Abstract

We present the class of Hida-Matérn kernels, which is the canonical family of covariance functions over the entire space of stationary Gauss-Markov Processes. It extends upon Matérn kernels, by allowing for flexible construction of priors over processes with oscillatory components. Any stationary kernel, including the widely used squared-exponential and spectral mixture kernels, are either directly within this class or are appropriate asymptotic limits, demonstrating the generality of this class. Taking advantage of its Markovian nature we show how to represent such processes as state space models using only the kernel and its derivatives. In turn this allows us to perform Gaussian Process inference more efficiently and side step the usual computational burdens. We also show how exploiting special properties of the state space representation enables improved numerical stability in addition to further reductions of computational complexity.

1 Introduction

The Gaussian process (GP) framework provides principled means to make inferences on functions [1]. Endowed with a calibrated measure of uncertainty, GPs fit well within the Bayesian paradigm and can be embedded into a broad class of models [2, 3, 4]. However, in practice the computational burden of using GPs typically limits their applicability to moderately sized datasets. Scalable frameworks such as inducing point methods, specially structured kernels, streaming approaches, state-space formulation of GPs and so on seek to remedy this weakness by reducing the computational complexity of GP inferences [5, 6, 7, 8]. While earlier literature surrounding GPs in the 1950s and 60s was largely focused on their theoretical properties, many of these findings have practical implications for developing more scalable inference frameworks.

Building on pioneering works by Takeyuki Hida and others, we introduce the class of Hida-Matérn kernels, which form a basis for translation invariant kernels and readily admit a state-space representation through which exact inference can be made in linear time.

When reasoning about a GP, inspection of its covariance function, or kernel, immediately reveals properties such as stationarity, periodicity, and differentiability. However, beyond the second order statistical structure of the process, the kernel alone fails to rigorously quantify aspects such as Markovianity, sample path properties, and uniqueness of representation. Building on P. Levy’s constructive formulation of GPs, or as it was called, a canonical representation [9, 10], T. Hida was able to broadly generalize characteristics of GPs [11, 12].

While in a practical sense many of these theoretical properties may be of little consequence, fully understanding the Markov properties of GPs can help alleviate the computational burden commonly

*Supported by NSF IIS-1845836 and NIH R01-EB026946

Preprint. Under review.
associated with them. For example, the GP with Matérn $\frac{1}{2}$ kernel (a.k.a. Ornstein-Uhlenbeck process) is well known to admit fast inference schemes thanks to its Markov property [13]. As we will see in Section 4, there exist other ways of defining a generalized GP Markov property, which in turn, provide a clear avenue to formulating them in terms of a state-space model (SSM). The SSM representation can then be used so that exact GP inference can be had in linear time.

More than this though, we will show that the class of GPs with admissible state-space representations turns out to be very broad. In fact, all stationary, real-valued, and finitely differentiable GPs, which we will refer to as Hida-Matérn GPs (H-M GPs), have canonical representations that must be linear combinations of basis functions initially derived by T. Hida in [11]. The derived covariance functions, or Hida-Matérn kernels, corresponding to these basis thus span the space of all such kernels governing H-M GPs. Moreover, due to the universality of Hida-Matérn kernels, translation-invariant covariance functions that govern GPs not in this class can be approximated arbitrarily well by linear combinations of Hida-Matérn kernels (Thm. 3.1).

Although state-space formulations of GPs have been examined extensively in recent literature, their formulation involves parametrizing a stochastic differential equation (SDE) whose stationary covariance matches that of the GP in question [14, 15, 16, 17]. In contrast, our approach only requires determining all derivatives of the covariance function. Furthermore, owing to Markov properties that will be discussed, the SSM formulation of any H-M GP is trivial to construct (Sec. 4.3).

To facilitate thinking beyond the second order structure of GPs, we re-introduce the importance of defining GP Markov properties through simple, yet enlightening examples (Sec. 2). We then introduce the family of Hida-Matérn kernels, present their universality with respect to $L^2$ convergence in the space of translation-invariant kernels, how their Markovianity leads to simple SSM representations, and then how said representations open the door for computationally feasible GP inference (Sec. 3). We show that linear SDEs with stable dynamics must admit a solution that lies within the Hida-Matérn family, which consequently implies that the matrix exponential of such dynamics matrices has a closed form solution (Sec. 4). Finally, we show examples of approximating arbitrary kernels through linear combinations of Hida-Matérn kernels (Sec. 7.1), demonstrate how low-order Hida-Matérn kernels extrapolate well on the Mauna Loa CO$_2$ data set, and illustrate the scalability of our approach in speed comparisons against state of the art methods (Sec. 8).

2 Background: Canonical representation of stationary Gaussian processes

2.1 Two senses of Markovian GP

Let $f(t)$ be a GP indexed over time, that is, for any finite time indices $(t_1, \ldots, t_n)$, the joint distribution of $(f(t_1), \ldots, f(t_n))$ is normal [1]. The covariance function, $k(t, s) = \text{cov}(f(t), f(s))$, of a GP fully specifies its probabilistic structure. However, an alternative constructive formulation by P. Levy’s led to the development of a canonical representation for GPs as stochastic integrals with respect to a Brownian motion, from which new Markov properties were formulated.

**Markov in the restricted sense:** In the following, we consider centered stationary GPs which are fully characterized by a translation invariant covariance kernel $k(t, s) = k(\tau)$ where $\tau = |t - s|$. The most well known example of a Markovian GP is the Ornstein-Uhlenbeck (OU) process, with kernel $k(\tau) = \sigma^2 \exp(-\mu \tau)$ [13]. As we will see, it is how one defines a Markov property for GPs that allows for greater insight into their behavior. In this case, the OU process possesses the simplest Markov property in that, $p(f(t) \mid \sigma(f(s)): s < t) = p(f(t) \mid f(s))$, where $\sigma(f(s))$ represents all information known about the process up to time $s$. The fact the OU process is Markov is easily identified by writing down its corresponding SDE and associated solution

$$df(t) = -\mu f(t) dt + \sigma dW(t)$$

$$f(t) = \exp(-\mu(t-s))f(s) + \sigma \int_s^t \exp(-\mu(t - \tau))dW(\tau)$$

where $W(u)$ is the Wiener process (Brownian motion). Note that the solution, which is a Gaussian process, depends only on $f(s)$ given farther history $f(u)$, $\forall u < s$ [18]. Now, it is easy to determine

---

\footnote{rigorously, $\sigma(f(s))$ is the filtration of the process up until time $s$.}
that the conditional distribution, \( f(t) \mid f(s) \),

\[
p(f(t) \mid f(s)) = \mathcal{N}(f(t) \mid k(t - s)k(0)^{-1}f(s), k(0) - k(t - s)^2 k(0)^{-1})
\]  

(3)

Though this example is simple, it is pedagogical in that we can view the kernel not only as a function that tells us the covariance of the process between two time points, but as the operator which propagates the process forward in time. However, the OU process is not a very interesting GP – it is mean square differentiable nowhere, thus its sample functions are very rough [18, 13]. This leads us to P. Levy’s definition of \( N \)-ple Markov in the restricted sense, extending the Markov property to GPs of higher order differentiability [9, 10, 12].

**Definition 1** (\( N \)-ple Markov in the restricted sense). A GP, \( f(t) \), is called \( N \)-ple Markov in the restricted sense if it is exactly \( N \) -times differentiable in mean square and

\[
p(f(t) \mid \sigma(f(s)); s \leq t) = p(f(t) \mid f(s), f^{(1)}(s), \ldots, f^{(N-1)}(s))
\]  

(4)

where \( f^{(i)} \) denotes the \( i \)th mean square derivative of \( f \).

This definition has immediate consequences in reasoning about finitely differentiable GPs which we will show through a motivating example. First, note that a GP and all of its mean square derivatives are jointly Gaussian [18, 19] and define \( k(p)(\tau) = \frac{\partial^p}{\partial \tau^p} k(\tau) \) so that for a stationary GP,

\[
\text{cov}(f^{(p)}(t), f^{(q)}(s)) = (-1)^q k^{(p+q)}(\tau) \quad \text{with} \quad \tau = |t - s|
\]  

(5)

Now, consider a GP, \( f(t) \), with the Matérn \( 3/2 \) kernel parameterized by unit variance and length-scale, \( k(\tau) = (1 + \sqrt{3}\tau) \exp(-\sqrt{3}\tau) \) \[1\]. Since \( k(\cdot) \) is once differentiable, this GP has only one mean square derivative, \( f^{(1)}(t) \), so that \( p(f(t) \mid f(s), f^{(1)}(s); s < t) = p(f(t) \mid f(s), f^{(1)}(s)) \), making it a 2-ple Markov GP in the restricted sense. By the joint Gaussianity, the conditional density is now, considering a GP, \( f(t) \), with the Matérn \( 3/2 \) kernel parameterized by unit variance and length-scale, \( k(\tau) = (1 + \sqrt{3}\tau) \exp(-\sqrt{3}\tau) \) \[1\]. Since \( k(\cdot) \) is once differentiable, this GP has only one mean square derivative, \( f^{(1)}(t) \), so that \( p(f(t) \mid f(s), f^{(1)}(s); s < t) = p(f(t) \mid f(s), f^{(1)}(s)) \), making it a 2-ple Markov GP in the restricted sense. By the joint Gaussianity, the conditional density is

\[
p(f(t) \mid f(s), f^{(1)}(s)) = \mathcal{N}(f(t) \mid [k(t - s), -k^{(1)}(t - s)] \mathbf{K}^{S}(0)^{-1} \begin{bmatrix} f(s) \\ f^{(1)}(s) \end{bmatrix}, q(t - s))
\]  

(6)

where \( q(t - s) = k(0) - [k(t - s), -k^{(1)}(t - s)] \mathbf{K}^{S}(0)^{-1}[k(t - s), k^{(1)}(t - s)]^\top \) and \( [\mathbf{K}^{S}(0)]_{ij} = (-1)^j k^{(i+j)}(\tau) \big|_{\tau=0} \). Note how linear combinations of \( k(t - s) \) and \( k^{(1)}(t - s) \) fully describe how the process, \( f(t) \), evolves over time.

We can generalize this to arbitrary \( N \)-ple Markov GP in the restricted sense and its mean square derivative evolve over time. Define a vector representation \( f^S(t) \equiv [f(t), f^{(1)}(t), \ldots, f^{(N-1)}(t)]^\top \), which, like the OU process earlier, is not differentiable. The vector process \( f^S(t) \) is a 1-ple GP. Now, \( p(f^S(t) \mid f^S(s)) \) can be determined so that

\[
p(f^S(t) \mid f^S(s)) = \mathcal{N}(\mathbf{K}^{S}(t - s)\mathbf{K}^{S}(0)^{-1} f^S(s), \mathbf{Q}(t - s))
\]  

(7)

where \( [\mathbf{K}^{S}(\tau)]_{ij} = (-1)^j k^{(i+j)}(\tau) \). While Eq. (7) could be used to recursively determine the trajectory of \( f(t) \), Eq. (6) could not; retaining the most recent information about the process in conjunction with its mean square derivatives is key to inferring its future behavior.

**Markov in the Hida sense**: Consider a GP, \( f(t) = f_1(t) + f_2(t) \), where \( f_1(t) \) and \( f_2(t) \) are both GPs with differently parameterized Matérn \( 3/2 \) kernels. Then, \( f(t) \) is only once differentiable, however, it is not a 2-ple GP in the restricted sense because Def. 1 is not satisfied. With Levy’s definition being insufficient we now arrive at T. Hida’s more general description of a GP Markov property, which we refer to as \( N \)-ple Markov (in the Hida sense).

**Definition 2** (\( N \)-ple Markov in the Hida sense [12]). A GP, \( f(t) \), is called an \( N \)-ple Markov GP if it admits the following filtered white noise representation:

\[
f(t) = \int_0^t F(t-u)d\mathcal{W}(u)
\]  

(8)

\[
F(t-u) = \sum_{i=1}^{N} g_i(t)h_i(u) \quad u \leq t \quad g_i, h_i : \mathbb{R} \rightarrow \mathbb{C}
\]  

(9)
Brownian Motion

As an example, take any A remarkable fact is that the canonical filter of the process. Substitution of Eq. (9) into Eq. (8) shows \( f(t) \) is constructed as a linear combination of \( N \) additive random processes by writing.

\[
f(t) \triangleq \sum_{i=1}^{N} f_i(t) = \sum_{i=1}^{N} g_i(t) U_i(t)
\]

\[
U_i(t) = \int_{0}^{t} h_i(u) dW(u).
\]

Hence, the process \( f(t) \) would be described as \( N \)-ple Markov in the Hida sense. We see then that if a process is \( N \)-ple Markov (in the Hida sense) it may not be \( N \)-ple Markov in the restricted sense as Def. 2 does not require differentiability of the process. Both of these definitions will be the groundwork for constructing appropriate state space models that can be used for GP inference.

### 2.2 The complete canonical basis

A remarkable fact is that the canonical filter of any stationary \( N \)-ple Markov GP lies in the span of a known set of basis functions. The form of those basis functions as derived by Hida in [12] is given in the following theorem.

**Theorem 1** (Canonical Kernels (p. 102 [12])). If \( f(t) \) is a stationary \( N \)-ple Markov GP, then its canonical filter, \( F(t-u) \), can be represented by a linear combination of the basis

\[
F(t-u) = \sum_{k=1}^{m} c_k (t-u)^{p_k} e^{-\mu_k (t-u)}
\]

where \( \mu_k \in \mathbb{C} \) with \( \text{Re}(\mu_k) > 0 \), \( p_i \in \mathbb{N} \), \( \sum_k p_i = N \), and \( c_k \in \mathbb{C} \).

**Remark 1.** This basis coincides with the basis over elements of the transition matrix of linear dynamical systems in Jordan form [20].

In order to relate a canonical filter of this form to the stationary covariance of the process consider the limit as time approaches infinity, so that any transient behavior dies out; then the covariance between \( f(t+\tau) \) and \( f(t) \) will depend only on the difference \( \tau \) so that we have,

\[
k(\tau) = \lim_{t \to \infty} \mathbb{E} [ f(t+\tau) f(t)^* ]
\]

\[
= \lim_{t \to \infty} \int_{0}^{t} \int_{0}^{t+\tau} F(t-u) F^*(t+\tau-s) dW(u) dW(s)
\]

\[
= \int_{\tau}^{\infty} F(u) F^*(u-\tau) du
\]

As an example, take \( m = 1 \), \( p_1 = 1 \), \( \mu_1 = \sqrt{3} \), and \( c_1 = 12\sqrt{3} \), then \( F(t-u) = 12\sqrt{3} (t-u) \exp(-\sqrt{3}(t-u)) \). By plugging this canonical filter into the equation above, we find that the stationary covariance of this process equals the Matérn \( \frac{3}{2} \) kernel with unit length-scale and variance. This invites the obvious question, what are the equivalent set of basis functions that span the space of admissible covariance functions over stationary, finitely differentiable GPs.
Figure 2: **Left**: Commutative diagram showing how the canonical filter is related to the covariance function and PSD. **Right**: Kernels whom are limit points of Hida-Matérn kernels. The cosine (spectral delta) kernel, squared exponential kernel, and Gabor kernel and their PSD as limits of Hida-Matérn kernels.

### 3 The Hida-Matérn Kernel

Leveraging the basis over canonical filters describing stationary and Markovian GPs, we can determine a corresponding basis over admissible covariance functions. Let $f_p(t)$ be a GP with canonical filter given by a single basis as described in Thm. 1, i.e. $F_p(\tau) = \tau^p e^{-\mu \tau}$, then $f_p(t)$ has power spectral density (PSD), $S_p(\omega)$, given by $S_p(\omega) = \hat{F}_p(\omega) \hat{F}_p(\omega)^*$ with $\hat{F}_p(\omega)$ being the Fourier transform of $F_p(\tau)$. In the case that $\mu$ is complex, then $S_p(\omega)$ would not be the PSD of a real-valued GP as it would not be purely real and symmetric.

Keeping this in mind, we can determine a basis over real-valued covariance functions by isolating the real and symmetric parts of $S_p(\omega)$, giving us $S(\omega)$. Then, by invoking Bochner’s theorem we can arrive at $k(\tau)$ through taking the inverse Fourier transform of $S(\omega)$ [1]. The result is stated below, and the full derivation may be found in the Appendix.

**Proposition 1.** The real-valued covariance function, and PSD corresponding to a canonical basis, $F_p(\tau) = \tau^p e^{-\mu \tau}$, with $\mu = a + jb$, are

$$k_{H,p}(\tau; a, b) = \cos(br) k_{\text{Mat}}(\tau; l = 2 \sqrt{\frac{p}{a}}, \nu = p + \frac{1}{2})$$

$$S_{H,p}(\omega; a, b) = (p!)^2 \left[ \left( \frac{1}{(\omega - b)^2 + a^2} \right)^{p+1} + \left( \frac{1}{(\omega + b)^2 + a^2} \right)^{p+1} \right]$$

where $k_{\text{Mat}}(\tau; l, \nu)$ is the general Matérn covariance kernel of order $\nu$ and length-scale $l$. Although simple, these kernels, which we denote Hida-Matérn kernels, span the space of stationary and finitely differentiable GP covariance functions. Similar to a standard Matérn kernel the parameter $p$ controls the differentiability/smoothness, $a$ is the inverse length-scale, and $b$ controls the center of the PSD. To conceptualize the Markov property better take $p = N$, and $b \neq 0$, then a GP with covariance function $k_{H,N}(\tau; a, b)$ will be $2N$-ple Markov in the Hida sense; however, when $b = 0$, then this covariance function coincides exactly with the Matérn kernels and such a GP would be $N$-ple Markov in the restricted sense.

**Theorem 3.1** (Stationary Hida-Matérn kernels forms a universal class). For any fixed value of $p$, Hida-Matérn kernels are dense in the space of square integrable stationary kernels, hence universal.
3.1 The family of Hida-Matérn kernels

With the Hida-Matérn kernel defined, we can now consider the family formed by their linear combinations. Doing so, a mixture of Hida-Matérn (MHM) kernels can be defined as,

$$k_{H,N,p,c}(\tau) = \sum_{i=1}^{L} c_i k_{H,p_i}(\tau)$$

(16)

$$p = (p_1 \cdots p_L)\top \in \mathbb{R}^M, \quad \sum_{i=1}^{L} p_i = N, \quad i \in \mathbb{N}$$

(17)

$$c = (c_1 \cdots c_L)\top \in \mathbb{R}^L, \quad c_i \geq 0, \quad \forall i$$

(18)

where \( p \) specifies the mixands smoothness, \( c \) their respective weights, and \( N \) the order of the Markov property in the Hida sense. Note how based on the value of each \( b_i \), if \( N = \sum_{i=1}^{L} p_i \), then the process can be anywhere from \( N \)-ple Markov to \( 2N \)-ple Markov in the Hida sense. Since Hida-Matérn kernels with \( p \) fixed form a universal class, the Hida-Matérn mixture kernels do as well.

4 Hida-Matérn State Space Representations

The intuition regarding Markov GPs that we developed earlier will help now in constructing a state-space representation of GPs that can be described by the Hida-Matérn class of kernels [18]. Though the state-space representation of GPs has been used successfully in the literature, construction of an appropriate SSM usually begins by correctly parameterizing a linear SDE whose solution has a stationary distribution coinciding with the GP of interest [8, 14, 17]. In contrast, the construction we present leverages the \( N \)-ple Markov property and allows for a state-space representation that only depends on the kernel and its derivatives. To start, consider GPs that are \( N \)-ple Markov in the restricted sense, which for now, limits our scope to Hida-Matérn kernels of order \( N \) with \( b = 0 \) (or Matérn kernels of order \( \nu = p + \frac{1}{2} \) for \( p \in \mathbb{N} \)). In the course of doing so, we will gradually expand upon this construction to handle the full family of Hida-Matérn kernels.

4.1 SSMs for \( N \)-ple GPs in the restricted sense

Let \( f(t) \) be an \( N - 1 \) times differentiable, and stationary GP with kernel \( k_{H,N}(\tau; a, 0) \), then \( f(t) \) is an \( N \)-ple Markov GP. By consolidating \( f(t) \) and its mean square derivatives into the vector process

$$\mathbf{f}^S(t) = \begin{pmatrix} f(t) & f^{(1)}(t) & \cdots & f^{(N-1)}(t) \end{pmatrix}\top \in \mathbb{R}^N$$

(19)

we have that \( \mathbf{f}^S(t) \) is a 1-ple Markov GP in the restricted sense so that \( p(\mathbf{f}^S(t)|\mathbf{f}^S(s); s < t) = p(\mathbf{f}^S(t)|\mathbf{f}^S(s)) \). Recognizing \( \mathbf{f}^S(t) \) is a multioutput GP described by the kernel \( \mathbf{K}^S(\tau) \), with \( [\mathbf{K}^S(\tau)]_{ij} = (-1)^j k^{(i+j)}(\tau) \), allows us to write that conditional distribution as follows,

$$p(\mathbf{f}^S(t)|\mathbf{f}^S(s)) \sim \mathcal{N}((\mathbf{f}^S(t)|\mathbf{A}(\tau)\mathbf{f}^S(s), \mathbf{Q}(\tau))$$

(20)

$$\mathbf{A}(\tau) = \mathbf{K}^S(\tau)\mathbf{K}^S(0)^{-1}\mathbf{f}^S(s)$$

(21)

$$\mathbf{Q}(\tau) = \mathbf{K}^S(0) - \mathbf{K}^S(\tau)\mathbf{K}^S(0)^{-1}\mathbf{K}^S(\tau)\top$$

(22)

where \( \tau = |t-s| \) with \( t > s \). [21] It’s easy to see now that the conditional density in Eq. (20) can be rewritten as a difference equation so that \( \mathbf{f}^S(t) \) is equivalently described by,

$$\mathbf{f}^S(t) = \mathbf{A}(\tau)\mathbf{f}^S(s) + \epsilon(\tau)$$

(23)

$$\epsilon(\tau) \sim \mathcal{N}(\epsilon(\tau)|0, \mathbf{Q}(\tau))$$

(24)

The main object of interest \( f(t) \) is easily extracted from \( \mathbf{f}^S(t) \) by letting \( \mathbf{h} = (1 0 \cdots 0)\top \) so that \( f(t) = \mathbf{h}\top \mathbf{f}^S(t) \). Thus, we can easily reason about the behavior of \( f(t) \) through linear combinations of its past mean square derivatives or about \( \mathbf{f}^S(t) \) as a whole. Much like our motivating examples earlier, \( \mathbf{K}^S(\tau) \) alone characterizes how the process propagates forward through time. In order to motivate the practical utility of this construction we proceed by considering standard GP regression.
4.2 GP Regression with finite SSMs

In a GP regression setting with noisy observations, \( \{y_i(t_i)\}_{i=1}^M \), and stationary GP \( f(t) \sim \mathcal{GP}(0, k(t)) \) that is \( N - 1 \) times differentiable, the standard generative model, as in [1], can be recast as a linear Gaussian SSM using Eq. (23) so that,

\[
\begin{align*}
    f^S(t_{i+1}) &= A(\tau_i) f^S(t_i) + \epsilon(\tau_i) \\
    y(t_i) &= h^\top f^S(t_i) + \nu
\end{align*}
\]

with \( t_{i+1} > t_i \) \( \forall i \), \( \tau_i = |t_{i+1} - t_i| \), and \( \nu \sim \mathcal{N}(0, \sigma^2) \). Since the aforementioned SSM is stable it admits a stationary covariance \( P_{\infty} = \lim_{t \to \infty} \mathbb{E}[f^S(t)f^S(t)^\top] \) [22]. Ensuring the process begins in the stationary state amounts to specifying \( f^S(t_0) \sim \mathcal{N}(0, P_{\infty}) \). For SSMs, determining \( P_{\infty} \) involves finding the solution of the continuous/discrete Lyapunov equation, however in the Appendix we show that \( P_{\infty} = K^S(0) \). In this form, the usual Kalman filtering and smoothing algorithms can be used to recover the posterior in \( O(MN^3) \) time, where \( M \) is the number of data points.

While the Kalman filtering algorithm is computationally efficient, ensuring that it is numerically stable can often be difficult. As a result, ill numerical conditioning will be exacerbated by the fact that elements of \( K^S(\tau) \) are increasing in magnitude towards the bottom right, which we illustrate in Fig. 5. Approaches such as balancing or Nordsieck coordinate transformations, often used to combat this problem, formulate a surrogate SSM that can be used for equivalent inference [23, 24, 15, 25]. Similar in spirit to those approaches, we propose a correlation transform, taking advantage of the SSMs formulation in terms of covariances. Concretely, consider a linear transformation of the original process, \( z^S(t) = Cf^S(t) \) where

\[
    [C]_{ii} = 1/\sqrt{[K^S(0)]_{ii}}
\]

so that \( K^S_z(\tau) \triangleq \text{cov}(z^S(t + \tau), z^S(t)^\top) = CK^S(\tau)C^\top \). An alternative SSM is formed by substituting \( K^S_z(\tau) \) for \( K^S(\tau) \) and rewriting the observation equation as \( y(t_i) = h^\top C^{-1}z^S(t_i) + \nu \). Whereas \( K^S(0) \) would be ill conditioned and thus cause problems when taking its inverse, \( K^S_z(0) \) will have a significantly lower condition number, and we can expect the better numerical conditioning to provide more accurate inference. After inference pertaining to \( z^S(t) \) is made, properties of interest related to \( f^S(t) \) are easily recovered.

4.3 SSMs for general N-ple GPs

Having introduced how an SSM for a GP with Hida-Matérn kernel of order \( N \) when \( b = 0 \) can be constructed, we are now in position to consider the general case. In order to make this jump, first observe that when \( b \neq 0 \), a GP, \( f(t) \), with kernel \( k_{H,N}(\tau; a, b) \), will be a \( 2N \)-ple process in the Hida sense but not the restricted sense. We can see this most readily by breaking the cosine term into a sum of two complex exponentials so that,

\[
    k_{H,N}(\tau; a, b) = \frac{1}{2}e^{ib\tau}k_{H,N}(\tau; a, 0) + \frac{1}{2}e^{-ib\tau}k_{H,N}(\tau; a, 0)
\]

\[
    = \Re\{e^{ib\tau}k_{H,N}(\tau; a, 0)\}
\]

which shows \( k_{H,N}(\tau; a, b) \) is \( N \) times differentiable, but the sum of complex conjugate Hida-Matérn kernels of order \( N \). However, it is apparent that the representation given by Eq. (27) is
redundant due to the complex conjugacy. Hence, we can consider a complex GP, $z(t)$, with kernel $k_z(\tau; a, b) = e^{i\theta}k_{H,N}(\tau; a, 0)$, so that an equivalent description of $f(t)$ is given by the SSM,

$$z^S(t) = A_z(\tau)z^S(s) + \epsilon_z(\tau)$$

$$f(t) = \text{Re}\{h^Tz^S(t)\}$$

$$\epsilon_z(\tau) \sim N(\epsilon_z(\tau)|0, Q_z(\tau))$$

where

$$A_z(\tau) = K_z^S(\tau)K_z^S(0)^{-1}$$

$$Q_z(\tau) = K_z^S(0) - K_z^S(\tau)K_z^S(0)^{-1}K_z^S(\tau)^H$$

and similar to earlier, $z^S(t) = (z(t) \cdots z^{(N-1)}(t))^T$, and $[K_z^S(\tau)]_{ij} = (-1)^{i+j}k_z^{(i+j)}(\tau)$. This shows that we can reason about these particular $2N$-ple Markov GPs, in the Hida sense, with an $N$-dimensional state space model by exploiting the complex-conjugate symmetries present. Sans the necessity of having to work with complex numbers, this formulation lends itself to GP regression exactly as described in Section 4.2.

### 4.4 SSMs for Hida-Matérn mixtures

Now that we have explored how to formulate the SSM describing a GP whose covariance function is an elementary Hida-Matérn kernel, it is a trivial extension to consider the SSM formulation for the Hida-Matérn mixture kernel, $k_{H,N,p,e}(\tau)$.

$$K_{H,N,p,e}^S(\tau) = \text{diag}(c_1K_{H,p_1}^S(\tau), c_2K_{H,p_2}^S(\tau), \cdots, c_MK_{H,p_M}^S(\tau))$$

so that $K_{H,N,p,e}^S(\tau) \in \mathbb{R}^{N \times N}$, and $K_{H,p}^S(\tau) \in \mathbb{R}^p \times p$. Thus, we can engineer arbitrarily complex kernels as linear combinations of Hida-Matérn kernels, yet work with them in the same manner by constructing an SSM with $K_{H,N,p,e}^S(\tau)$.

### 5 Connection with the SDEs driven by Brownian motion

The solution of all linear SDEs driven by Brownian motion is a simple Markov GP [18, 26, 27]. After having explored the properties of being Markov in the restricted sense, and Markov in the Hida sense, it is natural to ask how they relate to the solution of a linear SDE. Let’s take a linear SDE, driven by Brownian motion,

$$df^S(t) = Ff^S(t) dt + LdW(t)$$

with $f^S(t) \in \mathbb{R}^N$, $L \in \mathbb{R}^{N \times M}$ and $W(t)$ an $M$ dimensional Brownian motion. The solution of this SDE can be expressed as

$$f^S(t) = \Phi(t)f^S(s) + \int_0^t \Phi(t-u)L dW(u)$$

where $\Phi(t) = \exp(Ft)$ is the state transition matrix of the system and $\tau = t - s$ [18]. For the SDE solution of $f^S(t)$ and Eq. (23) to be consistent with one another it has to be true that $\Phi(t) = \exp(Ft) = K^S(\tau)K^S(0)^{-1}$.

Now, note that if we were to take the Jordan decomposition of $F$, it will have as many Jordan blocks as mixands; each Jordan blocks size corresponding to the order of a single mixand. Let $G$ be an alternative dynamics matrix, but with the same Jordan decomposition as $F$ so that $F = JCJ^{-1}$ and $G = DJD^{-1}$. With $C$ and $D$ both full rank, there exists a linear transformation $X$ so that $XC = D$; meaning that $\exp(Gt) = \exp(DJD^{-1}t) = D\exp(Jt)D^{-1} = XC\exp(Jt)C^{-1}X^{-1}$. Thus, if $f^S(t)$ is a GP described by an SDE with the dynamics $F$, then $g^S(t) = Xf^S(t)$ is a GP described by the dynamics $G$, resulting in the following Lemma.

**Lemma 5.1.** If given a linear and finite-dimensional SDE, written as

$$df^S(t) = Ff^S(t) dt + LdW(t)$$

where $F \in \mathbb{R}^{N \times N}$, $L \in \mathbb{R}^{N \times M}$, and $W(t) \in \mathbb{R}^M$ is an $M$-dimensional Brownian motion process. If the solution, $f^S(t) \in \mathbb{R}^N$, then, it must be the case that Eq. (37) corresponds to a linear transformation of the SSM formulation of an $N$-ple GP in the Hida-Matérn family.
In other words, a linear SDE with dynamics matrix $F$ having $L$ Jordan blocks is equivalent to a linear transformation of the SSM formulation of a Hida-Matérn GP with $L$ mixands. With this, we now have a way to consider kernels that map $\mathbb{R} \rightarrow \mathbb{R}^D$.

5.1 Multi-output Hida-Matérn kernels

To reiterate the previous realization, any linear and finite dimensional SDE can be interpreted as a linear change of coordinates of the SSM representation of a Hida-Matérn mixture. The Jordan block structure of the associated dynamics matrix will dictate the number of mixands and their Markov order, while the eigenvalues of those blocks are determined by each mixands hyperparameters.

To consider multi-output GPs take a MHM GP, $f(t)$, with kernel $k_{H,N,p,c}(\tau)$, but ignore the observation equation so that we are only concerned with $f^S(t)$. As we said, $f^S(t)$ has an equivalent SDE formulation, with associated dynamics matrix $F$ having a particular Jordan block structure. Taking $g^S(t) = Xf^S(t)$, we have a new GP of the same dimension, and its SSM formulation becomes immediate as

$$
\text{cov}(g^S(t + \tau), g^S(t)) = \text{cov}(Xf^S(t + \tau), f^S(t)X^\top) = XK^S(\tau)X^\top
$$

(38)

Now, we have that $g^S(t)$ is a multi-output MHM (MO-MHM) GP, i.e. $g^S(t) : \mathbb{R} \rightarrow \mathbb{R}^d$. Taking this one step farther, the observations can be modified so that rather than projecting the latent process onto one dimension via $h$, we project it to $\mathbb{R}^D$ via $H \in \mathbb{R}^{d \times D}$. The augmented SSM becomes

$$
g^S(t) = A_p(\tau)g^S(s) + \epsilon_p(\tau)
$$

(39)

$$
y(t) = Hg^S(t) + \nu
$$

(40)

Whereas before, we had latents $f^S(t)$ that represented the process and its mean square derivatives, the latents, $g^S(t)$, are their linear combinations as mapped by $X$. We note that these linear transformations do not alter the Markov property of the process, they simply transform it to a different coordinate system. From the discussion earlier, it is also clear that the dynamics matrix of the SDE describing $g^S(t)$ has the same Jordan block structure as that of $f^S(t)$.

6 Numerical and Algorithmic Properties

In Section 4.2 we noted that the structure of $K^S_{H,p}(\tau)$ can lead to technical difficulties which make naive state space inference infeasible. The proposed correlation transform amends the ill numerical conditioning that manifests itself in a naive implementation and makes it possible to work with higher order Hida-Matérn kernels. There are additional precautions we may take that guarantee higher numerical stability as well as reductions in computational complexity.

6.1 Special structure of $K^S_{H,p}(\tau)$

As the building blocks of $K^S_{H,p}(\tau)$ are covariances between the process and its mean square derivatives, one may suspect that $K^S_{H,p}(\tau)$ has some special structure that can be exploited. In fact, many of the properties $K^S_{H,p}(\tau)$ exhibit are similar to those of matrices explored in [28].

For any multioutput covariance matrix formed from a single Hida-Matérn kernel we have that the $i^{th}$ off diagonal, up to appropriate sign flips, will only consist of the $i^{th}$ derivative of the kernel evaluated at $\tau$. Hence, the multioutput covariance kernel is composed of only $2N - 1$ unique elements, leading to memory requirements that scale linearly with the order of the kernel used.
Let’s now consider the inversion of $K^S_{H,N}(0)$. If the correlation transform is properly used, then this inverse should not be a source of much trouble. However, when $b = 0$, the multioutput covariance kernel will be identically 0 at all indices $i, j$ such that $i + j$ is odd. To see this, note that the PSD of the Hida-Matérn kernel is real and symmetric, and that $|K^S_{H,N}(0)|_{ij}$ exactly coincides with the $(i + j)^{th}$ moment of $S(\omega)$, which is 0 for $i + j$ odd. Recognizing this, elementary row and column operations can be used to transform the covariance matrix into the block matrix

$$R_L \cdots R_1 K^S_{H,N}(0) C_1 \cdots C_L = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

with $R_i$ and $C_i$ being elementary row/column operations. The inverse of the altered matrix is then easily taken block by block. Subsequent application of the inverse row and column operations return the desired inverse. When $b \neq 0$ we do not have this sparsity present as we have chosen to work with the reduced, yet equivalent, SSM. Understanding its structure can still help us to achieve improved numerical stability by enforcing properties we expect that numerical noise may break. For example, if $b \neq 0$ then $K^S_{H,N}(0)$ should be purely imaginary at indices such that $i + j$ is odd. Hence, its inverse should retain this structure and if numerical noise causes these entries to become real, then they are easily masked.

![Condition No. Comparison](image)

Figure 5: Illustrative figure showing the ill numerical conditioning if a coordinate transformation is not used for a Hida-Matérn kernel $k_{H,S}(\tau; 1, 0)$. To the left of condition number plots are example matrices for $\tau = 0.5$; in the orange box (□) is the matrix before the correlation transform, above is after. Values of $\tau$ plotted are 0.0, 0.001, 0.01, 0.1, and 0.5. **left:** $K(\tau)$ clearly benefits from the correlation transform, as evident in both the heatmap of the matrix and its condition number as a function of $\tau$. **middle:** $A(\tau)$ without the correlation transform has elements of increasing magnitude in the bottom left; the correlation transform amends this by creating a more homogenous matrix in terms of relative magnitudes. **right:** $Q(\tau)$ benefits the least from the correlation transform, however, it is clearly more balanced as seen from the heatmap.

6.2 Kalman updates

In performing Kalman filtering over $M$ time-steps, the updated covariance, $P_m$ at time-step $m$, is not guaranteed to retain the positive semidefinite structure of a proper covariance matrix. Often, the Joseph form of the covariance update or square root filtering can be used to ensure that positive semidefiniteness is not lost. However, since $h$ is sparsely populated (i.e. a kernel which is the sum of $L$ Hida-Matérn kernels will have $L$ nonzero values), the equations for the updated mean and covariance, $m_m$ and $P_m$, can be simplified for additional computational savings and superior numerical stability.

7 Relation to Other Kernels

Previously we noted that many frequently used kernels either reside in the Hida-Matérn family or can be approximated appropriately. Take as an example the squared exponential kernel, it is well known that the standard Matérn family of kernels with smoothness parameter $\nu$ approaches a squared exponential as $\nu \to \infty$. Remembering this, it is clear that the Spectral Mixture family of kernels is an asymptotic limit of a Hida-Matérn mixture as the smoothness parameter $p_i \to \infty$, $\forall i$ [29].

The cosine kernel, $k(\tau) = \sigma^2 \cos(b\tau)$, is also not directly within the Hida-Matérn family, but it is approached in the limit $a \to 0$ for $\sigma^2 k_{H,P}(\tau; a, b)$. As another example, take the periodic kernel defined as $k(\tau) = \sigma^2 \exp\left(-\frac{2}{\pi} \sin^{-1}\left(b \frac{\tau}{2}\right)\right)$, which can be expanded using its Taylor series expansion.
representation [16]

\[
k_{PER}(\tau) = \sigma^2 \exp \left( - \frac{2 \sin^2(\omega_0 \frac{\pi}{2})}{t^2} \right)
\]

(42)

\[
= \exp(-l^2) \sum_{q=0}^{\infty} \frac{1}{q!} \cos^q(\omega_0 \tau)
\]

(43)

\[
= \exp(-l^2) \sum_{q=0}^{\infty} \sum_{v=0}^{q} \left( \begin{array}{c} q \\ v \end{array} \right) \frac{1}{q!2^q} \text{Re} \left( e^{i\omega_0 \tau(q-2v)} \right)
\]

(44)

\[
\approx \exp(-l^2) \sum_{q=0}^{L} \sum_{v=0}^{q} \left( \begin{array}{c} q \\ v \end{array} \right) \frac{1}{q!2^q} \text{Re} \left( e^{i\omega_0 \tau(q-2v)} \right)
\]

(45)

\[
= \exp(-l^2) \sum_{q=0}^{L} \sum_{v=0}^{q} \left( \begin{array}{c} q \\ v \end{array} \right) \frac{1}{q!2^q} k_{H,p}(\tau; a \to 0, b = \omega_0(q - 2\nu))
\]

(46)

Which illustrates how a kernel such as the periodic covariance can be decomposed such that it can be approximated by a Hida-Matérn mixture.

**Remark 2.** Certain kernels such as the squared exponential do not fall strictly under this class as they have a countably infinite number of derivatives and are analytic. Such GPs are regarded as “completely deterministic” [9]– meaning if observed for an infinitesimal amount of time their future behavior is in theory completely known [13].

| kernel | \(k(\tau)\) |
|--------|-------------|
| Squared Exp. | \(\sigma^2 \exp(-\frac{1}{2\pi^2} \tau^2)\) |
| Rational Quadr. | \((1 + \frac{x^2}{2\alpha^2})^{-\alpha}\) |
| Gabor. | \(\sigma^2 \cos(2\pi b) \exp(-\frac{1}{2\pi^2} \tau^2)\) |
| Sinc | \(\sigma^2 \text{sinc}(\Delta \tau) \cos(2\pi b \tau)\) |

Table 1: Functional form of stationary kernels used throughout the paper. \(\alpha > 0, l > 0, \nu > 0\), and \(K_p\) is the modified Bessel function of the second kind. See [30] for the Sinc kernel, [29] for the Spectral Mixture kernel, and [1] for further details on other kernels.

Now consider the LEG family of kernels introduced in [31]. Take \(z(t)\) to be the solution of a linear SDE driven by Brownian motion. Linear transformation of \(z(t)\) and addition of Gaussian noise gives an observed process \(x(t)\) so that

\[
dz(t) = -\frac{1}{2} \mathbf{G} z(t) dt + \mathbf{N} dW(s)
\]

(47)

\[
x(t) = \mathbf{B} z(t) + \mathbf{A} \epsilon(t)
\]

(48)

\[
\epsilon(t) \sim \mathcal{N}(\epsilon(t)|0, \mathbf{I})
\]

(49)

In which case it is said that \(x(t) \sim \text{LEG} (\mathbf{N}, \mathbf{R}, \mathbf{B}, \mathbf{A})\) where \(\mathbf{G} = \mathbf{N}^T + \mathbf{R} - \mathbf{R}^T\). The equivalence to a SIMO Hida-Matérn kernel is immediate from the discussion earlier on SDEs; the number of mixands determined by the Jordan block structure of \(\mathbf{G}\), with their hyperparameters determined by the eigenvalues of \(\mathbf{G}\).

To make the equivalence more concrete, decompose \(-\frac{1}{2} \mathbf{G}\) into \(\text{CJC}^{-1}\) where \(\mathbf{J}\) is a Jordan block matrix. Say that \(\mathbf{J}\) has \(m\) blocks, each of size \(p_i\), \(i = 1, \ldots, m\). Then, a GP \(f(t)\), with Hida-Matérn mixture kernel containing \(m\) mixands, the \(i\)-th having order \(p_i\), has an SSM formulation equivalent to the solution of a linear SDE with dynamics matrix \(\mathbf{F}\) such that \(\mathbf{F} = \text{DJD}^{-1}\). Taking \(\mathbf{X} = \mathbf{C}\), the SSM formulation of the vector process \(g^\mathbf{S}(t) = \mathbf{X} f^\mathbf{S}(t)\) will then be the solution of the SDE as defined in Eq. (47). Further, setting \(\mathbf{H} = \mathbf{B}\) and taking the covariance of \(\nu\) to be \(\mathbf{A} \mathbf{A}^T\) shows GPs defined either way are equivalent.

11
7.1 Approximating arbitrary kernels

Imagine a scenario where a kernel, $k_{\text{ref}}(\tau; \theta)$, has been designed, its hyperparameters chosen, and we wish to make GP inference under this kernel. If the data set is small, exact GP inference can be used, however, scalability quickly becomes a concern and only approximate methods are viable. When the kernel of interest can be approximated well by linear combinations of Hida-Matérn kernels, then the SSM formulation presented is an appealing avenue.

Under this scenario, the question becomes how can the parameters of a Hida-Matérn mixture, $k_{H,N,p,c}(\tau)$, be estimated so that it closely matches $k_{\text{ref}}(\tau; \theta)$. A practical, yet simple manner of estimating these hyperparameters is to minimize the squared loss between the reference kernel and the Hida-Matérn mixture with respect to the mixtures hyperparameters. This results in the following optimization problem,

$$\nu = \arg\min_{\nu} \int (k_{\text{ref}}(\tau; \theta) - k_{H,N,p,c}(\tau))^2 d\tau \tag{50}$$

where $\nu$ contains all hyperparameters of the Hida-Matérn mixture to optimize. Through Parseval’s Theorem, we can see that the objective in Eq. (50) not only minimizes the squared distance between the mixture and target kernels but also the squared distance of their PSDs.

In Fig. 6 are plotted Hida-Matérn mixtures containing four mixands fit to various stationary kernels. Each mixand is a second order Hida-Matérn kernel and its parameters ($a, b, \sigma$) were optimized so that the squared distance between the mixture and reference kernel were minimized.

8 Experiments

8.1 Mauna Loa Carbon Dioxide

Here, we examine the predictive capabilities and expressivity of the Hida-Matérn family by applying it to the popular Mauna Lua CO$_2$ dataset. The long upward trend present in this data as well as the yearly periodic component mean that an appropriate linear combination of Hida-Matérn kernels will need to be constructed. To demonstrate that a low order SSM is sufficient for this dataset we construct a simple kernel that is the sum of two order 3 Hida-Matérns, i.e.

$$k_{H,6}(\tau) = c_1 k_{H,3}(\tau; a_1, b) + c_2 k_{H,3}(\tau; a_2, 0)$$
We use data recorded up until 2004 after which only predictions are made. Examining the posterior predictive we see that the Hida-Matérn kernels used in conjunction are able to sufficiently extrapolate future predictions.

8.2 Scalability of SSM representation

From a practical standpoint, one of the most useful implications that arises from the N-ple GP Markov characterization is the ability to quickly formulate a corresponding state space model amenable for inference. As a result, it is trivial to form appropriate state space models that can be used to recover exact GP inference over large datasets where naive GP regression would be impractical due to the cubic scaling of the computational complexity.

In this toy example, we consider 50,000 data points along a uniform grid generated by the sum of two Spectral Mixture kernels. [29] We compare posterior inference made with the SSM formulated using various Hida-Matérn kernels, inducing points, KISS-GP [7], and random Fourier features.

As the data points are distributed on a uniform grid we expect that inference using the SSM and Hida-Matérn kernels should be the fastest as $K^S(\Delta)$ only need be computed once. Indeed, this is the case and it also has the lowest marginal KLD to the true posterior even though there is a model mismatch (as the Spectral Mixture is only an asymptote of the Hida-Matérn family).
9 Conclusion

We showed how viewing GPs through the lens of their Markov property has both theoretical and practical consequences. We reintroduced results from Hida that all finitely differentiable stationary GPs are Markovian and their kernel must admit a decomposition in terms of linear combinations of the derived Hida-Matérn kernels. As a consequence, Hida-Matérn GPs can be simply rewritten as a linear Gaussian state-space model. The SSM representations enabled us to make exact GP inference in linear time for any 1-dimensional stationary GP whose kernel is in the Hida-Matérn family. As a byproduct of the admitted SSM representation, we also fleshed out connections to SDEs whose solutions are GPs, showing that the fundamental matrix solution of those systems has a closed form representation. Finally, we showed many commonplace kernels either reside directly within the Hida-Matérn family or can be seen as appropriate asymptotic limits. The Hida-Matérn kernel provides a unifying framework that bridges linear models used in the statistical signal processing literature and the nonlinear kernel methods in the machine learning literature.

References

[1] C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. The MIT Press, November 2005.
[2] T. Karaletsos and T. D. Bui. Hierarchical gaussian process priors for bayesian neural network weights. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 17141–17152. Curran Associates, Inc., 2020.
[3] M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary gaussian process classification. Journal of Machine Learning Research, 6(57):1679–1704, 2005.
[4] Y. C. Ng, N. Colombo, and R. Silva. Bayesian semi-supervised learning with graph gaussian processes. In S. Bengio, H. Wallach, H. Larochelle, et al., editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.
[5] T. D. Bui, C. Nguyen, and R. E. Turner. Streaming sparse gaussian process approximations. In I. Guyon, U. V. Luxburg, S. Bengio, et al., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.
[6] M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In D. van Dyk and M. Welling, editors, Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, volume 5 of Proceedings of Machine Learning Research, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.
[7] A. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (KISS-GP). In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 1775–1784, Lille, France, 2015. PMLR.
[8] J. Hartikainen and S. Sarkka. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In 2010 IEEE International Workshop on Machine Learning for Signal Processing, pages 379–384. IEEE, 2010.
[9] P. Lévy. A special problem of brownian motion, and a general theory of gaussian random functions. In J. Neyman, editor, Contributions to Probability Theory, pages 133–176. University of California Press, 1956.
[10] P. Lévy. Wiener’s Random Function, and Other Laplacian Random Functions. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pages 171–187, 1951.
[11] T. Hida. Canonical representations of Gaussian processes and their applications. Memoirs of the College of Science, University of Kyoto. Series A: Mathematics, 33(1):109 – 155, 1960.
[12] T. Hida and M. Hitsuda. Gaussian Processes. American Mathematical Society, 1993.
[13] M. L. Stein. Interpolation of Spatial Data. Springer Series in Statistics. Springer New York, 1999.
[14] A. Solin, J. Hensman, and R. E. Turner. Infinite-horizon gaussian processes. In S. Bengio, H. Wallach, H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] A. Corenflos, Z. Zhao, and S. Särkkä. Temporal Gaussian Process Regression in Logarithmic Time. *arXiv:2102.09964 [cs, stat]*, May 2021.

[16] A. Solin and S. Särkkä. Explicit Link Between Periodic Covariance Functions and State Space Models. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 904–912, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[17] A. Solin. *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. PhD thesis, Aalto University, 2016.

[18] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Courier Corporation, January 2007.

[19] S. Särkkä. Linear operators and stochastic partial differential equations in gaussian process regression. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 151–158. Springer Berlin Heidelberg, 2011.

[20] R. W. Brockett. *Finite Dimensional Linear Systems*. Society for Industrial and Applied Mathematics, May 2015.

[21] M. A. Álvarez and N. D. Lawrence. Computationally Efficient Convolved Multiple Output Gaussian Processes. *Journal of Machine Learning Research*, 12(41):1459–1500, 2011.

[22] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., 1979.

[23] M. R. Osborne. On Nordsieck’s method for the numerical solution of ordinary differential equations. *BIT Numerical Mathematics*, 6(1):51–57, 1966.

[24] A. Nordsieck. On Numerical Integration of Ordinary Differential Equations. *Mathematics of Computation*, 16(77):22–49, 1962.

[25] N. Krämer and P. Hennig. Stable implementation of probabilistic ode solvers, 2020.

[26] S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 1 edition, 2019.

[27] B. Oksendal. *Stochastic Differential Equations (3rd Ed.): An Introduction with Applications*. Springer-Verlag, 1992.

[28] G. Strang and S. MacNamara. Functions of difference matrices are toeplitz plus hankel. *SIAM Review*, 56:525–546, 08 2014.

[29] A. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. *International conference on machine*, 2013.

[30] F. Tobar. Band-limited gaussian processes: The sinc kernel. In H. Wallach, H. Larochelle, A. Beygelzimer, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] J. Loper, D. Blei, J. P. Cunningham, and L. Paninski. Linear-time inference for gaussian processes on one dimension, 2021.

[32] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS’07, pages 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc.

[33] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In S. Bengio, H. Wallach, H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[34] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Signal Processing Series. Pearson, 3rd edition, 2014.

[35] H. Bateman. *Tables of Integral Transforms Volume 1*. Bateman Manuscript Project. McGraw-Hill, 1st edition, 1954.
[36] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing*. Pearson Prentice Hall, 4th ed edition, 2007.

[37] W. Rudin. *Functional Analysis*. McGraw-Hill, Inc, 1991.

[38] Y.-L. K. Samo and S. Roberts. Generalized Spectral Kernels. 2015.

[39] T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall Information and System Sciences. Prentice Hall, New Jersey, 2000.

[40] K. Ogata. *Modern Control Engineering*. Prentice-Hall Electrical Engineering Series. Instrumentation and Controls Series. Prentice-Hall, Boston, 5th ed edition, 2010.
10 Appendix

A Experimental Details

A.1 Mauna Loa

The prior GP for analyzing the Mauna Loa data set is set to

\[ k_{H,6}(\tau) = c_1 k_{H,3}(\tau; a_1, b) + c_2 k_{H,3}(\tau; a_2, 0) \]

with \( c_1 = 0.05^2, a_1 = 1/25, b = 2\pi, \) and \( c_2 = 2.3^2, a_2 = 1/100, \) i.e. one kernel which is periodic with a short lengthscale to capture the yearly periodic trend and another which is not periodic with a long lengthscale to capture the linear trend. This setup is similar to those commonly used in the literature, where the first additive kernel would usually be a Matérn/Squared Exponential multiplying the periodic covariance function \([1, 15, 16]\).

The GP is fit using the data from 1974-2004, and predictions are made for the time window from 2004-2020 as shown in Fig. 7. From the fit, we can see that the sum of these two low order Hida-Matérn kernels form a covariance function such that the resulting GP inference can make predictions that capture both the seasonal and linear trends in the data.

A.2 Toy Dataset

For the toy dataset, 50,000 observations, with uniform spacing of 0.05, are generated according to a prior GP whose covariance function is the sum of two spectral mixture kernels, i.e., \([?]\)

\[ k(\tau) = k_{SM1}(\tau; c_1, l_1, \omega_1) + k_{SM2}(\tau; c_2, l_2, \omega_2) \]  

with \( k_{SMi} = c_i \exp\left(-\frac{\tau^2}{2l_i^2}\right) \cos(\omega_i \tau) \)

where the hyperparameters are \( c_1 = 1.5^2, l_1 = 2.0, \omega_1 = 2\pi \cdot 0.01, \) and \( c_2 = 1.5^2, l_2 = 2.0, \omega_2 = 2\pi \cdot 0.05. \) For comparison, we consider SVGPR, KISSGP, as well as random Fourier features \([6, ?, 32]\). Experiments for SVGPR and KISSGP are ran using GPyTorch with all hyperparameter optimization done before computing the wall-clock time of the calculation for the posterior distribution over the grid \([33]\). To subserve any numerical difficulties that would arise from calculating the KLD between the posterior under each method and exact GP inference we instead consider the average KLD per grid point, or average marginalized KLD. As this inference is over a grid, we note that the SSM inference only need calculate the state space matrices once and can then sweep through the entire data set.

B Derivation of the Hida-Matérn kernel starting from the basis of canonical filters

Let \( f_p(\tau) \) be a canonical filter of order \( p \), i.e.,

\[ f_p(\tau) = \tau^p \exp(-\mu \tau) \]

the covariance function, \( r_p(\tau) \), of a process described with such a canonical filter is

\[ r_p(\tau) = \int_0^\infty f_p(t)f_p^*(t-\tau)dt \]

where * denotes complex conjugation. The PSD, \( S(\omega) \), of \( r_p(\tau) \) is related to the fourier transform of \( f_p(\tau) \) so that we have \( S(\omega) = F_p(\omega)F_p^*(-\omega) \) where \( F_p(\tau) = \mathcal{F}[f_p(\tau)] \). For convenience we proceed by working with the Fourier cosine and Fourier sine transforms denoted by

\[ F_{C,p}(\omega) = \mathcal{F}^C[f_p(\tau)] = \int_0^\infty f_p(\tau) \cos(\omega \tau) d\tau \]

\[ F_{S,p}(\omega) = \mathcal{F}^S[f_p(\tau)] = \int_0^\infty f_p(\tau) \sin(\omega \tau) d\tau \]

so that \( F_p(\omega) = F_{C,p}(\omega) - jF_{S,p}(\omega) \) \([34]\). In general by working with this basis we are not guaranteed that the resulting power spectral density of the covariance function will be real and symmetric. To
enforce this constraint the imaginary part of \( F_p(\omega)F_p^*(\omega) \) needs to be isolated. Through the Fourier sine and cosine transformations we have

\[
S(\omega) = (F_{C,p}(\omega) + jF_{S,p}(\omega))(F_{C,p}(\omega) + jF_{S,p}(\omega))^* \tag{54}
\]

\[
= F_{C,p}(\omega)F_{C,p}(\omega) + jF_{S,p}(\omega)F_{S,p}(\omega) - jF_{C,p}(\omega)F_{S,p}(\omega) + jF_{S,p}(\omega)F_{C,p}(\omega) \tag{55}
\]

\[
= \|F_{C,p}(\omega)\|^2 + \|F_{S,p}(\omega)\|^2 - jF_{C,p}(\omega)F_{S,p}(\omega) + jF_{S,p}(\omega)F_{C,p}(\omega) \tag{56}
\]

Now, \( \|F_{C,p}(\omega)\|^2 + \|F_{S,p}(\omega)\|^2 \) must be isolated. For analytic purposes, this will be more easily achieved by considering the following identity,

\[
2(\|F_{C,p}(\omega)\|^2 + \|F_{S,p}(\omega)\|^2) = (F_{C,p}(\omega) + jF_{S,p}(\omega))(F_{C,p}(\omega) + jF_{S,p}(\omega))^* \tag{57}
\]

+ \( (F_{C,p}(\omega) - jF_{S,p}(\omega))(F_{C,p}(\omega) - jF_{S,p}(\omega))^* \)

With a canonical filter given in the form of Eq. (53) we have that it’s Fourier cosine and sine transforms respectively are [35]

\[
F_p^C(\omega) = p! \left( \frac{\mu}{\mu^2 + \omega^2} \right)^2 \sum_{m=0}^{[0.5(p+1)]} (-1)^m \left( \frac{p+1}{2m} \right) \left( \frac{\omega}{\mu} \right)^{2m} \tag{58}
\]

\[
F_p^S(\omega) = p! \left( \frac{\mu}{\mu^2 + \omega^2} \right)^2 \sum_{m=0}^{[0.5p]} (-1)^m \left( \frac{p+1}{2m+1} \right) \left( \frac{\omega}{\mu} \right)^{2m+1} \tag{59}
\]

Thus for the subtractive term in Eq. (57),

\[
F_{C,p}(\omega) - jF_{S,p}(\omega) = p! \left( \frac{\mu}{\mu^2 + \omega^2} \right)^2 \sum_{m=0}^{p+1} (-j)^m \left( \frac{p+1}{m} \right) \left( \frac{\omega}{\mu} \right)^m \tag{60}
\]

\[
= p! \left( \frac{\mu}{\mu^2 + \omega^2} \right)^2 \left( 1 - j \frac{\omega}{\mu} \right)^{p+1} \tag{61}
\]

and similarly for the additive term in Eq. (57)

\[
F_{C,p}(\omega) + jF_{S,p}(\omega) = p! \left( \frac{\mu}{\mu^2 + \omega^2} \right)^2 \left( 1 + j \frac{\omega}{\mu} \right)^{p+1} \tag{62}
\]

Using these expansions Eq. (57) then becomes

\[
2(\|F_{C,p}(\omega)\|^2 + \|F_{S,p}(\omega)\|^2) = (p!)^2 \left[ \left( \frac{\mu}{\mu^2 + \omega^2} \right) \left( \frac{\mu}{\mu^2 + \omega^2} \right)^* \right]^{p+1}
\]

\[
\times \left[ \left( 1 - j \frac{\omega}{\mu} \right) \left( 1 - j \frac{\omega}{\mu} \right)^* \right]^{p+1} + \left[ \left( 1 + j \frac{\omega}{\mu} \right) \left( 1 + j \frac{\omega}{\mu} \right)^* \right]^{p+1} \tag{63}
\]

which upon some simplification we obtain

\[
S(\omega) = \frac{1}{2(k!)^2} \left[ \frac{1}{(a^2 + b^2)^{p+1}} (a^2 + (\omega - jb)^2)^{p+1} + (a^2 + (\omega + jb)^2)^{p+1} \right] \frac{1}{(a^2 + b^2)^{4} + (a^2 - b^2)\omega^2 + \omega^4}^{p+1} \tag{64}
\]

\[
= \frac{1}{2} \left( \frac{p!}{2} \right)^2 \left[ \frac{(a^2 + (\omega - b)^2)^{p+1} + (a^2 + (\omega + b)^2)^{p+1}}{(a^2 + b^2 - 2b\omega + \omega^2)(a^2 + b^2 + 2b\omega + \omega^2)^{p+1}} \right] \tag{65}
\]

\[
= \frac{1}{2} \left( \frac{p!}{2} \right)^2 \left[ \frac{(a^2 + (\omega - b)^2)^{p+1} + (a^2 + (\omega + b)^2)^{p+1}}{(a^2 + (\omega - b)^2)(a^2 + (\omega + b)^2)^{p+1}} \right] \tag{66}
\]

where partial fraction simplification was used in the last line [36]. Inspection shows that this PSD is very similar in terms of each summand to the PSD of the Matérn kernel. Substituting \( \zeta_1 = \omega + b \) and \( \zeta_2 = \omega - b \) the inverse Fourier transform is easily found.
C Proof of Lemma 5.1

Though much of this proof was outlined in the main text, here we discuss this lemma in more detail.

Lemma 5.1. If given a linear and finite-dimensional SDE, written as

$$dF^S(t) = F^S(t) dt + L dW(t)$$

(37)

where $F \in \mathbb{R}^{N \times N}$, $L \in \mathbb{R}^{N \times M}$, and $dW(t) \in \mathbb{R}^M$ is an $M$-dimensional Brownian motion process. If the solution, $F^S(t) \in \mathbb{R}^N$, then, it must be the case that Eq. (37) corresponds to a linear transformation of the SSM formulation of an $N$-ple GP in the Hida-Matérn family.

Proof. We begin with the observation that the solution of any linear SDE driven by Brownian motion is both Gaussian and Markov in the sense that $p(F^S(t)|F^S(s); s < t) = p(F^S(t)|F^S(s))$ (sometimes referred to as Markov in the simple sense) [18, 12].

For a linear SDE, the fundamental matrix solution, $\Phi(t, u) = \Phi(t - u) = \exp(F(t - u))$ tells us how the process propagates forward [26]. If we take the Jordan composition of $F = CJC^{-1}$; with $J$ a Jordan block matrix, and $C$ a non-singular matrix of the same dimension, then the matrix exponential can be rewritten so that $\exp(F\tau) = \exp(CJC^{-1}\tau) = C \exp(J\tau)C^{-1}$. For concreteness, let's also say that $J$ has an Jordan blocks, each of size $p_i$.

Take $g(t)$ to be a Hida-Matérn mixture kernel, with $m$ mixands of Markov order $p_i$, then its SSM formulation has transition matrix $A(\tau) = K^S(\tau)K^S(0)^{-1}$. In addition, $g^S(t)$ being a vector GP must have a representation in terms of a linear SDE, so for now we will say that SDE has transition matrix $G$; meaning $A(\tau) = K^S(\tau)K^S(0)^{-1} = \exp(G\tau)$. Since the kernel of $g(t)$ contains $m$ mixands of order $p_i$, $G$ will admit a Jordan decomposition $G = DJD^{-1}$ with the same Jordan block structure as $F$.

As there must exist some $X$ such that $XD = C$, lets take $F^S(t) = Xg^S(t)$. Then the transition matrix of $F^S(t)$ would be $XK^S(\tau)K^S(0)^{-1} = X \exp(G\tau)X^{-1} = XD \exp(J\tau)D^{-1}X^{-1} = C \exp(J\tau)C^{-1} = \exp(F\tau)$. Which is what we desired.

D Proof of Theorem 3.1

The class of Hida-Matérn kernels form a basis that can approximate any $L_2$ integrable kernel.

Theorem 3.1 (Stationary Hida-Matérn kernels forms a universal class). For any fixed value of $p$, Hida-Matérn kernels are dense in the space of square integrable stationary kernels, hence universal.

Proof. From Wiener’s Tauberian theorem, we have that if $S \in L^2(\mathbb{R})$ is square integrable then the span of the translations $S(\omega + b)$ is dense in $L^2(\mathbb{R})$ if and only if the real zeros of the Fourier transform of $S$ form a set of Lebesgue measure 0 [37].

Take a Hida-Matérn kernel, fix $a$ and $p$, then let $k_H(\tau; a, b)$ denote the PSD, or the Fourier transform of $k_H(\tau)$ as $S_H(\omega) = \sum_{i=1}^{L} c_i S_H(\omega; a, b)$. We can write, $k_H(\tau; a, b) = \exp(-j \tau \omega)k_H(\tau; 0, 0) + \exp(j \tau \omega)k_H(\tau; 0, 0)$. By the frequency shifting property of the Fourier transform we have that $S_H(\omega; a, b) = S_H(\omega - b; a, 0) + S_H(\omega + b; a, 0)$. Furthermore, since $S_H(\omega; a, 0)$ is symmetric about the origin, we have that $F[S_H(\omega; a, 0)] = F^{-1}[S_H(\omega; a, 0)]$. Recognizing that the second term results in the non-oscillatory Hida-Matérn kernel/Matérn kernel in the time domain now makes it clear that the Fourier transform of $S_H(\omega; a, 0)$ is strictly positive and so its real zeros have Lebesgue measure 0.

Now, using Wiener’s Tauberian theorem, we have that that the span of translations of $S_H(\omega; a, 0)$ is dense in $L^2(\mathbb{R})$. So, if we have some square integrable kernel $k(\tau)$ with Fourier transform $S(\omega)$, then from Wiener’s Tauberian theorem we should be able to find a linear combination of Hida-Matérns such that

$$\left( \int |S(\omega) - \sum_{i=1}^{L} c_i S_H(\omega; a, b) |^2 d\omega \right)^{1/2} \to 0$$

(67)
However, by now using Parseval’s theorem we also have that
\[ \int |S(\omega) - \sum_{i=1}^{L} c_i S_{H,p}(\omega; a, b)|^2 d\omega = \int |k(\tau) - \sum_{i=1}^{L} c_i k_{H,p}(\tau; a, b)|^2 d\tau \quad (68) \]
which also means that the class of Hida-Matérn kernels are dense in the space of \( L^2(\mathbb{R}) \).

**Remark.** We note a similar way of proving pointwise convergence for any stationary, real valued, positive semidefinite kernel modulated by a cosine was used in [38].

### E Kalman filtering for GP inference with Hida-Matérn kernels

**Algorithm 1:** Kalman filtering algorithm for GP regression with Hida-Matérn kernels

**input:** \( \{t_i, y_i\}_{i=1}^{L}, \quad K^S(\tau), \quad \sigma^2 \)

**begin**
\[ \mathcal{I} \leftarrow \text{where}(h == 0) \]
\[ P_0 \leftarrow P_\infty \]
\[ m_0 \sim \mathcal{N}(m_0 | 0, P_\infty) \]

Kalman Filtering

**for** \( i = 1, \ldots, L \) **do**
\[ \tau_i \leftarrow t_i - t_{i-1} \]
\[ A_i(\tau_i) \leftarrow K^S(\tau_i)K^S(0)^{-1} \]
\[ Q(\tau) \leftarrow K^S(0) - K^S(\tau)K^S(0)^{-1}K^S(\tau)^H \]
\[ P_i^- \leftarrow A_i(\tau_i)P_{i-1}A_i(\tau_i)^H + Q(\tau_i) \]
\[ m_i^- \leftarrow A_i m_{i-1} \]
\[ \alpha \leftarrow \left( \sum_{k,l \in \mathcal{I}} P_i^- [k, l] + \sigma^2 \right)^{-1} \]
\[ \beta \leftarrow \left( y_i - \sum_{k \in \mathcal{I}} P_i^- [:, k] \right) \]
\[ m_i \leftarrow m_i^- + \alpha \beta \sum_{k \in \mathcal{I}} P_i^- [:, k] \]
\[ P_i \leftarrow P_i^- - \alpha \sum_{k \in \mathcal{I}} P_i^- [:, k]P_i^- [:, k]^T \]

**end**

RTS Smoothing

**for** \( i = L - 1, \ldots, 1 \) **do**
\[ G \leftarrow P_i A_i^T [P_i^T]^{-1} \]
\[ m_i \leftarrow m_i + G (m_{i+1} - m_i^-) \]
\[ P_i \leftarrow P_i + G (P_{i+1} - P_i^-) G^T \]

**end**

**return** \( \{m_i, P_i\}_{i=1}^{L} \)

**end**

### F Low Rank Kalman Updates

Taking advantage of the sparsity of the observation extraction vector \( h \) can also aid in reducing numerical noise by recognizing that its sparsity leads to low rank Kalman updates of the predicted covariance and simplified equations for the updates of the mean.
**Kalman Equations** Take $m_k^-$ and $P_k^-$ to be the mean and covariance prediction at $t_k$ with $m_k$ and $P_k$ to be the updated mean and covariance and $t_k$. Then with $A(\Delta_k) = K^S(t_{k+1} - t_k)K^S(0)^{-1}$ and $Q(\Delta_k) = K^S(0) - K^S(t_{k+1} - t_k)K^S(0)^{-1}K^S(t_{k+1} - t_k)^T$ the Kalman recursions follow

$$
m_k^- = A(\Delta_k)m_{k-1}
$$
$$
P_k^- = Q(\Delta_k) + A(\Delta_k)P_k A(\Delta_k)^T
$$
$$
u_k = y_k - h^T m_k^-
$$
$$S_k = h^T P_k^- h + R
$$
$$K_k = P_k^- h S_k^{-1}
$$

$$
m_k = m_k^- + K_k \nu_k
$$
$$P_k = P_k^- - K_k h^T P_k^-
$$

$$= (I - K_k h^T)P_k^- (I - K_k h^T)^T + K_k R K_k^T$$ (Joseph form)

In general we would not use the standard covariance update because numerically it will not guarantee $P_k$ is PSD. With that said, let $Z$ be the set of indices where $h$ has non-zero elements and let’s first expand the equation for $m_k$

$$m_k = m_k^- + K_k \nu_k
$$

$$= m_k^- + P_k^- h S_k^{-1} \nu_k
$$

$$= m_k^- + P_k^- h (h^T P_k^- h + R)^{-1} \nu_k
$$

$$= m_k^- + P_k^- h (h^T P_k^- h + R)^{-1} (y_k - h^T m_k^-)
$$

$$= m_k^- + \sum_{i \in Z} P_k^- [:, i] \left( \sum_{i,j \in Z} P_k^- [i,j] + R \right)^{-1} \left( y_k - \sum_{i \in Z} m_k^- [i] \right)
$$

Now take $\alpha = (\sum_{i,j \in Z} P_k^- [i,j] + R)^{-1}$ and $\beta = (y_k - \sum_{i \in Z} m_k [i])$ and we get that

$$m_k = m_k^- + \alpha \beta \sum_{i \in Z} P_k^- [:, i]
$$

Making it obvious that it is sufficient to work with select columns and elements of $P_k^-$ and that the update is simply the sum of select scaled columns of $P_k$.

Let’s do the same for the updated covariance,

$$P_k = P_k^- - K_k h^T P_k^-
$$

$$= P_k^- - P_k^- h S_k^{-1} h^T P_k^-
$$

$$= P_k^- - P_k^- h (h^T P_k^- h + R)^{-1} h^T P_k^-
$$

$$= P_k^- - \alpha P_k^- h h^T P_k^-
$$

$$= P_k^- - \alpha \sum_{i \in Z} P_k^- [:, i] P_k^- [:, i]^T
$$

Now, it is obvious that the updated covariance $P_k$ consists of subtracting a rank($|Z|$) matrix from $P_k^-$ which is simply the sum of outer products select columns of $P_k^-$.

**G Naive mean-square processes**

Consider a situation where we would like to perform inference with the prior GP having kernel $k_{H,3}(r; a, 0)$ (the Matérn $\frac{3}{2}$) but, computational resources restrict us to 2-dimensional state space.
With that said, it does not mean that naive mean square processes are not worth any further examination. Where
As matrices partitioned so that
The question is whether or not this marginalized model can serve as a way to achieve approximate
is just the appropriate block of
Once
K
p
K
in continuous time – for it to be true that, differentiable GPs may admit some favorable characteristics or interesting behavior. Thus, before
Minimizing Stationary multioutput covariance
First, decompose the vector process so that
\[ p(z(t_{k+1})|z(t_k)) = \int p(z(t_{k+1})|z(t_k), g(t_k))p(g(t_k)|z(t_k))dg(t_k) \] (69)
\[ \Lambda(\Delta_k) = A_{00}(\Delta_k) + A_{01}(\Delta_k)K_{10}^S(0)K_{00}^S(0) \] (70)
\[ \Sigma(\Delta_k) = Q_{00}(\Delta_k) + A_{01}(\Delta_k)M A_{01}(\Delta_k)^T \] (72)
\[ M = (K_{11}^S(0) - K_{10}^S(0)K_{00}^S(0)^{-1}K_{10}^S(0)^T) \] (73)
with matrices partitioned so that \( \Xi = \begin{pmatrix} \Xi_{00} & \Xi_{01} \\ \Xi_{10} & \Xi_{11} \end{pmatrix} \) with \( \Xi_{00} \in \mathbb{R}^{M \times M}, \Xi_{11} \in \mathbb{R}^{N-M \times N-M} \), and \( \Delta_k = t_{k+1} - t_k \). Similar to the unmarginnalized process we recover the following SSM describing the vector process \( z(t) \)
\[ z(t_{k+1}) = \Lambda(\Delta_k)z(t_k) + q(t_k) \] (74)
\[ q(t_k) \sim \mathcal{N}(q(t_k)|0, \Sigma(\Delta_k)) \] (75)
As \( z(t) \) is \( N-M \) times differentiable by construction, we can see why such a process can not exist in continuous time – for it to be true that, \( p(z(t)|\sigma(z(s)); s < t) = p(z(t)|z(s)), \) \( z(t) \) can not be differentiable. However, it is perfectly valid to construct an SSM such as Eq. (74) in discrete time. The question is whether or not this marginalized model can serve as a way to achieve approximate inference faithfully with respect to the prior GP in mind.
It turns out that performing the described one step marginalization of higher order derivatives does not aid in propagating their uncertainty. Indeed, as is shown in Section J, the marginalized transition matrix in Eq. (71) and the covariance of the noise term in Eq. (72) only depend on \( K_{00}^S(\tau) \), or the block of the multioutput covariance associated with \( z(t) \). More concretely,
\[ \Lambda(\Delta_k) = K_{00}^S(\tau)K_{00}^S(0)^{-1} \] (76)
\[ \Sigma(\Delta_k) = K_{00}(0) - K_{00}(\Delta_k)K_{00}^S(0)^{-1}K_{00}(\Delta_k)^T \] (77)
G.2 Approximating arbitrary priors with naive mean square processes
With that said, it does not mean that naive mean square processes are not worth any further examination. We can posit that the family of naive mean square processes when considering arbitrary differentiable GPs may admit some favorable characteristics or interesting behavior. Thus, before ruling out naive mean square processes as viable alternatives or priors for inference, let us examine them to approximate some prior GP of interest.
First though, we must ask – what is a reasonable objective function to minimize so that the hyperparameters of these naive mean square processes can be set to best approximate the desired prior.
Minimizing Stationary multioutput covariance: One choice is choosing the hyperparameters such that the up-projected stationary covariance of the low dimensional model try to most closely match the exact the true stationary covariance. Let \( K^S(0; \theta) \) be the stationary covariance of the SSM formed by \( p(F^S(t)|F^S(s); \theta) \) and \( K_{1L}^S(0; \nu) \) be the stationary covariance of the SSM formed by \( p(z(t)|z(s); \nu) \), where \( \theta \) is fixed and \( \nu \) is free to optimize.
Once \( K_{1L}^S(0; \nu) \) is found we can posit how to project it to the same space as \( K^S(0; \theta) \). As \( K_{1L}^S(0; \nu) \) is just the appropriate block of \( K^S(0; \theta) \), the estimator of \( K^S(0; \theta) \) from \( K_{1L}^S(0; \nu) \) involves only
linear transformations found through manipulations of Gaussians so that the hyperparameters, $\nu$, can be found as

$$
\nu = \arg\min_{\nu} \left\| \mathbf{K}^S(0; \theta) - \left( \mathbf{K}^S_L(0; \nu) \mathbf{C} \right) \right\|^2
$$

(78)

Where $\mathbf{C} = \mathbf{K}^S_{10}(0; \nu) \mathbf{K}^S_{00}(0; \nu)^{-1}$, so that $\mathbf{C} \in \mathbb{R}^{M \times N-M}$. Since $\mathbf{K}^S(0; \theta)$ contains all information about the moments of the PSD, $S(\omega)$, we may think of the objective as a moment matching problem with respect to $S(\omega)$.

**Minimizing the $L_2$ distance between the true kernel and approximate kernel:** Another alternative for selecting the hyperparameters is to choose a bin size and then minimize the $L_2$ distance between the approximate kernel and the desired one. Take a partition over the interval $[0, T]$, with $0 = t_0 < t_1 < \cdots < t_T = T$, and $T$ large enough such that the kernel of interest has decayed sufficiently close to 0. Then with, $\Sigma(t_i) \triangleq \text{Cov}(\mathbf{z}(t_i), \mathbf{z}(t_{i-1}))$, we have that $\Sigma(t_i) = \mathbf{K}^S_{00}(0) \prod_{i=1}^{T} \Lambda(t_i - t_{i-1})$, so that the optimal hyperparameters with respect to the $L_2$ norm between the approximate kernel and true kernel should satisfy

$$
\nu = \arg\min_{\nu} \sum_{i} \left( \mathbf{k}(t_i; \theta) - \mathbf{h}^\top \Sigma(t_i; \nu) \mathbf{h} \right)^2
$$

(79)

### G.3 Example Approximations

To examine what the resulting inference using a low dimensional naive mean square process, let’s first consider the Matérn $\frac{7}{2}$, which admits exact inference for an SSM that is 4-dimensional.

![Figure 9: 2nd, 3rd, and 4th order models of the Matérn $\frac{7}{2}$. Column A): Posterior fit over noisy observations of a sample from the true kernel when the stationary objective is used. Column B): Posterior fit over noisy observations of a sample from the true kernel when the $L_2$ objective is used. Column C): The kernel under the inferred hyperparameters for both objectives for a spacing of 0.01. Column D): The PSD found from taking the Fourier transform of the autocorrelation. Exact inference can be seen in the 4th order model fit](image)

As can be seen in Fig. 9 as the order of the naive SSM increases, the approximate covariance becomes more like the true covariance function. In addition, the stationary objective and $L_2$ objective result in different sets of hyperparameters that minimize the associated cost functions.

Now, we can consider something infinitely differentiable so let’s take the familiar squared exponential covariance function. For this example, under the stationary covariance objective we use up to the 15th order approximate SSM and consider the described stationary and $L_2$ objectives as well as leaving the hyperparameters true to the GP prior.
Considering another infinitely differentiable covariance function, we examine the posterior fits and approximate discrete time kernel, with the reference being the sinc kernel [30].

Figure 10: The squared exponential. Column A 1/2): Posterior fits with the SSM constructed using the original hyperparameters , Column B 1/2): Posterior fits constructed using the stationary objective to set hyperparameters, Column C 1/2): Posterior fits constructed using the L2 objective to set hyperparameters .Column D 1/2): Autocorrelation functions under each different setting of the hyperparameters as well as the true autocorrelation.

Figure 11: 4th, 5th, and 6th order approximations to the sinc kernel. Column A): Posterior fit under the stationary objective, Column B): Posterior fit under the L2 objective, Column C): Autocorrelation of the discrete process,Column D): PSD corresponding to said autocorrelation.
H Stationary Covariance of an \( N \)-ple Markov GP

Though the fact that the stationary covariance of the continuous time representation of a GP as in Eq. (23) being \( K^S(0) \) is intuitive, it is also true that the stationary covariance of the discretized model as presented for standard GP regression is \( K^S(0) \) — independent of the spacing of the observations.

**Proposition 2.** For state space models as defined, the stationary marginal covariance of the process, \( f^S(t) \), denoted \( P_\infty \) is exactly \( K^S(0) \).

**Proof.** The stationary marginal covariance satisfies the discrete Lyapunov equation,

\[
A(\Delta)P_\infty A(\Delta)^T + Q(\Delta) - P_\infty = 0
\]

By substituting \( K^S(\Delta)K^S(0)^{-1} \) for \( A(\Delta) \) and expanding \( Q(\Delta) \) we get that

\[
K^S(\Delta)K^S(0)^{-1}P_\infty K^S(0)^{-1}K^S(\Delta)^T + K^S(0) - K^S(\Delta)K^S(0)^{-1}K^S(\Delta)^T - P_\infty = 0
\]

which can be factored as

\[
K^S(\Delta)(K^S(0)^{-1}P_\infty K^S(0)^{-1} - K^S(0)^{-1})K^S(\Delta)^T + K^S(0) - P_\infty = 0
\]

Letting \( Y = K^S(0)^{-1}P_\infty K^S(0)^{-1} - K^S(0)^{-1} \) gives us that \( P_\infty = K^S(0)(Y + K^S(0)^{-1})K^S(0) \)

upon whose substitution we find

\[
K^S(\Delta)YK^S(\Delta)^T + K^S(0) - K^S(0)YK^S(0) - K^S(0) = 0
\]

\[
K^S(\Delta)YK^S(\Delta)^T - K^S(0)YK^S(0) = 0
\]

Vectorizing Eq. 84 we now find that

\[
(K^S(\Delta) \otimes K^S(\Delta)^T - K^S(0) \otimes K^S(0)) \vec{Y} = 0
\]

which means that since a unique solution of \( \vec{Y} \) must exist, then the only possibility is that \( \vec{Y} = 0 \) since its premultiplier is of full rank and has no nullspace. This then means that

\[
\vec{Y} = 0
\]

\[
K^S(0)^{-1}P_\infty K^S(0)^{-1} - K^S(0)^{-1} = 0
\]

giving us the result

\[
P_\infty = K^S(0)
\]

**Remark.** We see that the stationary covariance is invariant to the choice of \( \Delta \), as such we could have used the continuous or discrete lyapunov equations to solve for the stationary covariance. Indeed, plugging in this solution to the continuous time Lyapunov equation is consistent.

\[\square\]

I Properties of \( K^S(\tau) \)

The multioutput kernel associated with an \( N \)-ple GP in the restricted sense has many nice properties. Some of these allow for increased numerical stability, which is especially important when using filtering algorithms for inference as ill conditioned SSMs can lead to poor inference [39, 40]. We now detail some of them that were mentioned but glossed over in the main text, and try to discuss any relevant connections to filtering or other related areas.

I.1 Unique Elements of \( K^S(\tau) \)

First, we note the number of unique elements in \( K^S(\tau) \) is \( 2N - 1 \). This is one of the more easily seen properties as the \( i \)th off diagonal of \( K^S(\tau) \) corresponds to the \( i \)th derivative of \( k(\tau) \) up to a sign change.
I.2 Unique Elements of $K^S(0)$

In the case that $\tau = 0$ and $b = 0$ the multioutput kernel evaluated at 0 should have only $N$ unique elements. To see this recall that $k^{q,r}(0) = (-1)^r (-j)^{q+r} \int \omega^{q+r} S(\omega) d\omega = (-j)^{q+r} \int \omega^{q+r} S(\omega) d\omega$. Now, since $S(\omega)$ is real and symmetric, its odd order moments are 0, therefore, only the entries of $K^S(0)$ with $q + r$ even will be non-zero.

Still considering the case that $b = 0$ what is nice about this property is that when $N$ is even $N^2/2$ elements will be 0, and when $N$ is odd then $(N^2 - 1)/2$ elements will be 0 – in either case this can be leveraged for faster and more stable inverses. Due to this sparsity, through elementary row and column operations, whether $N$ is even or odd, $K^S(0)$ can be transformed into a block matrix and each block may be inverted separately. Applying the inverse elementary row/column swapping matrices will then yield $K^S(0)^{-1}$.

J Marginalization of the naive model does not propagate uncertainty

Proposition 3. For the one step marginalization of the naive mean square process, we have that the state transition matrix and noise covariance are respectively

$$A^M(\Delta) = K^S_{00}(\Delta)K^S_{00}(0)^{-1}$$

$$Q^M(\Delta) = K^S_{00}(0) - K^S_{00}(\Delta)K^S_{00}(0)^{-1}K^S_{00}(\Delta)^T$$

Proof. We first expand the marginalized system in terms of its constituent blocks. For the transition matrix we have,

$$A^M(\Delta) = A_{00}(\Delta) + A_{01}(\Delta)K^S_{01}(0)^T K^S_{00}(0)^{-1}$$

and for the state noise we have that

$$Q^M(\Delta) = Q_{00}(\Delta) + A_{01}(\Delta) \left( K^S_{11}(0) - K^S_{01}(0)^T K^S_{00}(0)^{-1}K^S_{01}(0) \right) A_{01}(\Delta)^T$$

Using the matrix inversion lemma for block matrices we first find $A_{00}(\Delta)$, $A_{01}(\Delta)$, and $Q_{00}(\Delta)$.

To declutter their expansions let

$$B = \left( K^S_{11}(0) - K^S_{01}(0)^T K^S_{00}(0)^{-1}K^S_{01}(0) \right)^{-1}$$

$$C = K^S_{00}(0)^{-1}K^S_{01}(0)$$

and write,

$$A_{00}(\Delta) = K^S_{00}(\Delta)K^S_{00}(0)^{-1} + K^S_{00}(\Delta)CB^T - K^S_{01}(\Delta)BC^T$$

$$A_{01}(\Delta) = K^S_{01}(\Delta)B - K^S_{00}(\Delta)CB$$

$$Q_{00}(\Delta) = K^S_{00}(0) - K^S_{00}(\Delta)K^S_{00}(0)^{-1}K^S_{00}(\Delta)^T - K^S_{00}(\Delta)CB^T K^S_{00}(\Delta)^T + K^S_{01}(\Delta)BC^T K^S_{00}(\Delta)^T + K^S_{00}(\Delta)CK^S_{01}(\Delta)^T - K^S_{01}(\Delta)BK^S_{01}(\Delta)^T$$

Plugging in first for $A^M(\Delta) = A_{00}(\Delta) + A_{01}(\Delta)C^T$
\[ A^M(\Delta) = K_{00}^S(\Delta)K_{00}^S(0)^{-1} + K_{00}^S(\Delta)BC^T - K_{01}^S(\Delta)BC^T \]
\[ + K_{01}^S(\Delta)BC^T \]
\[ - K_{00}^S(\Delta)CBC^T \]

With most the terms cancelling out, we arrive to the satisfying conclusion that

\[ A^M(\Delta) = K_{00}^S(\Delta)K_{00}^S(0)^{-1} \]  \hspace{1cm} (97)

Now, we expand \( Q^M(\Delta) = Q_{00}(\Delta) + A_{01}(\Delta)B^{-1}A_{01}(\Delta)^T \) first by examining the second term,

\[ A_{01}(\Delta)B^{-1}A_{01}(\Delta)^T = (K_{01}^S(\Delta)B - K_{00}^S(\Delta)CB)B^{-1}(BK_{01}^S(\Delta)^T - BC^TK_{00}^S(\Delta)^T) \]
\[ = K_{01}^S(\Delta)BK_{01}^S(\Delta)^T \]
\[ - K_{00}^S(\Delta)BC^TK_{00}^S(\Delta)^T \]
\[ - K_{00}^S(\Delta)CBK_{01}^S(\Delta)^T \]
\[ + K_{00}^S(\Delta)CBC^TK_{00}^S(\Delta)^T \]

Quick inspection shows when added to \( Q^M(\Delta) \) these terms will be cancelled out and finally we get that

\[ Q^M(\Delta) = K_{00}^S(0) - K_{00}^S(\Delta)K_{00}^S(0)^{-1}K_{00}^S(\Delta)^T \]  \hspace{1cm} (98)

\[ \square \]