

Robert Piché

Tampere University of Technology
33101 Tampere, Finland

ABSTRACT

In this paper we study the accuracy and convergence of state-space approximations of Gaussian processes (GPs) with squared exponential (SE) covariance functions. This kind of approximations is important in construction of Kalman filtering and smoothing based GP regression algorithms, which have a linear (as opposed to conventional cubic) computational complexity in the number of training samples. We start by deriving general conditions for a spectral density approximation to give a uniform convergence of the mean and covariance functions. We then show that the previously proposed reciprocal Taylor series approximation gives such uniform convergence. We then derive new approximations based on Padé approximants of the exponential function as well as approximations inspired by the central limit theorem, and prove their uniform convergence. Finally, we compare accuracy of the different approximations numerically.

***Index Terms*—** Gaussian process regression, state-space approximation, squared exponential, Kalman filter and smoother, Padé approximant, central limit theorem

1. INTRODUCTION

Gaussian process (GP) regression (e.g. [1]) is concerned with estimating the value of an unknown function $f(t)$ at a given input value t (i.e. test point) based on a finite number of noisy training samples observed from it. The difference to classical regression is that instead of postulating a parametric regression function $f_\theta(t;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^D$ are parameters to be fitted, in GP regression we use a Gaussian process prior with a given covariance function $k(t,t')$ to model the unknown functions $f(t)$. Here we concentrate on the case where the input is univariate $t \in \mathbb{R}$. The univariate case is important in the signal processing context where we are modeling temporal signals and hence the input variable can be considered to be time.

A GP regression problem [1] with the noisy measurements $y_j \in \mathbb{R}$, $j = 1,\ldots,N$ of the unknown function $f(t)$ at times t_j can be written as:

$$\begin{aligned} f(t) &\sim \mathrm{GP}(0, k(t,t')), \\ y_j &= f(t_j) + \epsilon_j, \end{aligned} \tag{1}$$

where the errors ϵ_j are independent zero mean Gaussian with variances σ_j^2. This model as well as the results of this article can be easily extended in various ways, for example, to non-zero mean, or to vector or correlated measurements.

The a posteriori process, that is, the process conditioned on the given set of measurements $\mathbf{y} = \begin{pmatrix} y_1 & \ldots & y_N \end{pmatrix}^T$, is also Gaussian and has the following mean and variance functions [1]:

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

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_N^2)$, $\mathbf{K} = k(t_{1:N}, t_{1:N})$ is an $N \times N$ matrix and $\mathbf{k}(t) = k(t, t_{1:N})$ is a column vector. In principle, formulas (2) give the full solution to the GP regression problem (with fixed hyperparameters), but, because of the $O(N^3)$ complexity of the matrix inversions, the formulas can only be applied directly to small data sets.

State-space representations of Gaussian processes [2–4] have recently been proposed as a solution to the above-mentioned computational scaling problem, other approaches being, for example, sparse, reduced rank, and fast Fourier transform (FFT) based approximations (see, e.g., [5–10] and references therein). The state-space methods are based on constructing a state-space model (see [2] for details)

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{A}\,\mathbf{x}(t) + \mathbf{L}\,w(t), \\ y_j &= \mathbf{H}\,\mathbf{x}(t_j) + \epsilon_j, \end{aligned} \tag{3}$$

where $\mathbf{A}$, $\mathbf{L}$, and $\mathbf{H}$ are given matrices and $w(t)$ is a white noise process with spectral density q, such that the state-inference problem in the above model is equivalent to a GP regression problem (1). It turns out [2–4] that the state-inference problem can be solved efficiently using classical Kalman filtering and smoothing [11–14], which has linear complexity $O(N)$ with respect to the number of measurements. The state-space GP methods have also been extended to non-Gaussian and non-linear settings and they have found many applications in location sensing, physics, and medicine [4, 15–18].

The main challenge in the state-space approach is the construction of the state-space model from a given covariance function prescription. As discussed in, for example, [2–4], such a state-space representation is possible exactly if and only if the spectral density of the Gaussian process is a rational function. Unfortunately, this is not the case for the most commonly used square exponential covariance function

$$k(t,t') = s^2\, \exp\left(-\frac{(t-t')^2}{2\ell^2}\right). \tag{4}$$

The spectral density is

$$\begin{aligned} S(\omega) &= \int k(t, t+\tau)\, \exp(-i\,\omega\,\tau)\, d\tau \\ &= s^2\, \sqrt{2\pi}\,\ell\, \exp\left(-\frac{\ell^2\,\omega^2}{2}\right), \end{aligned} \tag{5}$$

*Thanks to Academy of Finland (projects 266940, 273475) for funding.

978-1-4799-3694-6/14/$31.00 ©2014 IEEE

which is clearly not a rational function.

However, a simple and accurate approximation, as already pointed out in [19] and later used in [2–4], is the following rational approximation based on the nth order truncated Taylor series expansion of the exponential function in the denominator:

$$S_n(\omega) = \frac{s^2\,(2\pi)^{1/2}\,\ell}{1 + \left(\frac{\ell^2\,\omega^2}{2}\right) + \frac{1}{2!}\left(\frac{\ell^2\,\omega^2}{2}\right)^2 + \cdots + \frac{1}{n!}\left(\frac{\ell^2\,\omega^2}{2}\right)^n}. \tag{6}$$

This approximation leads to a $n-1$ times differentiable process that approximates the infinitely differentiable squared exponential process. For any finite n this approximate system can be converted into an equivalent n-dimensional state-space model with white noise input [2].

It is desirable that the above approximation should converge to the exact spectral density as $n \to \infty$. We also want the corresponding covariance function approximation, which by the Wiener-Khinchin theorem (e.g., [1, 2]) is given as

$$k_n(t,t') = \frac{1}{2\pi}\int S_n(\omega)\,\exp(i\,\omega\,(t-t'))\,\mathrm{d}\omega, \tag{7}$$

to converge to the exact covariance. The aim of this article is to prove that these convergences indeed happen and to introduce some alternative approximations that also converge to the exact covariance. We also show that these convergences imply convergence of the posterior mean and covariance. Additionally, we numerically evaluate how fast the errors diminish as the approximation order increases.

2. UNIFORM CONVERGENCE OF COVARIANCE FUNCTIONS AND GP REGRESSORS

In this section we derive two general theorems that will be used in the subsequent sections to show the convergence of different approximations. The first theorem is concerned with the convergence of the a priori covariance function and the second one with the convergence of the Gaussian process posterior mean and covariance functions.

Theorem 2.1 (Convergence of a priori covariance function)**.** *Consider a stationary covariance function $k(t,t')$ having spectral density $S(\omega)$ and a sequence of covariance functions $k_n(t,t')$ having spectral densities $S_n(\omega)$ for $n \in \mathbb{Z}_+$. If*

1. *$S_n(\omega)$ converges point-wise to $S(\omega)$ with $n \to \infty$; and*
2. *There exists a function $\bar{S}(\omega) \geq 0$ with $\int \bar{S}(\omega)\,\mathrm{d}w < \infty$ such that $S_n(\omega) \leq \bar{S}(\omega)$ for all $n \in \mathbb{Z}_+$ and $\omega \in \mathbb{R}$,*

then the covariance function approximation $k_n(t,t')$ converges uniformly to $k(t,t')$. Furthermore, the covariance functions are uniformly bounded.

Proof. The Lebesgue dominated convergence theorem (e.g. [20]) implies that given the conditions 1 and 2 above, together with the positivity of the spectral densities, we have

$$\lim_{n\to\infty}\int |S_n(\omega) - S(\omega)|\,\mathrm{d}\omega \to 0. \tag{8}$$

Using (7) and $|\exp(i\,\omega\,(t-t'))| \leq 1$ together with elementary inequalities gives

$$|k_n(t,t') - k(t,t')| \leq \frac{1}{2\pi}\int |S_n(\omega) - S(\omega)|\,\mathrm{d}\omega. \tag{9}$$

The convergence results by taking the limit $n \to \infty$. The convergence is uniform, because the bound on the right hand side is independent of t, t'. The uniform boundedness of the covariance functions k_n follows from

$$|k_n(t,t')| \leq \frac{1}{2\pi}\int S_n(\omega)\,\mathrm{d}\omega \leq \frac{1}{2\pi}\int \bar{S}(\omega)\,\mathrm{d}\omega < \infty. \tag{10}$$

The uniform boundedness of k results from the integrability of S implied by the dominated convergence theorem and a similar argument as above. □

Theorem 2.2 (Convergence of posterior mean and variance)**.** *If conditions 1 and 2 in Theorem 2.1 hold, and the joint noise covariance $\mathbf{\Sigma}$ is strictly positive definite, then the corresponding posterior mean and variance $\mu_n(t)$ and $V_n(t)$ converge uniformly to $\mu(t)$ and $V(t)$, respectively.*

Proof. For the mean we get by simple manipulation and using the triangle inequality:

$$\begin{aligned}
&|\mu_n(t) - \mu(t)| \\
&= |\mathbf{k}_n^\mathsf{T}(t)\ (\mathbf{K}_n + \mathbf{\Sigma})^{-1}\ \mathbf{y} - \mathbf{k}^\mathsf{T}(t)\ (\mathbf{K} + \mathbf{\Sigma})^{-1}\ \mathbf{y}| \\
&\leq |(\mathbf{k}_n(t) - \mathbf{k}(t))^\mathsf{T}\ (\mathbf{K}_n + \mathbf{\Sigma})^{-1}\ \mathbf{y}| \\
&\qquad + |\mathbf{k}^\mathsf{T}(t)\ (\mathbf{K}_n + \mathbf{\Sigma})^{-1}\ (\mathbf{K} - \mathbf{K}_n)\ (\mathbf{K} + \mathbf{\Sigma})^{-1}\ \mathbf{y}|.
\end{aligned} \tag{11}$$

Because $\mathbf{k}_n(t)$ converges uniformly to $\mathbf{k}(t)$ and $\mathbf{K}_n$ converges to $\mathbf{K}$, and the remaining vectors and matrices are uniformly bounded, μ_n also converges uniformly to μ.

For the variance we similarly get

$$\begin{aligned}
&|V_n(t) - V(t)| \\
&\leq |k_n(t,t) - k(t,t)| + |(\mathbf{k}(t) - \mathbf{k}_n(t))^\mathsf{T}\ (\mathbf{K}_n + \mathbf{\Sigma})^{-1}\ \mathbf{k}_n(t)| \\
&\qquad + |\mathbf{k}^\mathsf{T}(t)\ (\mathbf{K} + \mathbf{\Sigma})^{-1}\ (\mathbf{K} - \mathbf{K}_n)\ (\mathbf{K}_n + \mathbf{\Sigma})^{-1}\ \mathbf{k}_n(t)| \\
&\qquad + |\mathbf{k}^\mathsf{T}(t)\ (\mathbf{K} + \mathbf{\Sigma})^{-1}\ (\mathbf{k}(t) - \mathbf{k}_n(t))|,
\end{aligned} \tag{12}$$

where the result follows with the same arguments as for the mean. □

3. TAYLOR SERIES EXPANSIONS

In this section we show the convergence of the approximation (6), which we call the Taylor series approximation because it is obtained by approximating the exponential in $1/S(\omega) \propto \exp(\ell^2\,\omega^2/2)$ using the Taylor series expansion

$$\exp(x) = 1 + x + \frac{1}{2!}\,x^2 + \frac{1}{3!}\,x^3 + \cdots \tag{13}$$

truncated after the Mth term.

Theorem 3.1 (Convergence of the Taylor series approximation)**.** *Consider the sequence of approximations*

$$S_{[M]}(\omega) = \frac{s^2\,(2\pi)^{1/2}\,\ell}{1 + \left(\frac{\ell^2\,\omega^2}{2}\right) + \frac{1}{2!}\left(\frac{\ell^2\,\omega^2}{2}\right)^2 + \cdots + \frac{1}{M!}\left(\frac{\ell^2\,\omega^2}{2}\right)^M}. \tag{14}$$

When $M \to \infty$, the a priori covariance function converges uniformly to the squared exponential (4)*. Also the posterior mean and covariance functions converge uniformly to the GP regression solution with the squared exponential* (4) *covariance function.*

Proof. By Taylor's theorem, the sequence $S_{[M]}(\omega)$ converges point-wise to the spectral density (5). Furthermore, each sequence member is clearly dominated by the integrable function

$$\bar{S}(\omega) = \frac{s^2\,(2\pi)^{1/2}\,\ell}{1 + \left(\frac{\ell^2\,\omega^2}{2}\right)}, \tag{15}$$

and hence by Theorem 2.1 the prior covariance function converges uniformly and by Theorem 2.2 the posterior mean and covariance converge uniformly. □

In fact, the sequence of approximations in (14) converges uniformly, not only point-wise, to $S(\omega)$, which results from Theorem 2.3 and Equation (4.4) in [21]. That theorem also implies that the error should vanish exponentially in the order M, more precisely, $\sup_\omega |S_{[M]}(\omega) - S(\omega)| \leq S(0)\,2^{-M}$.

4. PADÉ APPROXIMANTS

The Padé approximant [22,23] of the exponential function is a rational function, denoted

$$\exp_{[L/M]}(x) = \frac{B(x)}{A(x)} = \frac{b_0 + b_1\,x + \cdots + b_L\,x^L}{1 + a_1\,x + \cdots + a_M\,x^M}. \tag{16}$$

The coefficients are defined by the condition that the Taylor series of (16) agrees with (13) to as high a degree as possible. Perron's formula for the coefficients in (16) is

$$a_j = (-1)^j\,\frac{(L+M-j)!\,M!}{(L+M)!\,j!\,(M-j)!}, \quad j = 1, \ldots, M, \tag{17a}$$

$$b_j = \frac{(L+M-j)!\,L!}{(L+M)!\,j!\,(L-j)!}, \quad j = 0, \ldots, L. \tag{17b}$$

The Padé approximants of the spectral density (5) are

$$S_{[L/M]}(\omega) = s^2\,(2\pi)^{1/2}\,\ell\,\exp_{[L/M]}\left(-\frac{\ell^2\,\omega^2}{2}\right). \tag{18}$$

In particular, the Taylor series approximation $S_{[M]}$ in (14) is the Padé approximant $S_{[0/M]}$. Other Padé approximants are

$$S_{[1/2]}(\omega) = \frac{(2\pi)^{1/2}\,\ell\,s^2\,\left(1 - \frac{\ell^2\,\omega^2}{6}\right)}{\frac{\ell^4\,\omega^4}{24} + \frac{\ell^2\,\omega^2}{3} + 1}, \tag{19a}$$

$$S_{[1/3]}(\omega) = \frac{(2\pi)^{1/2}\,\ell\,s^2\,\left(1 - \frac{\ell^2\,\omega^2}{8}\right)}{\frac{\ell^6\,\omega^6}{192} + \frac{\ell^4\,\omega^4}{16} + \frac{3\,\ell^2\,\omega^2}{8} + 1}, \tag{19b}$$

$$S_{[2/3]}(\omega) = \frac{(2\pi)^{1/2}\,\ell\,s^2\,\left(\frac{\ell^4\,\omega^4}{80} - \frac{\ell^2\,\omega^2}{5} + 1\right)}{\frac{\ell^6\,\omega^6}{480} + \frac{3\,\ell^4\,\omega^4}{80} + \frac{3\,\ell^2\,\omega^2}{10} + 1}, \tag{19c}$$

$$S_{[1/4]}(\omega) = \frac{(2\pi)^{1/2}\,\ell\,s^2\,\left(1 - \frac{\ell^2\,\omega^2}{10}\right)}{\frac{\ell^8\,\omega^8}{1920} + \frac{\ell^6\,\omega^6}{120} + \frac{3\,\ell^4\,\omega^4}{40} + \frac{2\,\ell^2\,\omega^2}{5} + 1}, \tag{19d}$$

$$S_{[2/4]}(\omega) = \frac{(2\pi)^{1/2}\,\ell\,s^2\,\left(\frac{\ell^4\,\omega^4}{120} - \frac{\ell^2\,\omega^2}{6} + 1\right)}{\frac{\ell^8\,\omega^8}{5760} + \frac{\ell^6\,\omega^6}{240} + \frac{\ell^4\,\omega^4}{20} + \frac{\ell^2\,\omega^2}{3} + 1}. \tag{19e}$$

An L/M Padé approximant corresponds to a $M - L - 1$ times differentiable Gaussian process. Hence the processes $S_{[1/2]}$ and $S_{[2/3]}$ are not differentiable, the processes $S_{[1/3]}$ and $S_{[2/4]}$ are once differentiable, and $S_{[1/4]}$ is three times differentiable.

However, not all Padé approximants are valid (i.e. non-negative) spectral densities. For example, $S_{[1/M]}(\omega) < 0$ for $\omega^2\ell^2 > 2(1 + M)$ and so no $S_{[1/M]}$ is valid. On the other hand, it follows from the formula at the end of [22, §66] that all $S_{[2n/M]}$ are valid spectral densities. Among the approximants in (19), $S_{[2/4]}$ is the best in the sense that it is valid, it is differentiable, and it is fairly low order. For this reason, we will use this particular Padé approximant as the basis for other approximations in the next section.

The advantageous properties of the $S_{[2/4]}$ approximant motivate us to define a sequence of approximants as follows:

$$S_n(\omega) = S_{[2n/4n]}(\omega). \tag{20}$$

The convergence of this sequence of approximations is established by the following theorem.

Theorem 4.1 (Convergence of the $S_{[2n/4n]}$ approximation)**.** *For the sequence of spectral density approximations $S_n(\omega)$ of* (20)*, the prior covariance function approximation as well as the corresponding posterior mean and variance approximations converge uniformly to the squared exponential* (4) *counterparts.*

Proof. Theorem 2.3 in [21] implies that $\sup_\omega |S_{[2n/4n]}(\omega) - S(\omega)| \leq S(0)\,2^{-2n}$ and hence the approximation $S_{[2n/4n]}$ converges point-wise. The formula at the end of [22, §66] implies that these approximants are always positive and hence valid spectral densities. Finally, lemma A.1 (Appendix A) implies that $S_{[2n/4n]}(\omega) \leq \bar{S}(\omega)$, where $\bar{S}$ is defined in (15). □

5. CENTRAL LIMIT THEOREM APPROXIMATIONS

In the familiar exponential function formula

$$\exp(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \tag{21}$$

the term $1 + \frac{x}{n}$ can be recognized as a first order approximation to $\exp(x/n)$. Hence this formula is closely related to the identity

$$\exp(x) = \exp(x/n)^n \tag{22}$$

which corresponds to the "scaling and squaring" approximation that is used in numerical computation of matrix exponentials [24].

In fact, the identity (21) can be generalised to

$$\exp(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n} + o\left(\frac{1}{n}\right)\right)^n, \tag{23}$$

where the term $o(1/n)$ can be an arbitrary expression that goes to zero faster than $1/n$, with any dependence on x (recall that we are considering point-wise convergence). Thus, we can put, for example, any Taylor series expansion of the exponential (13) of order greater than one inside the parenthesis and still have the convergence to the exponential function — and the convergence rate could be expected to be higher with a higher order Taylor polynomial.

The identity (23) is also closely related to the central limit theorem (CLT, see, e.g., [25]), because if we replace x with $-\omega^2/2$, we get

$$\exp(-\omega^2/2) = \lim_{n\to\infty}\left(1 - \frac{\omega^2}{2n} + o\left(\frac{1}{n}\right)\right)^n. \tag{24}$$

If we now interpret the term in the parenthesis as a Taylor series expansion of the characteristic function of a random variable $X_i/\sqrt{n}$, it says that the sum of n such independent random variables $\sum_{i=1}^n X_i$ (whose characteristic function is the product the individual characteric functions) converges to a standard Gaussian. By generalizing this idea we get the following theorem.

Theorem 5.1 (Convergence of CLT approximation I)**.** *Consider a non-negative integrable base approximation to the squared exponential covariance function of the form*

$$\hat{S}(\omega) = s^2 \, (2\pi)^{1/2} \, \ell \times \left(\frac{1 + b_1 \left(\frac{\ell^2 \, \omega^2}{2}\right) + \cdots + b_L \left(\frac{\ell^2 \, \omega^2}{2}\right)^L}{1 + a_1 \left(\frac{\ell^2 \, \omega^2}{2}\right) + \cdots + a_M \left(\frac{\ell^2 \, \omega^2}{2}\right)^M} \right). \tag{25}$$

with $L < M$. If for $n = 2, 3, \ldots$ we form a sequence of new approximations via

$$S_n(\omega) = \hat{S}(0) \left(\frac{\hat{S}(n^{-1/2} \, \omega)}{\hat{S}(0)} \right)^n, \tag{26}$$

then the resulting approximation converges uniformly to the squared exponential covariance function (4) *in the limit $n \to \infty$, provided that $a_1 - b_1 = 1$ and if there exists an integrable function $\bar{S}(\omega)$ such that $S_n(\omega) \le \bar{S}(\omega)$ for all n. The corresponding posterior mean and variance also converge uniformly.*

Proof. We have

$$S_n(\omega) = s^2 \, (2\pi)^{1/2} \, \ell \times \left(\frac{1 + b_1 \left(\frac{\ell^2 \, \omega^2}{2n}\right) + \cdots + b_L \left(\frac{\ell^2 \, \omega^2}{2n}\right)^L}{1 + a_1 \left(\frac{\ell^2 \, \omega^2}{2n}\right) + \cdots + a_M \left(\frac{\ell^2 \, \omega^2}{2n}\right)^M} \right)^n \tag{27}$$

which converges point-wise to

$$\lim_{n \to \infty} S_n(\omega) = s^2 \, (2\pi)^{1/2} \, \ell \, \exp\left((b_1 - a_1) \left(\frac{\ell^2 \, \omega^2}{2} \right) \right). \tag{28}$$

The limit above equals to the squared exponential if $a_1 - b_1 = 1$. Provided that the mentioned integrable function exists, then the result follows from Theorems 2.1 and 2.2. □

Corollary 5.1 (Convergence of CLT approximation II)**.** *Assume that we form the sequence $S_n(\omega)$ as in Theorem 5.1 and that the base approximation* (25) *is bounded by $\bar{S}(\omega)$ defined in* (15)*. Then the corresponding mean and covariance functions converge uniformly to the squared exponential ones.*

Proof. We get

$$\begin{aligned} & s^2 \, (2\pi)^{1/2} \, \ell \left(\frac{1 + b_1 \left(\frac{\ell^2 \, \omega^2}{2n}\right) + \cdots + b_L \left(\frac{\ell^2 \, \omega^2}{2n}\right)^L}{1 + a_1 \left(\frac{\ell^2 \, \omega^2}{2n}\right) + \cdots + a_M \left(\frac{\ell^2 \, \omega^2}{2n}\right)^M} \right)^n \\ & \le \frac{s^2 \, (2\pi)^{1/2} \, \ell}{\left(1 + \left(\frac{\ell^2 \, \omega^2}{2n}\right)\right)^n} \qquad \text{[then use the binomial theorem]} \\ & = \frac{s^2 \, (2\pi)^{1/2} \, \ell}{\left(1 + n \left(\frac{\ell^2 \, \omega^2}{2n}\right) + \text{(positive terms)}\right)} \le \frac{s^2 \, (2\pi)^{1/2} \, \ell}{\left(1 + \left(\frac{\ell^2 \, \omega^2}{2}\right)\right)}. \end{aligned} \tag{29}$$

The result now follows from Theorem 5.1 and the integrability of the last bound. □

Example 5.1. *For the first order Taylor series base approximation we have*

$$S_n(\omega) = \frac{s^2 \, (2\pi)^{1/2} \, \ell}{\left(1 + \left(\frac{\ell^2 \, \omega^2}{2n}\right)\right)^n}. \tag{30}$$

The convergence now follows from Corollary 5.1 and hence the means and covariances converge uniformly to the squared exponential. This result is just the well-known result of convergence of the Matérn class of covariance functions to the squared exponential [1].

Example 5.2. *By using the* $[2/4]$ *Padé approximant* (19e) *as the base approximation we get*

$$S_n(\omega) = (2\pi)^{1/2} \, \ell \, s^2 \left(\frac{\frac{\ell^4 \, \omega^4}{120 n^2} - \frac{\ell^2 \, \omega^2}{6n} + 1}{\frac{\ell^8 \, \omega^8}{5760 n^4} + \frac{\ell^6 \, \omega^6}{240 n^3} + \frac{\ell^4 \, \omega^4}{20 n^2} + \frac{\ell^2 \, \omega^2}{3n} + 1} \right)^n. \tag{31}$$

$S_{[2/4]}$ *is bounded by* (15) *and thus the convergence follows from Corollary 5.1.*

Example 5.3. *It is also possible to construct approximations using other than Taylor series or Padé approximants as the base approximation. For example,*

$$\hat{S}(\omega) = \frac{s^2 \, (2\pi)^{1/2} \, \ell}{1 + \frac{\ell^2 \, \omega^2}{2} + \frac{\beta^2}{2!} \left(\frac{\ell^2 \, \omega^2}{2}\right)^2} \tag{32}$$

with $\beta = (\pi - 1)/\sqrt{2}$ satisfies the conditions of Theorem 5.1 and matches the process variance, that is, $\hat{k}(t, t) = \frac{1}{2\pi} \int \hat{S}(\omega) \, \mathrm{d}\omega = s^2$. When $s = \ell = 1$, the corresponding covariance function differs from $k(t, t + \tau)$ by at most 0.05*, whereas the maximum difference for second-order Taylor is* 0.141.

6. NUMERICAL EVALUATION OF ACCURACY

In this section we evaluate the accuracy of the proposed approximations numerically. Because the computational requirements of the Kalman filter and smoother are proportional to size of the state-vector, it is natural to measure the accuracy as function of state dimensionality. Furthermore, as we are mainly interested in the uniform errors in the covariance function, we use the maximum error over all the input values as the accuracy measure.

The following approximations were tested (with scaling parameter values $s = \ell = 1$):

- Taylor n: nth order Taylor series from Theorem 3.1.
- Taylor $[1]^{\mathrm{n}}$: CLT approximation with a first order Taylor series base approximation from Example 5.1.
- Taylor $[2]^{\mathrm{n}}$: Same as the above, but with a second order Taylor series base approximation.
- Pade $[2/4]^{\mathrm{n}}$: CLT approximation with a Padé $2/4$ base approximation from Example 5.2.
- Pade [2n/4n]: Padé $2n/4n$ approximation from Example 4.1.
- $\mathrm{MTaylor}^{\mathrm{n}}$: CLT approximation with a modified second order Taylor series base approximation from Example 5.3.

The maximum absolute errors in the prior covariance functions resulting from the approximations are shown in Figure 1. The results are given as function of the state dimensionality, because it determines the computational requirements of the method. The state

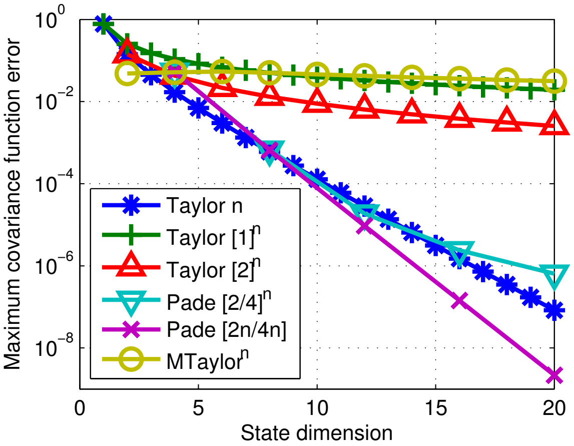

Fig. 1. Maximum errors in a priori covariance function as function of the state dimensionality.

dimensionality is determined by the denominator order and hence with different methods we get different state-dimensions with the same n. For example, with Taylor series, the state dimension is exactly n, but with Pade $[2n/4n]$ it is $4n$. In the figure, it can be seen that the Taylor series approximations and the Padé $2n/4n$ approximants have an exponentially decaying error as expected. The Taylor series approximation works surprisingly well and has a lower error than the Padé $2n/4n$ approximant until the state dimension 8, after which the Padé approximant is better.

In contrast, the errors of the CLT-type approximations do not have an exponential convergence rate although the Padé $2/4$ base approximation leads to quite low errors with the tested state-dimensionalities and with certain state-dimensionalities it has a lower error than the Taylor series approximation above. The modified Taylor series base approximation leads to the lowest error among them with small state dimensions, but its error increases in the beginning and then diminishes very slowly with the state dimension. The Taylor series base approximations do indeed converge faster, but the convergence rates are still quite low.

We also tested the effect of the approximations to the posterior mean and variance in a simple Gaussian process regression example shown in Figure 2. The resulting errors are shown in Figure 3. The results are similar to the a priori covariance function results with a few differences: with the low state dimesionalities, the modified Taylor series base approximation error first diminishes and then grows a bit before starting a slow descent. Furthermore, both the Padé $2n/4n$ approximants and the Padé $2/4$ base approximants have a clearly lower error than the Taylor series approximation already at the state dimension 8 although the latter error later rises above the Taylor series error due to its sub-exponential convergence rate.

7. CONCLUSION AND DISCUSSION

In this paper we have studied the convergence and accuracy of Taylor series, Padé, and central limit theorem (CLT) based approximations to the squared exponential function in the context of Gaussian process regression. We have proved the theoretical convergence of the previously proposed Taylor series approximations as well as of a number of new approximations. The numerical evaluation of the accuracy shows that the Taylor series expansion is a very good approximation, but there are some other approximations which come close or outperform the Taylor series approximations in approximation error. In particular, Padé approximants can be more accurate in this sense.

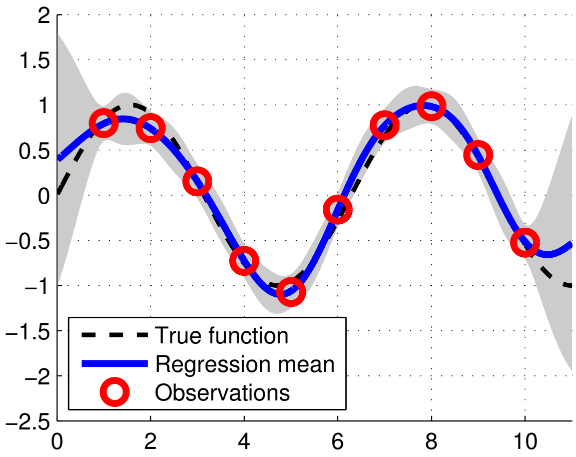

Fig. 2. Example of GP regression used for evaluating the posterior approximations. Dashed line is the true function and solid line is the posterior mean. Shaded area depicts the 95% (point-wise) confidence region.

However, the approximation error is not the only criterion that is important in the selection of the appropriate approximation. The CLT approximations have an advantage that because they are simple powers of a base approximation, forming the state-space model of an arbitrary order is easy provided that we can form it for the base approximation. This is advantageous when a symbolic spectral factorization is desired, which is the case, for example, in forming state space representations of spatio-temporal processes [3, 4].

A. UPPER BOUND FOR PADÉ 2N/4N

The following result is used in the proof of Theorem 4.1.

Lemma A.1. $\exp_{[0/1]}(-x) \geq \exp_{[2n/4n]}(-x)$ *for* $x \geq 0$.

Proof. For $n = 1$, we have

$$\begin{aligned}\exp_{[0/1]}&(-x) - \exp_{[2/4]}(-x)\\ &= \frac{\frac{1}{360}x^4 + \frac{1}{2}x^2}{(1+x)(\frac{1}{360}x^4 + \frac{1}{30}x^3 + \frac{1}{5}x^2 + \frac{2}{5}x + 1)} \geq 0.\end{aligned}$$

The inductive step

$$\exp_{[2n/4n]}(-x) - \exp_{[2(n+1)/4(n+1)]}(-x) \geq 0$$

follows from the inequalities

$$\begin{aligned}B_{[2n/4n]}&(-x)A_{[2n+2/4n+3]}(-x)\\ &- B_{[2n+2/4n+3]}(-x)A_{[2n/4n]}(-x) \geq 0, \qquad (33)\\ B_{[2n+2/4n+3]}&(-x)A_{[2n+2/4n+4]}(-x)\\ &- B_{[2n+2/4n+4]}(-x)A_{[2n+2/4n+3]}(-x)\\ &= \frac{(4n+3)!}{(6n+5)!}\frac{(2n+2)!}{(6n+6)!}x^{6n+6} \geq 0. \qquad (34)\end{aligned}$$

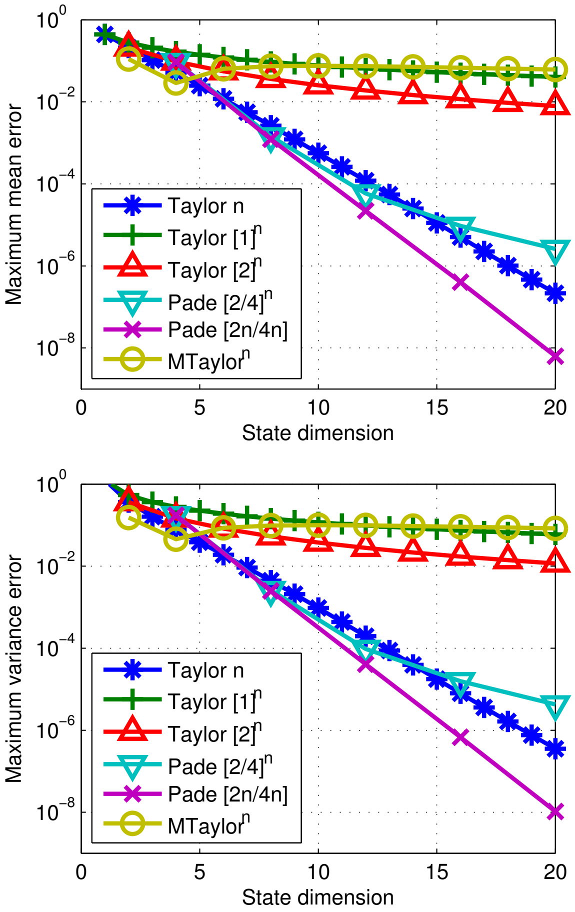

Fig. 3. Top: Maximum errors in posterior mean as function of the state dimensionality. Bottom: Maximum errors in the posterior variance.

The inequality (33) holds because the left hand side is a polynomial of degree $6n + 3$ whose three leading coefficients are positive and whose remaining coefficients can be shown, using arguments similar to [23, p. 94–95], to be zero. □

B. 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 *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 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 (AISTATS 2012)*, 2012, pp. 993–1001.

[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] J. Quiñonero-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.

[6] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, 2006, vol. 18, pp. 1259–1266.

[7] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[8] A. Solin and S. Särkkä, "Hilbert space methods for reduced-rank Gaussian process regression," 2014, arXiv:1401.5508.

[9] C. J. Paciorek, "Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package," *Journal of Statistical Software*, vol. 19, no. 2, pp. 1–38, 2007.

[10] J. Fritz, I. Neuweiler, and W. Nowak, "Application of FFT-based algorithms for large-scale universal kriging problems," *Mathematical Geosciences*, vol. 41, no. 5, pp. 509–533, 2009.

[11] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[12] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[13] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.

[14] S. Särkkä, *Bayesian filtering and smoothing*, Cambridge University Press, 2013.

[15] J. Hartikainen and S. Särkkä, "Sequential inference for latent force models," in *Proceedings of The 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

[16] J. Hartikainen, M. Seppänen, and S. Särkkä, "State-space inference for non-linear latent force models with application to satellite orbit prediction," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[17] J. Hartikainen, J. Riihimäki, and S. Särkkä, "Sparse spatio-temporal Gaussian processes with general likelihoods," in *Proceedings of ICANN'11*, 2011.

[18] A. Solin and S. Särkkä, "Infinite-dimensional Bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data," *Phys. Rev. E*, vol. 88, pp. 052909, Nov 2013.

[19] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Gordon and Breach, 1963.

[20] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, 3rd edition, 1987.

[21] E. B. Saff and R. S. Varga, "Convergence of Padé approximants to e^{-x} on unbounded sets," *Journal of Approximation Theory*, vol. 13, no. 4, April 1975.

[22] H. Padé, "Sur la représentation approchée d'une fonction par des fractions rationnelles," in *Annales scientifiques de l'École Normale Supérieure, Sér. 3*, vol. 9, pp. 3–93 (supplément). 1892, PhD Thesis.

[23] G. A. Baker and P. Graves-Morris, *Padé Approximants Part I: Basic Theory*, Addison-Wesley, 1981.

[24] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, third edition, 1996.

[25] D. W. Stroock, *Probability Theory: An Analytic View*, Cambridge University Press, 2nd edition, 2011.