Sparse Algorithms for Markovian Gaussian Processes

William J. Wilkinson
william.wilkinson@aalto.fi
Aalto University
The Alan Turing Institute

Arno Solin
arno.solin@aalto.fi
Aalto University

Vincent Adam
vincent.adam@secondmind.ai
Secondmind.ai

Abstract

Approximate Bayesian inference methods that scale to very large datasets are crucial in leveraging probabilistic models for real-world time series. Sparse Markovian Gaussian processes combine the use of inducing variables with efficient Kalman filter-like recursions, resulting in algorithms whose computational and memory requirements scale linearly in the number of inducing points, whilst also enabling parallel parameter updates and stochastic optimisation. Under this paradigm, we derive a general *site-based* approach to approximate inference, whereby we approximate the non-Gaussian likelihood with local Gaussian terms, called sites. Our approach results in a suite of novel sparse extensions to algorithms from both the machine learning and signal processing literature, including variational inference, expectation propagation, and the classical nonlinear Kalman smoothers. The derived methods are suited to large time series, and we also demonstrate their applicability to spatio-temporal data, where the model has separate inducing points in both time and space.

1 INTRODUCTION

Gaussian processes (GPs, Rasmussen and Williams, 2006) are distributions over functions, commonly used in probabilistic machine learning to endow latent functions in generative models with rich and interpretable priors. These priors provide strong inductive biases for regression tasks in the small data regime.

GPs with uni-dimensional input are especially well-

(a) $\longrightarrow$ Forward *filtering* pass

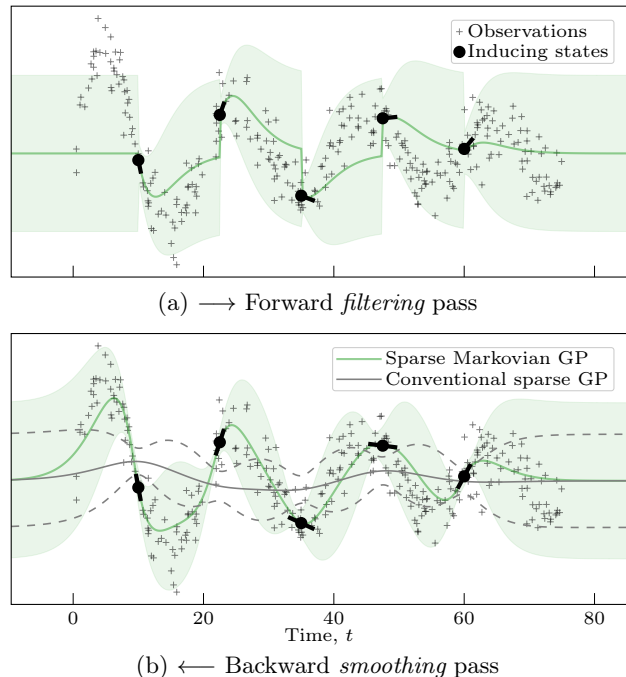(b) $\longleftarrow$ Backward *smoothing* pass

Figure 1: Inducing states with Markovian GPs: filtering **(a)** and smoothing (*i.e.*, posterior) distribution **(b)** of site-based sparse Markovian GP regression. The kernel is Matérn-5/2, so inducing states ($\blacktriangleright$) contain higher-order derivative information, making them more descriptive than conventional inducing function evaluations with the same number of inducing points. To achieve the same approximation accuracy, the conventional method would require the use of more function evaluations as inducing variables.

suited to modeling time series and spatio-temporal data. In this setting, the versatile class of *Markovian GPs* provides great computational advantages. These are GPs that can be rewritten in a stochastic differential equation (SDE) form (Särkkä and Solin, 2019) and, when marginalized to a discrete set of N ordered input locations, induce a sparse precision structure which enables efficient inference algorithms with linear time computational complexity, $\mathcal{O}(N)$, as opposed to the classic cubic time scaling, $\mathcal{O}(N^3)$, usually associated

with such models.

Sparse GPs (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006) are an alternative method for dealing with the computational intractability of GPs for large data sets, which exploit redundancy in the N data points to summarise the underlying function via a smaller set of M inducing points. This approach typically leads to inference algorithms with computational complexity $\mathcal{O}(NM^2 + M^3)$ (Hensman et al., 2013). Whilst conventional sparse GPs have been successfully applied in many domains, they are not naturally suited to time series, since the number of inducing points must grow in line with the number of time steps in order to describe all the variation in the data, which is prohibitive since their computational scaling is cubic in M.

For both of these schemes, methods have been devised to tackle intractable inference when using non-Gaussian likelihoods, making GPs applicable to large data sets where they provide principled uncertainty quantification and out-of-sample generalisation. These include variational inference (VI, Titsias, 2009; Hensman et al., 2013; Durrande et al., 2019) and expectation propagation (EP, Bui et al., 2017; Wilkinson et al., 2020).

Recently, sparsity and Markovianity have been combined under the variational inference framework to exploit the benefits of both approaches in a method named *doubly sparse variational GPs* (S²VGP, Adam et al., 2020), where the term *doubly sparse* comes from the fact that the method exploits the *sparse* precision matrix of states of a Markovian GP marginalized to a finite set of *sparse* inducing time points. This approach scales linearly in $(M + N)$, with M sequential computations and N independent ones, making it well suited to time series (see Fig. 1 for an illustration of the approach).

We generalise the doubly sparse approach by deriving *site-based* approximate inference algorithms for sparse Markovian GPs. These methods contrast the existing S²VGP method by parametrising the global approximate posterior via a set of *local* contributions from the data, in the same vein as EP. We derive four novel approximate inference algorithms based on this approach. These amount to doubly sparse extensions to conjugate-computation variational inference (CVI, Khan and Lin, 2017), power-expectation propagation (PEP, Minka, 2004), posterior linearisation (PL, García-Fernández et al., 2016), and the Extended Kalman smoother (EKS, Bell, 1994). We present these algorithms alongside the existing S²VGP approach, providing an overview of methods for inference in non-conjugate GP time series.

In our site-based algorithms, we leverage the idea of *site tying* (Li et al., 2015) in a principled way to reduce

the storage requirement of the algorithm from $\mathcal{O}(N)$ to $\mathcal{O}(Md^2)$, where d is the dimensionality of the state space representation of the Markovian GP, which is extremely efficient when $M \ll N$. We examine the properties of our new algorithms, and make detailed comparisons on multiple large time series data sets. In addition, we show how the methods can be applied to spatio-temporal data, where spatial inducing points are tracked over time by temporal processes which are summarised via a reduced number of time steps.

An efficient JAX (Bradbury et al., 2018) implementation of all site-based methods is provided at `https://github.com/AaltoML/Newt`.

2 BACKGROUND

Gaussian processes describe distributions over functions by stating that function evaluations at any finite collection of inputs are jointly Gaussian distributed. Given data comprising input-output pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N \in (\mathcal{X}, \mathbb{R})^N$, they are characterised completely by mean function $\mu(\mathbf{x})$ and covariance function $\kappa(\mathbf{x}, \mathbf{x}')$. A GP prior over a function, f, and the corresponding likelihood model for the observations, y_n, are written,

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')), \;\; \mathbf{y} \,|\, \mathbf{f} \sim \prod_{n=1}^N p(y_n \,|\, f_n), \quad (1)$$

where $\mathbf{f} = f(\mathbf{X})$ and $f_n = f(\mathbf{x}_n)$. If the likelihood is Gaussian, then the posterior, $p(\mathbf{f} \,|\, \mathbf{y})$, can be computed in closed form, at a computational cost $\mathcal{O}(N^3)$. However, we are interested in the general, non-conjugate, case which necessitates approximate inference methods.

2.1 Markovian Gaussian Processes

Markovian Gaussian processes are GPs with one-dimensional inputs, $x \in \mathbb{R}$, that have an equivalent linear time invariant (LTI) stochastic differential equation (SDE) representation with state dimension d:

$$\dot{\mathbf{s}}(x) = \mathbf{F}\mathbf{s}(x) + \mathbf{L}\boldsymbol{\varepsilon}(x), \qquad f(x) = \mathbf{H}\mathbf{s}(x), \qquad (2)$$

where $\boldsymbol{\varepsilon}(x) \in \mathbb{R}^e$ is a white noise process and $\mathbf{F} \in \mathbb{R}^{d \times d}$, $\mathbf{L} \in \mathbb{R}^{d \times e}$, $\mathbf{H} \in \mathbb{R}^{1 \times d}$ are the feedback, noise effect, and emission matrices. This representation supports linear time inference algorithms by explicitly performing inference over the larger set of random variables that constitute the discrete state space trajectory, $\mathbf{s} = \mathbf{s}(\mathbf{x})$, indexed at $\mathbf{x} = [x_1, \ldots, x_N] \in \mathbb{R}^N$. The majority of commonly used GP kernels (on one-dimensional inputs) admit the above form (Särkkä and Solin, 2019).

The solution to this LTI-SDE evaluated at $\mathbf{x}$ follows a

discrete-time linear system:

$$\mathbf{s}(x_{n+1}) = \mathbf{A}_{n,n+1}\mathbf{s}(x_n) + \mathbf{q}_n, \quad \mathbf{q}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n,n+1}),$$

$$\mathbf{s}(x_0) \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_0), \qquad f_n = \mathbf{H}\mathbf{s}(x_n), \qquad (3)$$

where state transitions, $\mathbf{A}_{n,n+1} \in \mathbb{R}^{d \times d}$, noise co-variances, $\mathbf{Q}_{n,n+1} \in \mathbb{R}^{d \times d}$, and stationary state co-variances, $\mathbf{P}_0 \in \mathbb{R}^{d \times d}$, can be computed analytically (see App. A.1). In the conjugate case, when $p(y_n \,|\, f_n = \mathbf{H}\mathbf{s}(x_n))$ is Gaussian, the posterior is a GP. Its marginal statistics and the marginal likelihood, $p(\mathbf{y})$ (used for optimising the model parameters), are available in closed form and can be computed efficiently using Kalman recursions with computational scaling $\mathcal{O}(Nd^3)$. For non-Gaussian likelihoods we resort to approximate inference, and various schemes have been proposed (Nickisch et al., 2018; Durrande et al., 2019; Wilkinson et al., 2020; Chang et al., 2020).

2.2 Sparse Gaussian Process Approximations

Sparse GPs are one of the most successful solutions to handling scalability issues and have allowed GPs to be applied to large data sets (see Bui et al., 2017, for a review). The *true* posterior process can be expressed as $p(f(\cdot)\,|\,\mathbf{y}) = \int p(f(\cdot), \mathbf{f}\,|\,\mathbf{y})\,\mathrm{d}\mathbf{f} = \int p(f(\cdot)\,|\,\mathbf{f})\,p(\mathbf{f}\,|\,\mathbf{y})\,\mathrm{d}\mathbf{f}$. Here we use $f(\cdot)$ to denote all possible function evaluations (including $\mathbf{u}$ and $\mathbf{f}$). This expression captures the information flow from the data $\mathbf{y}$ through function evaluations $\mathbf{f} = f(\mathbf{X})$. Sparse approximations build an *approximate* posterior process of the form,

$$q(f(\cdot)) = \int p(f(\cdot)\,|\,f(\mathbf{z}) = \mathbf{u})\,q(\mathbf{u})\,\mathrm{d}\mathbf{u}, \qquad (4)$$

where $\mathbf{z} \in \mathcal{X}^M$ are referred to as *pseudo-inputs*, and $q(\mathbf{u})$ can be interpreted as an approximate posterior on $\mathbf{u} = f(\mathbf{z}) \in \mathbb{R}^M$, i.e. $q(\mathbf{u}) \approx p(\mathbf{u}\,|\,\mathbf{y})$. This approach can be extended by conditioning the process using deterministic functions of the process $\mathbf{u} = \phi(f)$ (*e.g.*, Dutordoir et al., 2020; Hensman et al., 2018). Such approaches are referred to as *inter-domain*, and the search for good inter-domain features is driven by competing demands for good approximation accuracy, tractability and scalability. This paper is based on a particular inter-domain formulation dedicated to Markovian GPs whereby the inducing variables are inducing states (Adam et al., 2020, see Sec. 3). Given a choice of inducing variables, there are two main approaches to parametrizing the approximate posterior: a global approximation, and a local one.

Global approximate posterior One way to construct the approximate posterior is to choose $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{L}\mathbf{L}^\top)$ to be a free-form multivariate Gaussian, whose mean $\mathbf{m} \in \mathbb{R}^M$ and Cholesky factor of the covariance $\mathbf{L} \in \mathbb{R}^{M \times M}$ are optimised with respect to some objective.

Site-based approximate posterior An alternative, site-based approach to inference utilises the theoretical optimal form of the approximate posterior (Opper and Archambeau, 2009; Bui et al., 2017):

$$q(\mathbf{u}) \propto p(\mathbf{u}) \prod_n t_n(\mathbf{u}), \qquad (5)$$

i.e., we assume that it factorises as a product of the prior and the (possibly unnormalized) Gaussian sites parameterized in the natural form, $t_n(\mathbf{u}) = \tilde{\mathcal{N}}(\mathbf{u}; z_n, \boldsymbol{\lambda}_{1,n}, \boldsymbol{\lambda}_{2,n}) = z_n \exp(\mathbf{u}^\top \boldsymbol{\lambda}_{1,n} - 1/2\mathbf{u}\boldsymbol{\lambda}_{2,n}\mathbf{u}^\top)$, with $\boldsymbol{\lambda}_{1,n} \in \mathbb{R}^M$ and $\boldsymbol{\lambda}_{2,n} \in \mathbb{R}^{M \times M}$. These can be thought of as pseudo likelihood terms that describe the effect of the data on the posterior. This leads $q(\mathbf{u})$ to be Gaussian and its statistics can be computed in closed form. In the algorithms we describe, the optimal sites can be shown to be rank one, *i.e.*, $t_n(\mathbf{u}) = \tilde{\mathcal{N}}(\mathbf{W}_n\mathbf{u}; z_n, \lambda_{1,n}, \lambda_{2,n})$, where $\mathbf{W}_n$ is the conditional projection matrix: $\mathbb{E}_p[f_n\,|\,\mathbf{u}] = \mathbf{W}_n\mathbf{u}$. The sites can be updated either via gradient-based methods or via iterative deterministic algorithms.

2.3 Variational Inference

Global VI (SVGP) In the global variational approach to sparse GP inference (VI, Titsias, 2009), one attempts to directly learn $q(\mathbf{u})$ by minimizing the KL divergence between the approximate posterior $q(f)$, Eq. (4), and the true posterior, $\boldsymbol{\Delta} = \mathrm{KL}[q(f)\,\|\,p(f\,|\,\mathbf{y})]$, or equivalently by maximizing the variational objective, also called the evidence lower bound (ELBO),

$$\mathcal{L}(q) = \mathbb{E}_q \log p(\mathbf{y}\,|\,\mathbf{f}) - \mathrm{KL}[q(\mathbf{u})\,\|\,p(\mathbf{u})], \qquad (6)$$

which verifies $\log p(\mathbf{y}) - \mathcal{L}(q) = \boldsymbol{\Delta}$. This ELBO can used both for inference and learning. Evaluation of the KL divergence and of the expected log-likelihood terms, also called variational expectations, have respective computational costs of $\mathcal{O}(M^3)$ and $\mathcal{O}(NM^2)$, leading to overall computational complexity $\mathcal{O}(M^3 + NM^2)$ (Hensman et al., 2013).

Local VI (CVI) Conjugate-compuation variational inference (CVI, Khan and Lin, 2017) uses a mirror descent algorithm to derive a site-based algorithm that is equivalent to performing VI with natural-gradients (Salimbeni et al., 2018). To the best of our knowledge, CVI has not yet been applied to sparse GPs. To do so, the generative model must be split into a conjugate part (*i.e.*, the prior $p(\mathbf{u})$), and a non-conjugate part which gathers the remaining terms of the likelihood $p(\mathbf{y}\,|\,\mathbf{f})$ and the conditional prior $p(\mathbf{f}\,|\,\mathbf{u})$:

$$p(\mathbf{f}, \mathbf{u}, \mathbf{y}) = \underbrace{p(\mathbf{u})}_{p_c(\mathbf{u})} \underbrace{p(\mathbf{f}\,|\,\mathbf{u})\,p(\mathbf{y}\,|\,\mathbf{f})}_{p_{nc}(\mathbf{f},\mathbf{u})}. \qquad (7)$$

CVI approximates the non-conjugate part using Gaussian sites with the sufficient statistics of $p(\mathbf{u})$:

$\tilde{p}_{nc}(\mathbf{f}, \mathbf{u}) \approx p(\mathbf{f} \mid \mathbf{u}) t(\mathbf{u})$, where $t(\mathbf{u}) = \prod_{n=1}^{N} t_n(\mathbf{u})$, which turns out to be the same parametrisation as used in EP.

Letting $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}$ be the natural parameters of the prior $p(\mathbf{u})$ and sites $t(\mathbf{u})$ respectively, the natural parameters of $q(\mathbf{u})$ are $\boldsymbol{\Lambda} + \boldsymbol{\lambda}$. One can show that a natural gradient step on the variational parameters $\boldsymbol{\lambda}$ amounts to:

$$\mathbf{g} = \nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\mathbf{u})} \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})} \log p(\mathbf{y} \mid \mathbf{f}),$$
$$\boldsymbol{\lambda}^{(k+1)} = (1 - \rho) \boldsymbol{\lambda}^{(k)} + \rho \, \mathbf{g}, \tag{8}$$

where $\boldsymbol{\mu}$ are the expectation parameters of the posterior $q(\mathbf{u})$, k is the training iteration, and ρ is the step size. It should be noted that CVI is equivalent to SVGP with natural gradients as in Salimbeni et al. (2018). A natural gradient step in the SVGP approach requires switching between the natural and moment parameterisations of the global variational distribution $q(\mathbf{u})$ (and the gradients of these operations). This is more computationally costly and prone to numerical errors than the CVI derivation.

2.4 Expectation Propagation

The sparse variant of expectation propagation (EP, Minka, 2001; Bui et al., 2017) also uses a site-based approach, with posteriors $q(f)$ and $q(\mathbf{u})$ defined as in Eqs. (4) and (5) respectively. The EP algorithm aims to *globally* minimise the forward KL divergence, $\mathrm{KL}[p(f \mid \mathbf{y}) \,\|\, q(f)]$, but since this is intractable it instead updates each site separately in an iterative fashion by minimising *local* KL divergences, $t_n^{\mathrm{new}}(\mathbf{u}) =$

$$\underset{t_n^*(\mathbf{u})}{\arg \min} \, \overline{\mathrm{KL}} \left[q(f(\cdot), \mathbf{u}) \frac{p(y_n \mid f_n)}{t_n^{\mathrm{old}}(\mathbf{u})} \,\Big\|\, q(f(\cdot), \mathbf{u}) \frac{t_n^*(\mathbf{u})}{t_n^{\mathrm{old}}(\mathbf{u})} \right], \tag{9}$$

where $\overline{\mathrm{KL}}$ represents the KL divergence for unnormalised distributions. This is equivalent to matching the first two moments between the approximate joint (right) and the approximate joint in which one site is replaced with the true likelihood term (left). In other words, the local site $t_n(\mathbf{u})$ is optimized in the context of the leave-one-site-out posterior, $q(f(\cdot), \mathbf{u})/t_n^{\mathrm{old}}(\mathbf{u})$. Power expectation propagation (PEP, Minka, 2004) is a generalisation of EP that minimises the α-divergence, $\mathrm{D}_\alpha[p(f \mid \mathbf{y}) \,\|\, q(f)]$, usually implemented by raising the likelihood and site terms in Eq. (9) to a power of α.

2.5 Global, Local, and Intermediate Approximations

Minka (2001) showed that PEP corresponds to variational algorithms in the limit of $\alpha \to 0$. That is, for $\alpha \to 0$, if PEP converges, then it converges to the same optima as that given by optimising Eq. (6). This

result extends to the corresponding sparse VI and PEP algorithms (Bui et al., 2017).

In order to reduce the memory requirements associated with storing all the EP parameters, *tied* sites were introduced in an algorithm called stochastic expectation propagation (SEP, Li et al., 2015). In the most extreme instantiation of SEP, each of the N sites are set to correspond to a fraction of a global site $t_n(\mathbf{u}) = t(\mathbf{u})^{1/N}$. Intermediate algorithms are also possible in which subsets of data points are tied together. These algorithms, also referred to as average EP (Dehaene and Barthelmé, 2018), lead to an approximation whose memory requirement no longer scales with the number of data points.

In the opposite direction, efforts aimed at speeding up computation of the VI approximation have led to a localized (or de-globablized) variational posterior. For example, additional conditional independence assumptions between subsets of observations and subsets of the latent process have been proposed (Bui and Turner, 2014), leading to factors of the variational distribution impacting the posterior distribution locally.

2.6 Comparison and Performance Guarantees

Overall, sparse power EP and VI approaches are efficient and performant. The most recent survey and comparison of these methods (Bui et al., 2017) reports an overall slight advantage for EP in non-conjugate tasks. However, the VI algorithm is simpler, very modular, and has formed the basis of more extensions in the research community.

Many of the algorithms presented above are specific instances of broader classes of algorithms, and therefore inherit some general guarantees in terms of convergence or approximation error: *(i)* in the variational setting, sparse GPs come with guarantees on the quality of the posterior approximation as the number of inducing point is increased (Burt et al., 2019), *(ii)* iterative updates of CVI algorithms will increase the ELBO and converge under mild conditions (Khan et al., 2016, Prop. 2–3), *(iii)* recent convergence results also exist for EP (Dehaene and Barthelmé, 2018) under rather restrictive conditions. The search for guarantees for EP is an active research question.

3 INDUCING STATES FOR MARKOVIAN GP MODELS

Despite their success in the large data regime, the computational complexity of the above sparse approximations still makes them unsuitable for long (or unbounded) time series because in order to accurately approximate the posterior, the number of inducing

variables, M, needs to grow with the temporal horizon. Crucially, the posterior prediction of a single data point depends on the entire set of inducing variables $\mathbf{u}$ through the conditional $p(f(\cdot) \mid \mathbf{u})$, even those far apart in time.

In the following sections, we describe how the combination of Markovian GPs with sparse GPs, via state inducing features, further reduces the complexity of the algorithms, making them applicable to long time series.

3.1 State Inducing Features

A key property of the SDE formulation of Markovian GPs is that the state variables $\mathbf{s}(\mathbf{x})$, obtained by marginalizing the SDE to inputs $\mathbf{x} = (x_1, \ldots, x_n)$, have a Markovian property, *i.e.*, $p(\mathbf{s}(x_n) \mid \mathbf{s}(x_{n-1,\ldots,1})) = p(\mathbf{s}(x_n) \mid \mathbf{s}(x_{n-1}))$, which is another way of formulating the definition of the state $\mathbf{s}(x)$ as a summary of all the information necessary to predict the future beyond x. Thus, a natural choice of inducing variables for sparse inference with Markovian GPs is state evaluations, $\mathbf{u} = \mathbf{s}(\mathbf{z})$, indexed at M pseudo input locations $\mathbf{z} = (z_1, \ldots, z_M)$.

This leads to the conditional $f \mid \mathbf{u}$ being local, *i.e.*, if $z_m \leq x_n < z_{m+1}$, and noting $\mathbf{v}_{m(n)} = [\mathbf{u}_m, \mathbf{u}_{m+1}]$, then $p(f_n \mid \mathbf{u}) = p(f_n \mid \mathbf{v}_{m(n)}) = \mathcal{N}(f_n \mid \mathbf{W}_n \mathbf{v}_{m(n)}, \nu_n)$. The conditional is available in closed form via the statistics of the prior transitions, $p(\mathbf{s}_n \mid \mathbf{u}_m)$ and $p(\mathbf{u}_{m+1} \mid \mathbf{s}_n)$, and from the emission matrix $\mathbf{H}$ (see App. A.2). This makes marginal prediction $q(f_n)$, Eq. (4), cheap to evaluate since it only depends on the *local* marginal posterior $q(\mathbf{v}_{m(n)})$.

It should be noted that although the number of inducing points $\mathbf{z}$ is M, the number of inducing variables contained in $\mathbf{u}$ is Md, where d is the state dimension. Indeed each inducing state $\mathbf{s}(z_m)$ contains more information than a single inducing function evaluation $f(z_m) \in \mathbf{s}(z_m)$. In practice, fewer inducing inputs are needed when using inducing states than when using the classic inducing function evaluations (see Fig. 1 and Adam et al., 2020).

3.2 'Doubly Sparse' Variational Inference

The $\mathrm{S^2VGP}$ algorithm (Adam et al., 2020) parameterizes an approximate posterior over the inducing states, $\mathbf{u} = \mathbf{s}(z_1), \ldots, \mathbf{s}(z_M)$, as a linear Gaussian state space model: $q(\mathbf{u}) = q(\mathbf{u}_1) \prod_m q(\mathbf{u}_{m+1} \mid \mathbf{u}_m)$. This shared chain structure with the marginal prior $p(\mathbf{u})$ is optimal. The ELBO can be written as the sums:

$$\mathcal{L}(q) = \sum_n \mathbb{E}_{q(f_n)} \log p(y_n \mid f_n)$$
$$- \sum_m \mathrm{KL}\left[q(\mathbf{u}_{m+1} \mid \mathbf{u}_m) \,\|\, p(\mathbf{u}_{m+1} \mid \mathbf{u}_m)\right]. \quad (10)$$

The marginal posterior predictions $q(f_n)$ can be evaluated independently given the pairwise marginal on the inducing states $q(\mathbf{v}_m)$. These can be computed in linear time with chain length M using classic Kalman filtering algorithms (Särkkä, 2013) or linear algebra routines dedicated to banded matrices (Durrande et al., 2019). Evaluation of the KL divergence and the variational expectations have respective computational costs of $\mathcal{O}(Md^3)$ and $\mathcal{O}(Nd^3)$, leading to an overall computational complexity of $\mathcal{O}((N + M)d^3)$. More details on this algorithm are given in App. C.1.

4 SITE-BASED SPARSE MARKOVIAN GPs

In Sec. 2.2, we reviewed three common algorithms used to perform approximate inference given a sparse formulation of GPs: VI, CVI and PEP. In the special case of sparse Markovian GPs using inducing states, only the VI formulation (Sec. 3.2 and Adam et al., 2020) has been explored. CVI and PEP operate on the precision of the approximating distribution $q(\mathbf{u})$ and turn out to be ideally suited to the Markovian setting where this precision is sparse. In the following sections we describe how to adapt these algorithms to this setting and show how these methods inherit the favourable properties of their parents. We call these algorithms $\mathrm{S^2CVI}$ and $\mathrm{S^2PEP}$. We then go on to show that the doubly sparse approach is even more general, deriving the equivalent algorithms for the classical nonlinear Kalman smoothers, $\mathrm{S^2PL}$ and $\mathrm{S^2EKS}$.

Using state inducing features, the optimal sites for each data point x_n are only functions of the neighbouring states $\mathbf{v}_{m(n)}$ due the local structure of the conditional $f_n \mid \mathbf{u} = f_n \mid \mathbf{v}_{m(n)}$. The approximating distribution $q(\mathbf{u}) = p(\mathbf{u}) \prod_n t_n(\mathbf{v}_{m(n)})$ thus still has a chain structure which we exploit to efficiently compute the marginal statistics $q(\mathbf{v}_m)$ via, *e.g.*, filtering methods (see App. B for details).

A consequence of the locality of the sites is that all data points who lie in the same time segment between consecutive inducing inputs $[z_m, z_{m+1}]$ share the same support, $\mathbf{v}_m$. This provides a natural way to tie these sites together per segment $t_m(\mathbf{v}_m) = \prod_{n \in \mathcal{M}_m} t_n(\mathbf{v}_m)$, where $\mathcal{M}_m$ represents the indices to the data whose inputs fall in $[z_m, z_{m+1}]$. We adopt this approach, reducing our algorithms' memory requirements to $\mathcal{O}(Md^2)$, which is equivalent to $\mathrm{S^2VGP}$. A graphical representation of the approach is depicted in Fig. 2

4.1 Doubly Sparse CVI ($\mathrm{S^2CVI}$)

In the sparse state space formulation, the prior on the inducing states $p(\mathbf{u})$ has sufficient statis-
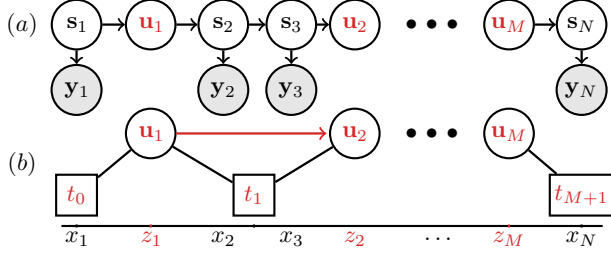
Figure 2: **(a)** Directed graphical model representing the joint prior over states indexed at x_n, associated data points y_n, and inducing states $\mathbf{u}_m$ indexed at z_m. **(b)** Graphical model for shared site based approximate posterior, with local sites $t_m(\mathbf{v}_m)$ represented as factor graph (undirected black edges). The prior over inducing states $p(\mathbf{u})$ is a state space model represented as a directed graphical model (red arrows).

tics $\phi(\mathbf{u}) = [(\mathbf{u}_k, \mathbf{u}_k\mathbf{u}_k^\top)_{k=1}^M, (\mathbf{u}_{k+1}\mathbf{u}_k^\top)_{k=1}^{M-1}]$ whose associated second-order natural parameters are the block-tridiagonal entries of the sparse precision matrix of $p(\mathbf{u})$. The non-tied formulation of CVI would introduce N sites, each dependent only on their nearest inducing states $t_n(\mathbf{v}_{m(n)})$. Indeed, because $p(f_n \,|\, \mathbf{u}) = p(f_n \,|\, \mathbf{v}_{m(n)})$, the site update $\mathbf{g}_n = \nabla_{\boldsymbol{\mu}}\mathbb{E}_{q(f_n)} \log p(y_n \,|\, f_n)$ in Eq. (8) is only non-zero for the natural parameter associated to the sufficient statistics of site t_n, i.e., $[\mathbf{v}_{m(n)}, \mathbf{v}_{m(n)}\mathbf{v}_{m(n)}^\top]$.

Here we use tied sites, parameterising $M + 1$ Gaussian sites $t_m(\mathbf{v}_m)$ with sufficient statistics $[\mathbf{v}_m, \mathbf{v}_m\mathbf{v}_m^\top]$ in their natural form. The edge cases are smaller sites over the first and last inducing states $t_0(\mathbf{u}_0 = \mathbf{s}(-\infty))$ and $t_{M+1}(\mathbf{u}_{M+1} = \mathbf{s}(\infty))$.

Unlike our presentation in Sec. 2.3, the sites are local, only depending on the states that directly neighbour them. The update rule is the same as in Eq. (8), but now using the fact that $p(f_n \,|\, \mathbf{u}) = p(f_n \,|\, \mathbf{v}_{m(n)})$, and that gradient $\mathbf{g}_n$ only contributes to site $t_{m(n)}(\mathbf{v}_{m(n)})$.

The updates can be written in terms of gradients of the variational expectations with respect to the mean and variance of the posterior marginal via the chain rule,

$$
\begin{aligned}
\mathbf{g}_{n,2} &= \mathbf{W}_n^\top \frac{\partial \mathcal{L}_n}{\partial \Sigma_n} \mathbf{W}_n, \\
\mathbf{g}_{n,1} &= \mathbf{W}_n^\top \frac{\partial \mathcal{L}_n}{\partial \mu_n} - 2\mathbf{W}_n^\top \frac{\partial \mathcal{L}_n}{\partial \Sigma_n} \mathbf{W}_n \boldsymbol{\mu}_{m(n)},
\end{aligned}
\tag{11}
$$

where $\mathcal{L}_n = \mathbb{E}_{q(f_n)} \log p(y_n \,|\, f_n)$ and $\mu_n = \mathbf{W}_n\boldsymbol{\mu}_{m(n)}$, $\Sigma_n = \mathbf{W}_n\boldsymbol{\Sigma}_{m(n)}\mathbf{W}_n^\top + \nu_n$ are the moments of $q(f_n)$. This algorithm is equivalent to the natural gradient approach presented in Adam et al. (2020). Its practical implementation is however simpler and less costly since there is no need to compute the KL term of the ELBO

to perform the update of the variational parameters. As a result it is also more numerically stable. The full details of the algorithm are given in App. C.2.

4.2 Doubly Sparse Power Expectation Propagation (S²PEP)

The S²PEP algorithm approximates the joint distribution over the states and the observations,

$$
\begin{aligned}
p(\mathbf{s}(\cdot), \mathbf{y}) &= p(\mathbf{u})\, p(\mathbf{s}(\cdot) \,|\, \mathbf{u}) \prod_n p(y_n \,|\, \mathbf{s}_n) \\
&\approx p(\mathbf{u})\, p(\mathbf{s}(\cdot) \,|\, \mathbf{u}) \prod_m t_m(\mathbf{v}_m) = q(\mathbf{s}(\cdot)), \quad (12)
\end{aligned}
$$

where $t_m(\mathbf{v}_m)$ is the tied site for all $x \in \mathcal{X}_m$, where $\mathcal{X}_m = \{x \in \mathbf{x} \,|\, z_m \leq x < z_{m+1}\}$. We can obtain the site for a single data point x_n as $t_n(\mathbf{v}_{m(n)}) = t_{m(n)}^{1/N_{m(n)}}(\mathbf{v}_{m(n)})$, where $N_{m(n)} = |\mathcal{X}_m|$ is the total number of data points in the neighbourhood. We now outline the PEP steps for updating the sites.

Cavity computation The leave-one-site-out posterior, i.e., the cavity, for a given data point y_n is determined first by computing the approximate posterior over the state with a fraction $\alpha/N_{m(n)}$ of the local site removed,

$$
\begin{aligned}
q^{\text{cavity}}(\mathbf{v}_{m(n)}) &\propto q(\mathbf{v}_{m(n)})\big/t_{m(n)}^{\alpha/N_{m(n)}}(\mathbf{v}_{m(n)}) \\
&= \mathcal{N}(\mathbf{v}_{m(n)} \,|\, \boldsymbol{\mu}_{m(n)}, \boldsymbol{\Sigma}_{m(n)}). \quad (13)
\end{aligned}
$$

The cavity on function evaluation f_n is obtained by marginalizing the joint cavity over f_n and $\mathbf{v}_{m(n)}$:

$$
\begin{aligned}
q^{\text{cavity}}(f_n) &= \int q^{\text{cavity}}(f_n, \mathbf{v}_{m(n)}) \, d\mathbf{v}_{m(n)} \\
&= \int p(f_n \,|\, \mathbf{v}_{m(n)})q^{\text{cavity}}(\mathbf{v}_{m(n)}) \, d\mathbf{v}_{m(n)} \quad (14) \\
&= \mathcal{N}(f_n \,|\, \mu_n = \mathbf{W}_n\boldsymbol{\mu}_{m(n)}, \mathbf{W}_n\boldsymbol{\Sigma}_{m(n)}\mathbf{W}_n^\top + \nu_n).
\end{aligned}
$$

Moment matching We next compute the moments of the so-called *tilted* distribution, i.e., the cavity combined with (a fraction of) the true likelihood function. As discussed in Bui et al. (2017) and Seeger (2005), the required moments can be conveniently obtained via the derivatives of the log-normaliser of the tilted distribution, $\log Z_n = \log \mathbb{E}_{q^{\text{cavity}}}[p^\alpha(y_n \,|\, f_n)]$, with respect to the cavity mean. Doing so provides the new marginal posterior moments $q(\mathbf{v}_{m(n)}) = \mathcal{N}(\mathbf{v}_{m(n)} \,|\, \boldsymbol{\mu}_{m(n)}^{\text{post}}, \boldsymbol{\Sigma}_{m(n)}^{\text{post}})$:

$$
\begin{aligned}
\boldsymbol{\mu}_{m(n)}^{\text{post}} &= \boldsymbol{\mu}_{m(n)} + \mathbf{W}_n \frac{\partial \log Z_n}{\partial \mu_n}, \\
\boldsymbol{\Sigma}_{m(n)}^{\text{post}} &= \boldsymbol{\Sigma}_{m(n)} + \mathbf{W}_n \frac{\partial^2 \log Z_n}{\partial \mu_n^2} \mathbf{W}_n^\top.
\end{aligned}
\tag{15}
$$

For Gaussian likelihoods, the above derivatives are available in closed form, whilst for non-conjugate models we must resort to numerical integration. Given the
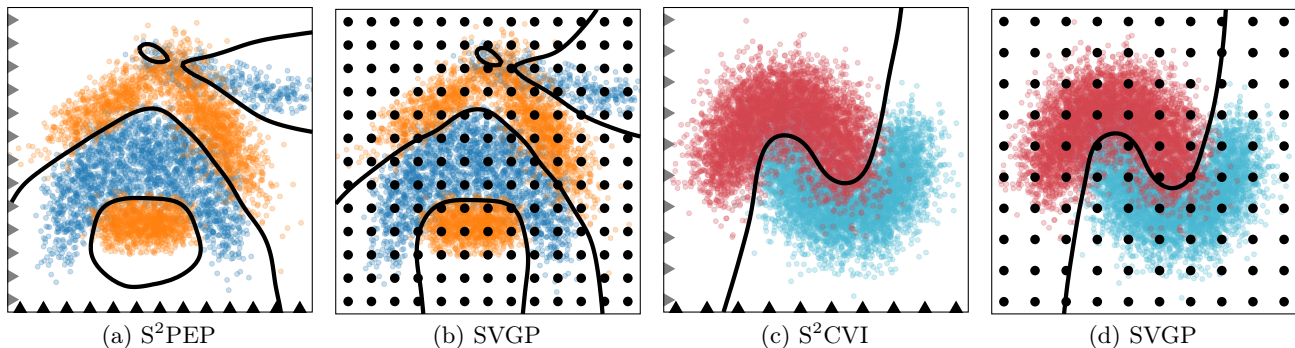
Figure 3: 2D classification tasks. For the doubly sparse methods, (a) and (c), inducing points are placed separately in the sequential dimension (x-axis, 'time', black triangles) and the 'spatial' dimension (y-axis, grey triangles).

new marginal posterior, we can finally compute the new tied site by removing the cavity from the posterior and combining it with a fraction of the old site (representing the other data points in the neighbourhood):

$$t_{m(n)}^{\text{new}}(\mathbf{v}_{m(n)}) = \left[t_{m(n)}^{\text{old}}(\mathbf{v}_{m(n)}) \right]^{1 - \frac{\alpha}{N_m}} \left[\frac{q(\mathbf{v}_{m(n)})}{q^{\text{cavity}}(\mathbf{v}_{m(n)})} \right]. \tag{16}$$

4.3 Doubly Sparse Posterior Linearisation (S²PL) and Nonlinear Kalman Smoothers

Site-based inference is in fact more general than just PEP and CVI. Wilkinson et al. (2020) showed that classical nonlinear Kalman smoothers, such as the Extended, Unscented and Gauss-Hermite smoothers, can also be formulated as site-based algorithms. These algorithms are based on various forms of linearisation of the likelihood model, and their approach is generalised and improved upon in a method called *posterior linearisation* (PL, García-Fernández et al., 2016). We derived a sparse extension to the posterior linearisation algorithms presented in Wilkinson et al. (2020), including a sparse version of the extended Kalman smoother (S²EKS). Details of the derivations are given in App. C.5.

4.4 Algorithmic Details

Approximate marginal likelihood For all our algorithms, the marginal likelihood can be written as: $p(\mathbf{y}) = p(y_1) \, p(y_2 \,|\, y_1) \, p(y_3 \,|\, \mathbf{y}_{1:2}) \prod_{n=4}^{N} p(y_n \,|\, \mathbf{y}_{1:n-1})$, and each term can be approximated during a forward filter pass through the data by noticing that,

$$p(y_n \,|\, \mathbf{y}_{1:n-1}) = \\ \int p(y_n \,|\, f_n = \mathbf{H}\mathbf{s}(x_n)) \, \tilde{p}(\mathbf{s}(x_n) \,|\, \mathbf{y}_{1:n-1}) \, \mathrm{d}\mathbf{s}(x_n), \tag{17}$$

where $\tilde{p}(\mathbf{s}(x_n) \,|\, \mathbf{y}_{1:n-1})$ is an approximate forward filter prediction calculated by replacing the likelihood term by the sites when filtering over the inducing states.

Alternative approximations to the marginal likelihood also exist. The PEP energy is obtained by marginalizing the approximate joint $\mathcal{Z}_{pep} = \int q(\mathbf{s}(\cdot)) \, \mathrm{d}\mathbf{s}(\cdot)$, as described in Bui et al. (2017) and in App. C.3.2. For CVI, the ELBO is typically used in place of the marginal likelihood (see App. C.2.1), and as with the PEP energy all its terms can be computed in $\mathcal{O}(Md^3)$.

Parallelizing the updates The sites may be updated one at a time as described, or they can be updated simultaneously as is done in parallel EP (Li et al., 2015). This particular setting is the closest to S²VGP in terms of both posterior approximation structure, storage and computational complexity. The ability to perform site updates in batches also facilitates stochastic optimisation, leading to overall computational complexity of $\mathcal{O}((M + N_*)d^3)$ for batch size N_*.

4.5 Spatio-temporal Gaussian Processes

As with standard filtering approaches to inference, our doubly sparse approach is compatible with spatio-temporal GP models (Särkkä et al., 2013; Tebbutt et al., 2021), allowing for analysis of data sets with input dimension greater than one. Here, we construct a sparse spatio-temporal GP, $f(x, \mathbf{r})$, with inducing inputs in space, $\mathbf{z}_r$, indexing a finite set of coupled inducing temporal GPs, $\mathbf{s}(x)$, and we also impose these temporal GPs to be sparse with inducing states $\mathbf{u}$ indexed at temporal inputs $\mathbf{z}_x$. Fig. 3 shows a demonstration of this approach on two-dimensional classification tasks.

We focus on the case of separable stationary spatio-temporal kernels where $\kappa(x, \mathbf{r}, x', \mathbf{r}') = \kappa_x(x - x') \, \kappa_r(\mathbf{r} - \mathbf{r}')$ and where κ_x is Markovian with state dimension d. A spatio-temporal GP with such a kernel has an equivalent representation as a stochastic partial differential equation (SPDE),

$$\frac{\partial \mathbf{s}(x, \cdot)}{\partial x} = \mathcal{A}_r \mathbf{s}(x, \cdot) + \mathcal{L}_r \mathbf{w}(x, \cdot), \\ f(x, \mathbf{r}) = \mathbf{H}_x \mathbf{s}(x, \mathbf{r}). \tag{18}$$
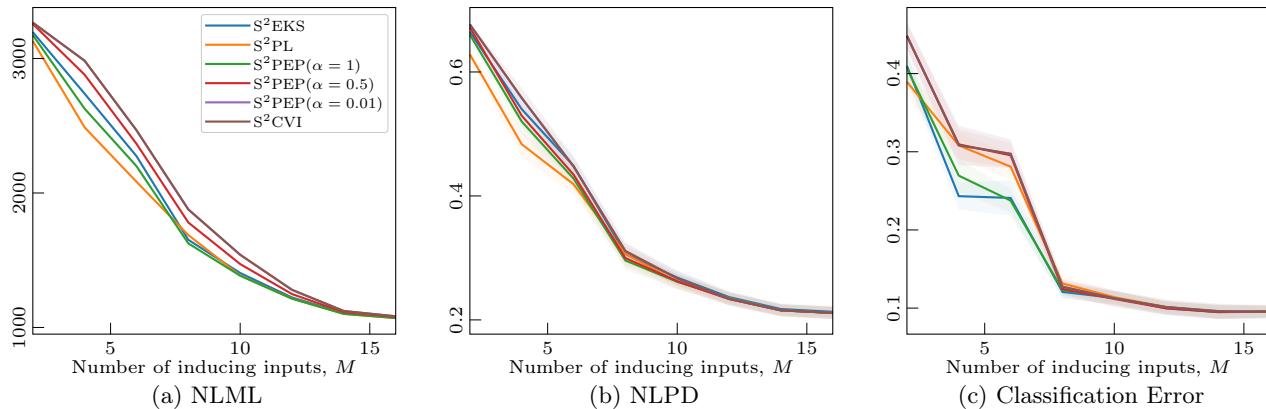
Figure 4: The Banana 2D classification task with varying number of inducing inputs (mean and standard deviation of 10-fold cross-validation). All methods improve monotonically as M increases. S^2PEP ($\alpha = 1$) and S^2PL slightly outperform the other methods when M is small. S^2CVI and S^2PEP ($\alpha = 0.01$) perform identically.

This infinite-dimensional SPDE marginalized to a finite set of M_z spatial locations $\mathbf{z}_r$ is a finite-dimensional SDE with state dimension $M_z d$. Noting $\mathbf{s}(x) = \boldsymbol{s}(x, \mathbf{z}_r)$ we have,

$$\frac{\mathrm{d}\mathbf{s}(x)}{\mathrm{d}x} = \mathbf{F}\mathbf{s}(x) + \mathbf{L}\boldsymbol{\varepsilon}(x). \tag{19}$$

The parameters of the SDE in Eq. (19) are given in Särkkä et al. (2013). Intuitively, state $\mathbf{s}$ splits into M_z correlated temporal processes $\mathbf{s} = [\mathbf{s}_1, \dots \mathbf{s}_{M_z}]$, whose marginal projection $f_i(\cdot) = \mathbf{H}_x \mathbf{s}_i(\cdot)$ verify

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[f_i(x), f_i(x')] &= \kappa_x(x - x')\kappa_r(0), \\
\mathbb{C}\mathrm{ov}[f_i(x), f_j(x)] &= \kappa_x(0)\kappa_r(\mathbf{z}_r^i - \mathbf{z}_r^j).
\end{aligned} \tag{20}$$

We can further marginalise this SDE to its values $\mathbf{u} = \mathbf{s}(\mathbf{z}_x)$ at M_x temporal inputs $\mathbf{z}_x$, leading to the discrete state-space model,

$$\begin{aligned}
\mathbf{u}_0 &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{z}_r, \mathbf{z}_r}^r \otimes \mathbf{P}_0^x), \\
\mathbf{u}_{m+1} &= \mathbf{A}_{m, m+1} \mathbf{u}_m + \mathbf{q}_m,
\end{aligned} \tag{21}$$

where $\mathbf{P}_0^x$ is the stationary covariance of $\mathbf{s}$ in Eq. (19). In our sparse algorithms for spatio-temporal models, we use $\mathbf{u}$ as inducing variables and we need the conditional $p(f_n \mid \mathbf{u})$ to make predictions about the process.

For a single data point, $(x_n, \mathbf{r}_n, y_n)$, and denoting $f_n = f(x_n, \mathbf{r}_n)$, there is a conditional independence property specific to separable kernels: $p(f_n \mid \mathbf{s}(\cdot)) = p(f_n \mid \mathbf{s}(x_n))$ (see Tebbutt et al., 2021). The conditional $p(f_n \mid \mathbf{u})$ is then obtained by marginalizing $\mathbf{s}(x_n)$ in the joint $p(f_n, \mathbf{s}(x_n) \mid \mathbf{u}) = p(f_n \mid \mathbf{s}(x_n))p(\mathbf{s}(x_n) \mid \mathbf{u})$, given by

$$\begin{aligned}
p(\mathbf{s}(x_n) \mid \mathbf{u}) &= \mathcal{N}(\mathbf{s}(x_n) \mid \mathbf{R}_n \mathbf{v}_{m(n)}, \mathbf{T}_n), \\
p(f_n \mid \mathbf{s}(x_n)) &= \mathcal{N}(f_n \mid \mathbf{B}(\mathbf{r}_n)\mathbf{s}(x_n), \mathbf{C}(\mathbf{r}_n)),
\end{aligned} \tag{22}$$

where $\mathbf{R}_n$ and $\mathbf{T}_n$ are defined as in App. A.2 and

$$\begin{aligned}
\mathbf{B}(\mathbf{r}_n) &= \left[\mathbf{K}_{\mathbf{r}_n, \mathbf{z}_r} \mathbf{K}_{\mathbf{z}_r, \mathbf{z}_r}^{-1}\right] \otimes \mathbf{H}, \\
\mathbf{C}(\mathbf{r}_n) &= \mathbf{K}_{\mathbf{r}_n, \mathbf{r}_n} - \mathbf{K}_{\mathbf{r}_n, \mathbf{z}_r} \mathbf{K}_{\mathbf{z}_r, \mathbf{z}_r}^{-1} \mathbf{K}_{\mathbf{z}_r, \mathbf{r}_n}.
\end{aligned} \tag{23}$$

There are several approximate inference approaches based on the SPDE formulation. *(i)* In Simpson et al. (2012), the SPDE is approximated via a local basis expansion where the associated weights are distributed as a Gaussian Markov random field. Its sparse precision matrix leads to efficient computations. However, the generative model (prior) is approximated, which is not the case in our approach. *(ii)* Global approximations based on the SPDE formulation (*e.g.*, Solin and Särkkä, 2020) also approximate the prior based on the truncation of an exact infinite expansion of the kernel. Our approach singles out a time dimension which turns the SPDE into a SDE with an infinite-dimensional state (Särkkä et al., 2013). Using a further sparse approximation to this infinite dimensional state leads to inference in an SDE with finite dimension.

5 EMPIRICAL ANALYSIS

In Fig. 4 we analyse the effect of increasing the number of inducing inputs, M, in a two-dimensional classification task. We observe that the training marginal likelihood (NLML), the test predictive density (NLPD), and the test classification error improve monotonically as M increases, as expected. When M is very small, the methods which use the EP energy for training (S^2PEP ($\alpha = 1$), S^2PL, S^2EKS) perform well. We provide a similar analysis of the more complicated Audio task in App. D, in which S^2PEP again requires few inducing inputs to obtain good results.

Table 1 analyses the practical performance of our site-based algorithms relative to the S^2VGP approach, and

Table 1: Normalised negative log predictive density (NLPD) results using 10-fold cross-validation. Mean and standard deviation shown (smaller is better). The banded matrix operations currently used for filtering in S^2VGP are unstable for large datasets, while SVGP does not scale to more than ~ 1000 inducing points.

	Motorcycle	Coal	Banana	Binary	Audio	Airline	Electricity
# data points, N	133	333	5300	10k	22k	36k	262k
# inducing inputs, M	30	15	15×15	1k	3k	4k	50k
Input dimension	1	1	2	1	1	1	1
Likelihood	Heteroscedastic	Poisson	Bernoulli	Bernoulli	Product	Poisson	Gaussian
S^2EKS	0.870±0.16	0.924±0.11	0.212±0.01	0.205±0.02	0.218±0.00	0.128±0.04	−0.085±0.02
S^2PL	0.892±0.15	0.925±0.11	0.211±0.01	0.189±0.02	0.213±0.11	0.128±0.04	−0.085±0.02
S^2PEP ($\alpha = 1$)	0.456±0.37	0.924±0.11	0.211±0.01	0.189±0.02	**−1.326±0.01**	0.128±0.04	−0.085±0.02
S^2PEP ($\alpha = 0.5$)	**0.420±0.35**	0.924±0.11	0.211±0.01	0.189±0.02	−0.624±0.07	0.128±0.04	−0.074±0.02
S^2PEP ($\alpha = 0.01$)	0.428±0.33	0.924±0.11	0.211±0.01	**0.188±0.02**	0.624±0.04	0.128±0.04	**−0.153±0.01**
S^2CVI	0.428±0.33	0.924±0.11	0.211±0.01	**0.188±0.02**	0.681±0.03	0.128±0.04	−0.152±0.01
S^2VGP	0.434±0.31	0.937±0.10	0.215±0.01	0.236±0.01	×	×	×
Sparse VI (SVGP)	0.440±0.30	0.954±0.12	0.226±0.01	0.207±0.02	×	×	×
Full EKS	0.871±0.16	0.924±0.11	0.212±0.01	0.205±0.02	0.412±0.01	0.128±0.04	−0.420±0.01
Full PL	0.893±0.15	0.924±0.12	0.211±0.01	0.189±0.02	−0.514±0.18	0.128±0.04	−0.420±0.01
Full PEP ($\alpha = 1$)	0.429±0.31	0.924±0.12	0.211±0.01	0.189±0.02	−1.441±0.02	0.128±0.04	−0.420±0.01
Full PEP ($\alpha = 0.5$)	0.422±0.31	0.924±0.12	0.211±0.01	0.189±0.02	−0.902±0.05	0.128±0.04	−0.420±0.01
Full PEP ($\alpha = 0.01$)	0.416±0.32	0.924±0.12	0.211±0.01	0.188±0.02	0.169±0.08	0.128±0.04	−0.420±0.01
Full CVI	0.415±0.32	0.924±0.12	0.211±0.01	0.188±0.02	0.671±0.03	0.128±0.04	−0.420±0.01

compared to their full (non-sparse) site-based equivalents. We use six of the benchmark tasks presented in Wilkinson et al. (2020), and add the additional Electricity task in order to show that the methods are applicable to extremely large time series. See App. D for further details on the data sets and full models used. Four of the data sets are sufficiently small that we are also able to compare against standard sparse VI (SVGP, using GPflow, Matthews et al., 2017).

We measure the negative log predictive density (NLPD), using 10-fold cross-validation, with each method run for 500 training iterations. Each iteration consists of an update to the sites, followed by a gradient step to update the hyperparameters using Adam. The ELBO is used as the training objective for the variational methods, and the power EP energy for PEP (see App. C.3.2). The PL and EKS methods also use the PEP energy ($\alpha = 1$, see App. C.5 for discussion). In the non-conjugate tasks, we use Gauss–Hermite integration with 20^q cubature points, where q is the dimensionality of the integral being approximated. However, in the Audio task this approach is not practical, and so we use the fifth-order Unscented transform (McNamee and Stenger, 1967).

Results We observe the performance to be highly model and task dependent. For many tasks, the methods all perform similarly. However, it is worth noting that S^2PEP performs well on the difficult Audio task which has a highly non-Gaussian likelihood model. It appears that the PEP energy is a good training objective in this case, whereas S^2CVI sometimes outperforms S^2PEP when the likelihood is simpler (*e.g.*, the Binary classification task, and the Electricity regression task). Fig. 3 illustrates the performance of the doubly sparse methods in comparison to SVGP on the 2D classification task. In the heteroscedastic noise task, S^2PEP

is the best performing sparse method. As expected, S^2PEP ($\alpha = 0.01$) and S^2CVI give similar results, except in the Audio task, where numerical integration error resulting from the three-dimensional cubature used for the updates causes the results to differ.

6 CONCLUSIONS

We have derived site-based inference methods for sparse Markovian GPs. In doing so, we have shown the generality of the sparse Markovian approach, and provided a suite of algorithms applicable to large temporal data. We also proposed a principled approach to site tying motivated by the specific structure of the prior on inducing states, resulting in methods with very efficient computational and memory scaling.

The site-based approach makes it possible to apply PEP (as well as the classical Kalman smoothers) in the doubly sparse framework, and this method outperforms existing approaches on some difficult non-conjugate tasks. Our new algorithms inherit many of the desirable properties of their full counterparts, including the ability to handle spatio-temporal models, resulting in a novel sparse approach in which inducing points in time and space are fully decoupled.

Acknowledgements

References

V. Adam, S. Eleftheriadis, A. Artemev, N. Durrande, and J. Hensman. Doubly sparse variational gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2874–2884. PMLR, 2020.

B. M. Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4 (3):626–636, 1994.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne. JAX: Composable transformations of Python+NumPy programs, 2018. `http://github.com/google/jax`.

T. D. Bui and R. E. Turner. Tree-structured Gaussian process approximations. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pages 2213–2221. Curran Associates, Inc., 2014.

T. D. Bui, J. Yan, and R. E. Turner. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research (JMLR)*, 18 (1):3649–3720, 2017.

D. Burt, C. E. Rasmussen, and M. Van Der Wilk. Rates of convergence for sparse variational gaussian process regression. In *International Conference on Machine Learning*, pages 862–871. PMLR, 2019.

P. E. Chang, W. J. Wilkinson, M. E. Khan, and A. Solin. Fast variational learning in state-space Gaussian process models. In *International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020.

G. Dehaene and S. Barthelmé. Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):199–217, 2018.

N. Durrande, V. Adam, L. Bordeaux, S. Eleftheriadis, and J. Hensman. Banded matrix operators for Gaussian Markov models in the automatic differentiation era. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2780–2789. PMLR, 2019.

V. Dutordoir, N. Durrande, and J. Hensman. Sparse Gaussian processes with spherical harmonic features. *arXiv preprint arXiv:2006.16649*, 2020.

Á. F. García-Fernández, L. Svensson, and S. Särkkä. Iterated posterior linearization smoother. *IEEE Transactions on Automatic Control*, 62(4):2056–2063, 2016.

Á. F. García-Fernández, F. Tronarp, and S. Särkkä. Gaussian process classification using posterior linearization. *IEEE Signal Processing Letters*, 26(5): 735–739, 2019.

G. Hébrail and A. Bérard. Individual household electric power consumption data set, 2012. URL `https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption`. Online: UCI Machine Learning Repository.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 282–290. AUAI Press, 2013.

J. Hensman, A. Matthews, and Z. Ghahramani. Scalable variational Gaussian process classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360. PMLR, 2015.

J. Hensman, N. Durrande, and A. Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 18:1–52, 2018.

M. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 878–887. PMLR, 2017.

M. E. Khan, R. Babanezhad, W. Lin, M. Schmidt, and M. Sugiyama. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 319–328, 2016.

Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2323–2331. Curran Associates, Inc., 2015.

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

J. McNamee and F. Stenger. Construction of fully symmetric numerical integration formulas. *Numerische Mathematik*, 10(4):327–344, 1967.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 17, pages 362–369. AUAI Press, 2001.

T. P. Minka. Power EP. Technical report, Microsoft Research, 2004. MSR-TR-2005-173.

H. Nickisch, A. Solin, and A. Grigorievskiy. State space Gaussian processes with non-Gaussian likelihood. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3789–3798. PMLR, 2018.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research (JMLR)*, 6(Dec):1939–1959, 2005.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006.

H. Salimbeni, S. Eleftheriadis, and J. Hensman. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 689–697. PMLR, 2018.

S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

S. Särkkä, A. Solin, and J. Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

M. Seeger. Expectation propagation for exponential families. Technical report, University of California at Berkeley, 2005.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21, 1985.

D. Simpson, F. Lindgren, and H. Rue. Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29, 2012.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 1257–1264, 2006.

A. Solin and S. Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30(2):419–446, 2020.

A. Solin, J. Hensman, and R. E. Turner. Infinite-horizon Gaussian processes. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 3486–3495. Curran Associates, Inc., 2018.

W. Tebbutt, A. Solin, and R. E. Turner. Combining pseudo-point and state space approximations for sum-separable gaussian processes. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=Ctq5FVu8KX.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR, 2009.

S. T. Tokdar and J. K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34–42, 2007.

V. Tolvanen, P. Jylänki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 14(Apr):1175–1179, 2013.

W. J. Wilkinson, M. R. Andersen, J. D. Reiss, D. Stowell, and A. Solin. End-to-end probabilistic inference for nonstationary audio analysis. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6776–6785. PMLR, 2019.

W. J. Wilkinson, P. E. Chang, M. R. Andersen, and A. Solin. State space expectation propagation: Efficient inference schemes for temporal Gaussian processes. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.