Combining Pseudo-Point and State Space Approximations for Sum-Separable Gaussian Processes

Anonymous Authors
Anonymous Institution

Abstract

Spatio-temporal Gaussian processes (GPs) are important probabilistic tools for inference and learning in climate science, epidemiology, or any time-driven general GP modelling problem. The current gold-standard methods for scaling GPs to large data sets are various flavours of pseudo-point methods. These methods do not cope well with long or unbounded temporal observation horizons, which undermines their efficiency and effectively turns the computational scaling back to cubic in the number of temporal observations. On the other hand, if the temporal part in the GP prior admits a Markov form, the inference can be sped up to linear in the number of temporal observations by using state space models. In this work we show how to combine the most widely used pseudo-point method, Titsias' variational approach, with the state space approximation framework. Our approach hinges on a surprising conditional independence property which applies to space–time separable GPs. By utilising pseudo-point approximations over space, and state space approximations through time, we are able to construct an approximation that is more scalable and widely applicable to spatio-temporal problems than either method on their own.

1. Introduction

Large spatio-temporal data containing millions of observations arise in various domains, such as climate science. While Gaussian process (GP) models (Rasmussen and Williams, 2006) are effective in such settings, their computational expense is prohibitive if they are employed naively. Consequently, approximation is necessary. In this work we combine the complementary strengths of pseudo-point (Quiñonero-Candela and Rasmussen, 2005; Bui et al., 2017) and state-space (Särkkä et al., 2013; Särkkä and Solin, 2019) approximations to tackle spatio-temporal problems.

At its core this work hinges on an observation made by O'Hagan (1998) that, if a GP's kernel is separable, then it possesses a surprising kind of conditional independence property. Interestingly, this observation appears to have gone largely unnoticed within the GP community. This facilitates the combination of pseudo-point and state-space approximations, resulting in algorithms that scale linearly in time. In particular, we show *(i)* how the conditional independence property can be exploited to significantly accelerate the variational inference scheme of Titsias (2009) for GPs with separable kernels and sum-separable kernels, *(ii)* how this can be straightforwardly combined with the Markov property exploited by state-space approximations (Särkkä and Solin, 2019) to obtain an accurate approximate inference algorithm for sum-separable spatio-temporal GPs, that scales linearly in time, and *(iii)* how the earlier work of Hartikainen et al. (2011) on this topic is more closely-related to Snelson and Ghahramani (2005) than previously realised.

2. Pseudo-Point Approximations

Consider a GP prior $f \sim \mathcal{GP}(m, \kappa)$ and an observation (likelihood) model with N observations $\mathbf{y} \in \mathbb{R}^N$ made at locations $\mathbf{x} \in \mathcal{X}^N$, with conditional distribution $p(\mathbf{y} \,|\, \mathbf{f}) = \prod_{n=1}^N p(\mathbf{y}_n \,|\, \mathbf{f}_n)$, $\mathbf{f}_n := f(\mathbf{x}_n)$. The seminal work of Titsias (2009), later clarified by de G. Matthews et al. (2016), introduced the following approximation to the posterior distribution over f:

$$q(f) = q(\mathbf{u})\, p(f_{\neq \mathbf{u}} \,|\, \mathbf{u}) \,, \tag{1}$$

where $\mathbf{u}_m := f(\mathbf{z}_m)$ are the *pseudo-points* for a collection of M *pseudo-inputs* $\mathbf{z}_{1:M}$, and $f_{\neq \mathbf{u}} := f \setminus \mathbf{u}$ are all of the random variables in f except those used as pseudo-points. We assume that $q(\mathbf{u})$ is Gaussian with mean $\hat{\mathbf{m}}_\mathbf{u}$ and covariance matrix $\hat{\mathbf{C}}_\mathbf{u}$; subject to the constraint imposed in Eq. (1) this family contains the optimal choice for $q(\mathbf{u})$ if each conditional $p(\mathbf{y}_n \,|\, f(\mathbf{x}_n))$ is Gaussian, and is otherwise the de-facto standard choice (Hensman et al., 2013). This yields the approximate posterior predictive distribution at any collection of test points $\mathbf{x}_*$

$$q(\mathbf{f}_*) = \mathcal{N}\Big(\mathbf{f}_*; \mathbf{m}_{\mathbf{f}_*} + \mathbf{C}_{\mathbf{f}_*\mathbf{u}} \Lambda_\mathbf{u} (\hat{\mathbf{m}}_\mathbf{u} - \mathbf{m}_\mathbf{u}), \mathbf{C}_{\mathbf{f}_*} - \mathbf{C}_{\mathbf{f}_*\mathbf{u}} \Lambda_\mathbf{u} \mathbf{C}_{\mathbf{u}\mathbf{f}_*} + \mathbf{C}_{\mathbf{f}_*\mathbf{u}} \Lambda_\mathbf{u} \hat{\mathbf{C}}_\mathbf{u} \Lambda_\mathbf{u} \mathbf{C}_{\mathbf{u}\mathbf{f}_*}\Big) \,, \tag{2}$$

where $\Lambda_\mathbf{u} := \mathbf{C}_\mathbf{u}^{-1}$ is the inverse of the covariance matrix between all pseudo-points, $\mathbf{C}_{\mathbf{f}_*\mathbf{u}}$ is the cross-covariance between the prediction points and pseudo-points under f, and $\mathbf{m}_\mathbf{u}$ and $\mathbf{m}_{\mathbf{f}_*}$ are the mean vectors at the pseudo points and prediction points respectively. Supposing that the conditionals are Gaussian, which we denote by $p(\mathbf{y} \,|\, \mathbf{f}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{S})$ for some positive-definite diagonal matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$, it is possible to optimise $q(\mathbf{u})$ analytically and obtain a closed-form expression for the ELBO, known as the *saturated bound*:

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y}; \mathbf{m}_\mathbf{f}, \mathbf{C}_{\mathbf{f}\mathbf{u}} \Lambda_\mathbf{u} \mathbf{C}_{\mathbf{u}\mathbf{f}} + \mathbf{S}) - \frac{1}{2} \mathrm{tr}\big(\mathbf{S}^{-1}(\mathbf{C}_\mathbf{f} - \mathbf{C}_{\mathbf{f}\mathbf{u}} \Lambda_\mathbf{u} \mathbf{C}_{\mathbf{u}\mathbf{f}})\big) \,. \tag{3}$$

Through the use of the matrix inversion and determinant lemmas, this quantity can be computed using only $\mathcal{O}(NM^2)$ operations. This is typically acceptable for regression tasks where the inputs are sampled i.i.d. as the value of M required as N increases generally seems not to grow too fast—indeed Burt et al. (2019) showed that if the inputs $\mathbf{x}_n$ are sampled i.i.d. from a Gaussian then the value of M required scales roughly logarithmically in N. However, Bui and Turner (2014) noted that this is typically not the case for time series problems, where the interval in which the observations live typically grows linearly in N. Moreover Tobar (2019) showed that to ensure the posterior approximation does not degrade, the density of the pseudo-points must not drop below a rate analogous to the Nyquist-Shannon rate. Consequently the number of pseudo-points M required to maintain a good approximation must grow lineary in N, so the cost of accurate approximate inference is really $\mathcal{O}(N^3)$ in this case.

3. State-Space Approximations to Sum-Separable Spatio-Temporal GPs

Särkkä et al. (2013) and Särkkä and Solin (2019) showed that given a separable spatio-temporal GP $f(\tau, \mathbf{r})$ can be approximated by another GP $\bar{f}(\tau, \mathbf{r}, d)$, where $d \in \{1, \dots, D\}$

picks one of D latent dimensions which render $\bar{f}$ Markov in τ, the first of which approximates f. This approximation can generally be made tight, and achieves equality in various interesting cases (e.g., Matérn-family GPs). $\bar{f}$ is specified implicitly through a linear stochastic differential equation, meaning that inference can be performed via efficient filtering / smoothing in a Linear-Gaussian State-Space Model (LGSSM). Let $\bar{\mathbf{f}}_t$ be the collection of random variables in $\bar{f}$ at inputs given by the Cartesian product between the singleton $\{t\}$, N_T arbitrary locations in space $\mathbf{r}_{1:N_T}$, and all of the latent dimensions $\{1, \dots, D\}$. Further, let the kernel of f be separable: $\kappa((\mathbf{r}, \tau), (\mathbf{r}', \tau')) = \kappa^{\mathbf{r}}(\mathbf{r}, \mathbf{r}')\, \kappa^{\tau}(\tau, \tau')$. Any collection of such finite dimensional marginals $\bar{\mathbf{f}} := \bar{\mathbf{f}}_{1:T}$, each using the same $\mathbf{r}_{1:N_T}$, form an LGSSM with $N_T D$-dimensional state-space, and dynamics

$$\bar{\mathbf{f}}_t = \mathbf{A}_t \bar{\mathbf{f}}_{t-1} + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{\mathbf{r}}_{\mathbf{f}} \otimes \mathbf{Q}_t)\,, \qquad \text{s.t.} \quad \mathbf{f}_{n,t} = \mathbf{H}\bar{\mathbf{f}}_{n,t}, \tag{4}$$

$$\mathbf{y}_{n,t} = \mathbf{f}_{n,t} + \mathbf{w}_{n,t}, \qquad \mathbf{w}_{n,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_{n,t})\,, \tag{5}$$

where $\mathbf{H} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{1 \times D}$ ignores all but the first dimension of $\bar{\mathbf{f}}_{n,t}$, $\mathbf{A}_t$ and $\mathbf{Q}_t$ are functions of κ^{τ}, and $\mathbf{C}^{\mathbf{r}}_{\mathbf{f}}$ is the covariance matrix associated with $\kappa^{\mathbf{r}}$ and $\mathbf{r}_{1:N_T}$. This formulation can be straightforwardly extended to sums of separable processes (App. B). While this formulation truly scales linearly in T it has two clear limitations, *(i)* all locations of observations must lie on a rectilinear time-space grid if any computational gains are to be achieved; and *(ii)* inference scales cubically in N_T, meaning that inference is rendered infeasible by time or memory constraints if a large number of spatial locations are observed.

4. Exploiting Separability to Obtain the Best of Both Worlds

We now combine the pseudo-point and state-space approximations, and show how a temporal conditional independence property means that the optimal approximate posterior is Markov. This in turn leads to a closed-form expression for the optimum under Gaussian observation models, and the existence of a simplified LGSSM in which exact inference yields optimal approximate inference in the original model.

Pseudo-Point Approximation of State-Space Augmentation We perform approximate inference in a GP f by applying the standard variational pseudo-point approximation (Sec. 2) to its state-space augmentation (Sec. 3) $\bar{f}$:

$$q(\bar{f}) := q(\bar{\mathbf{u}})\, p(\bar{f}_{\neq \bar{\mathbf{u}}} \,|\, \bar{\mathbf{u}})\,, \tag{6}$$

where the pseudo-points $\bar{\mathbf{u}} = \bar{\mathbf{u}}_{1:T}$ form a rectilinear grid of points in time, space, and the latent dimensions with the same structure as $\bar{\mathbf{f}}$ in Sec. 3, but replacing $\mathbf{r}_{1:N_T}$ with a collection of M_τ spatial pseudo-inputs, for a total of $T M_\tau D$ pseudo-points. $p(\bar{\mathbf{u}})$ is therefore Markov-through-time, so has block-tridiagonal precision $\Lambda_{\bar{\mathbf{u}}}$ (Grigorievskiy et al., 2017).

Crucially, we will now relax the assumption that $\mathbf{f}$ must associated with inputs on a rectilinear grid, requiring only that each observation is made at one of the T times at which we have placed pseudo-points. We denote the number of observations at time t by N_t.

Exploiting Conditional Independence Due to O'Hagan (1998)'s conditional independence property, $p(\bar{\mathbf{f}}_t \,|\, \bar{\mathbf{u}}) = p(\bar{\mathbf{f}}_t \,|\, \bar{\mathbf{u}}_t)$; see App. A for details. Consequently, the reconstruction terms in the ELBO depend only on $\bar{\mathbf{u}}_t$ as opposed to the entirety of $\bar{\mathbf{u}}$:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{n=1}^{N_t} \mathbb{E}_{q(\bar{\mathbf{u}}_t)} \left[\mathbb{E}_{p(\mathbf{f}_{n,t} \,|\, \bar{\mathbf{u}}_t)} [\log p(\mathbf{y}_{n,t} \,|\, \mathbf{f}_{n,t})] \right] - \mathcal{KL}[q(\bar{\mathbf{u}}) \,\|\, p(\bar{\mathbf{u}})] \tag{7}$$

This property yields substantial computational savings in both the separable and sum-separable case (App. B.1).

The Optimal Approximate Posterior is Markov As an immediate consequence of Eq. (7), by the same argument as that made by Opper and Archambeau (2009) the optimal approximate posterior precision has the form

$$\hat{\Lambda}^*_{\bar{\mathbf{u}}} = \Lambda_{\bar{\mathbf{u}}} + \underbrace{\begin{bmatrix} \mathbf{G}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{G}_T \end{bmatrix}}_{\mathbf{G}}, \ \mathbf{G}_t := -2 \nabla_{\hat{\mathbf{C}}_{\bar{\mathbf{u}}_t}} \sum_{n=1}^{N_t} \mathbb{E}_{q(\bar{\mathbf{u}}_t)} \left[\mathbb{E}_{p(\mathbf{f}_{n,t} \,|\, \bar{\mathbf{u}}_t)} [\log p(\mathbf{y}_{n,t} \,|\, \mathbf{f}_{n,t})] \right]. \tag{8}$$

$\mathbf{G}$ is block-diagonal with the same block-sizes as $\Lambda_{\bar{\mathbf{u}}}$, so $\hat{\Lambda}^*_{\bar{\mathbf{u}}}$ is also block-tridiagonal and the optimal approximate posterior is Markov; see App. C for details. Moreover, Ashman et al. (2020) (App. A) show that $\mathbf{G}_t$ can be written as a sum of N_t rank-1 matrices.

Optimal Approx. Posterior under Gaussian Observation Model In the case that each $p(\mathbf{y}_{n,t} \,|\, \mathbf{f}_{n,t}) = \mathcal{N}(\mathbf{y}_{n,t}; \mathbf{f}_{n,t}, \mathbf{S}_{n,t})$, an explicit expression for the optimal parameters of the approximate posterior $\mathcal{N}\left(\bar{\mathbf{u}}; \hat{\mathbf{m}}^*_{\bar{\mathbf{u}}}, [\hat{\Lambda}^*]^{-1}_{\bar{\mathbf{u}}}\right)$ can be obtained:

$$\hat{\Lambda}^*_{\bar{\mathbf{u}}} = \Lambda_{\bar{\mathbf{u}}} + \mathbf{G}, \text{ where } \mathbf{G}_t := \mathbf{H}^\top_{\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t \mathbf{f}_t} \mathbf{S}_t^{-1} \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t}, \tag{9}$$

$$\hat{\mathbf{m}}^* = \mathbf{m}_{\bar{\mathbf{u}}} + [\hat{\Lambda}^*]^{-1}_{\bar{\mathbf{u}}} \mathbf{H}^\top_{\mathbf{u}} \Lambda_{\mathbf{u}} \mathbf{C}_{\mathbf{u}\mathbf{f}} \mathbf{S}^{-1} (\mathbf{y} - \mathbf{m}_{\mathbf{f}}), \tag{10}$$

where $\mathbf{H}_{\mathbf{u}_t} := \mathbf{I}_{M_\tau} \otimes \mathbf{H}$, $\mathbf{H}_{\mathbf{u}} := \mathbf{I}_T \otimes \mathbf{H}_{\mathbf{u}_t}$, $\mathbf{S}_t := \text{blk-diag}(\mathbf{S}_{1,t}, \ldots, \mathbf{S}_{N_t,t})$, $\mathbf{S} := \text{blk-diag}(\mathbf{S}_1, \ldots, \mathbf{S}_T)$, and $\Lambda_{\mathbf{u}} \mathbf{C}_{\mathbf{u}\mathbf{f}}$ are block-diagonal; see App. A.3.

Approximate Inference via Exact Inference in an Approximate Model This optimal approximate posterior over $\bar{\mathbf{u}}$ is equal to the exact posterior under the simpler state-space model given by

$$\tilde{p}(\bar{\mathbf{u}}) := p(\bar{\mathbf{u}}), \quad \tilde{p}(\mathbf{y}_t \,|\, \bar{\mathbf{u}}_t) := \prod_{n=1}^{N_t} \mathcal{N}\left(\mathbf{y}_{n,t}; \mathbf{m}_{\mathbf{f}_{n,t}} + \mathbf{C}_{\mathbf{f}_{n,t}\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}} (\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}), \mathbf{S}_{n,t}\right), \tag{11}$$

that is, $\tilde{p}(\bar{\mathbf{u}} \,|\, \mathbf{y}) = \mathcal{N}\left(\bar{\mathbf{u}}; \hat{\mathbf{m}}^*_{\bar{\mathbf{u}}}, [\hat{\Lambda}^*]^{-1}_{\bar{\mathbf{u}}}\right)$. This is analogous to the relationship between the approximate model employed by the *Deterministic Training Conditional* (DTC) (Seeger et al., 2003), and the and the variational approximation (Titsias, 2009). Indeed, this model *is* precisely that employed by DTC, and it can be exploited both to perform approximate inference and the saturated bound; see App. D. This allows for approximate inference to be performed by running exact inference in an LGSSM, which scales *linearly* in time and can re-use existing code. For example, it is clear that the parallelised inference procedure derived by Särkkä and García-Fernández (2020) could be utilised in this approximation.

Computational Intensity The total number of flops required to compute the saturated ELBO is roughly $TM_\tau^3 + M_\tau^2 \sum_{t=1}^T N_t$. This is a great deal fewer when T is large than the $M^3 + M^2 N = M_\tau^3 T^3 + M_\tau^2 T^2 N$ required if the bound is computed naively.

Related Work The conditional independence property exploited to develop the variational approximation in this section also shines new light on the work of Hartikainen et al. (2011). Specifically, performing inference in the approximate model they introduce would yield FITC (Snelson and Ghahramani, 2005) *exactly*; see App. E.

The popular Kronecker-product methods for separable kernels explored by (Saatçi, 2012) are unable to handle heteroscedastic observation noise / missing data, scale cubically in time, and require observations to lie on a rectilinear grid. Our approach suffers none of these drawbacks.

5. Experiments: Proof-of-Concept

We conducted two simple proof-of-concept experiments on synthetic data with a separable GP to verify our approach to approximate inference. Further empirical evaluation of the method on real-world problems and sum-separable kernels is on-going. All timing experiments are conducted using a single thread on a 2019 MacBook Pro with 2.6 GHz CPU. In both experiments we consider quite a large temporal extent, but only moderate spatial, since we expect the proposed method to perform well in such situations. If the spatial extent of a data set of very large relative to the characteristic spatial variation, pseudo-point methods will struggle and, by extension, so will our method.

Arbitrary Spatial Locations Fig. 1 (lhs) shows how inputs were arranged for this experiment; at each time 10 spatial locations were sampled uniformly between 0 and 10. The spatial location of pseudo-inputs are regular between 0 and 10. The right hand side shows that when using pseudo-points we are indeed able to achieve substantial performance improvements relative to exact inference by utilising the state-space methodology, while retaining a tight bound. Approximating the log marginal likelihood well for $T = 10^5$ time points ($N = 10^6$ observations total), takes just over 10s in this example.

Grid-with-Missings Fig. 2 (lhs) shows how (pseudo-)inputs were arranged for this experiment for $M_\tau = 10$; the same 50 spatial locations are considered at each time point, but 5 of the observations are dropped at random, for a total of $N_t = 45$ observations per time point – our largest case therefore involves $N = 4.5 \times 10^6$ observations. The right of the figure shows that we able to compute a good approximation to the log marginal likelihood using roughly a third of the computation required by the standard state-space approach to inference. The extent of the improvement here will vary from setting-to-setting depending on how many pseudo-points are required at each point in time to account for the observations – for example, the gains relative to exact state-space inference will be greater if fewer pseudo-points-per-observation can be used.

6. Conclusion

This work shows that for GPs with sum-separable kernels, there exists a natural and efficient manner which to combine pseudo-point and state-space approximations. Preliminary

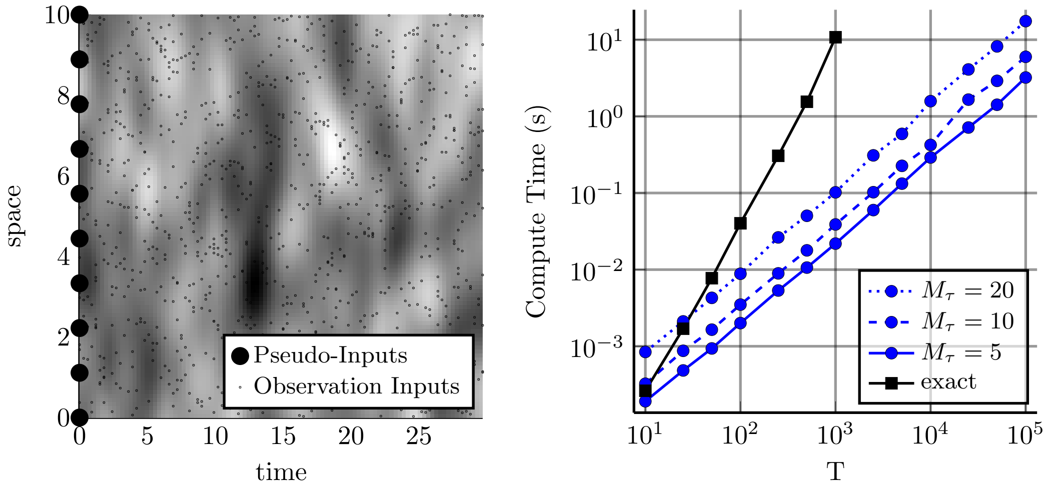

Figure 1: Arbitrary Spatial Locations experiment. Left: Locations of (pseudo-)inputs for $M_\tau = 10$. 10 locations in space chosen randomly at each time point. Right: Time to compute ELBO vs performing exact inference. ELBO tight for $M_\tau = 20$; see Fig. 3.

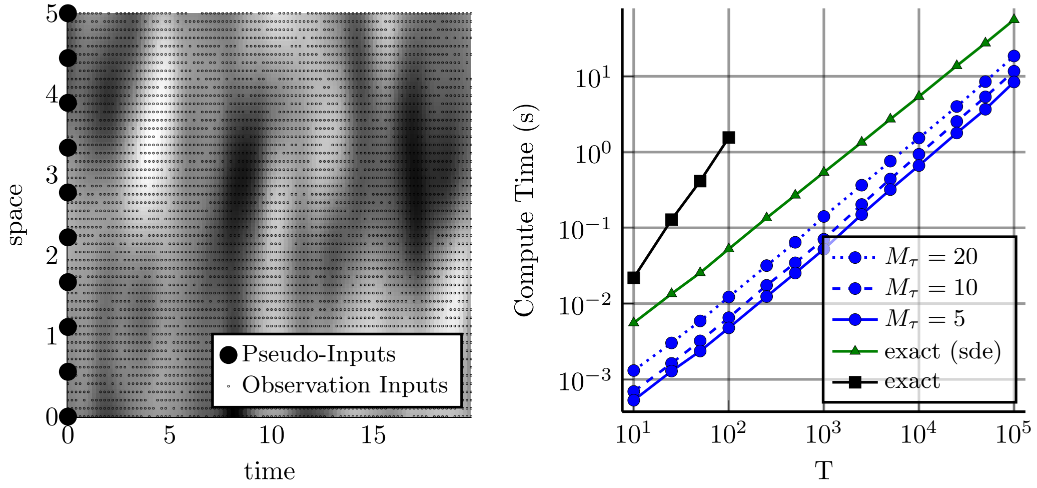

Figure 2: Grid-with-Missings experiment. Left: Locations of (pseudo-)inputs – note the grid structure with 50 observations per time point, of which 5 are missing. Right: Time to compute ELBO vs performing exact inference naively and via state-space methods (*sde*). ELBO tight for $M_\tau = 20$; see Fig. 3.

synthetic experiments show that approach can yield substantial improvements to computational efficiency whilst retaining a high degree of accuracy.

The methods presented can be straightforwardly combined with approximations for non-Gaussian likelihoods, such as those discussed by Wilkinson et al. (2020), Chang et al. (2020), and Ashman et al. (2020). Due to the Markov property there is a natural way to parametrise the approximate posterior (App. F) in such cases, that is analogous to parametrising the filtering distributions. This parametrisation might also be useful when dealing with problems in which mini-batching (Hensman et al., 2013) is desirable.

References

Matthew Ashman, Jonathan So, Will Tebbutt, Vincent Fortuin, Michael Pearce, and Richard E Turner. Sparse Gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.

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

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

David Burt, Carl Edward Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 862–871. PMLR, 2019.

Paul E Chang, William J Wilkinson, Mohammad Emtiyaz Khan, and Arno Solin. Fast variational learning in state-space gaussian process models. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 231–239. PMLR, 2016.

Alexander Grigorievskiy, Neil Lawrence, and Simo Särkkä. Parallelizable sparse inverse formulation Gaussian processes (SpInGP). In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

Jouni Hartikainen, Jaakko Riihimäki, and Simo Särkkä. Sparse spatio-temporal Gaussian processes with general likelihoods. In *International Conference on Artificial Neural Networks*, pages 193–200. Springer, 2011.

James Hensman, Nicolò Fusi, and Neil 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.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 4.4. MIT Press, 2012.

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

Anthony O'Hagan. A Markov property for covariance structures. *Statistics Research Report*, 98:13, 1998.

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

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Yunus Saatçi. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2012.

Simo Särkkä and Ángel F. García-Fernández. Temporal parallelization of Bayesian smoothers. *IEEE Transactions on Automatic Control*, 2020.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse gaussian process regression. Technical report, 2003.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264. MIT Press, 2005.

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

Felipe Tobar. Band-limited Gaussian processes: The sinc kernel. In *Advances in Neural Information Processing Systems*, pages 12749–12759. Curran Associates, Inc., 2019.

William J Wilkinson, Paul E Chang, Michael Riis Andersen, and Arno Solin. State space expectation propagation: Efficient inference schemes for temporal Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Appendix A. Conditional Independence in Separable Gaussian models

This section splits cleanly into four subsections. Everything in here follows from the definitions presented in this paper, a main result from O'Hagan (1998), the properties of the Kronecker product, and basic linear algebra. The second subsection contains the key result.

A.1. Kronecker Structure in the State-Space Prior

In the separable case, there is Kronecker structure that persists through the LGSSM prior. Our starting point is an LGSSM whose transition dynamics are of the form

$$\begin{aligned} \bar{\mathbf{f}}_t &= [\mathbf{I}_N \otimes \mathbf{A}_t] \, \bar{\mathbf{f}}_{t-1} + \varepsilon_t, && (12) \\ \varepsilon_t &\sim \mathcal{N}(0, \mathbf{C}^{\mathbf{r}} \otimes \mathbf{Q}_t) \,, && (13) \\ \bar{\mathbf{f}}_0 &\sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}^{\mathbf{r}} \otimes \mathbf{C}^{\tau}_0) \,, && (14) \end{aligned}$$

where $t \in \{1, ..., T\}$. Here $\mathbf{f}$ might comprise the finite-dimensional marginals over e.g. f and its derivatives at both observation locations and pseudo-inputs. Note that this is the standard form of the LGSSM that separable spatio-temporal GPs are converted to.

First consider the marginal covariance matrix of $\mathbf{f}_t$:

$$\mathbb{V}\left[\bar{\mathbf{f}}_t\right] = [\mathbf{I}_N \otimes \mathbf{A}_t] \, \mathbb{V}\left[\bar{\mathbf{f}}_{t-1}\right] [\mathbf{I}_N \otimes \mathbf{A}_t]^\top + \mathbf{C}^{\mathbf{r}} \otimes \mathbf{Q}_t. \tag{15}$$

Adopting the inductive hypothesis $\mathbb{V}\left[\bar{\mathbf{f}}_{t-1}\right] = \mathbf{C}^{\mathbf{r}} \otimes \mathbf{C}^{\tau}_{t-1}$, it is clear from Eq. (14) that the base case holds:

$$\mathbb{V}\left[\bar{\mathbf{f}}_0\right] = \mathbf{C}^{\mathbf{r}} \otimes \mathbf{C}^{\tau}_0 \tag{16}$$

Furthermore

$$\mathbb{V}\left[\bar{\mathbf{f}}_t\right] = \mathbf{C}^{\mathbf{r}} \otimes \underbrace{\left[\mathbf{A}_t \mathbf{C}^{\tau}_{t-1} \mathbf{A}_t^\top + \mathbf{Q}_t\right]}_{\mathbf{C}^{\tau}_t}, \tag{17}$$

so the Kronecker structure of the marginal covariance at time t is clear, as is the method by which it can be computed.

Now consider the covariance between two time points, $t < t'$ w.l.o.g:

$$\begin{aligned} \mathbb{V}\left[\bar{\mathbf{f}}_{t'}, \bar{\mathbf{f}}_t\right] &= \mathbb{E}\left[\left\{[\mathbf{I}_N \otimes \mathbf{A}_{t'}] \, (\bar{\mathbf{f}}_{t'-1} - \mathbf{m}_{t'-1}) + \varepsilon_{t'}\right\} (\bar{\mathbf{f}}_t - \mathbf{m}_t)^\top\right] \\ &= [\mathbf{I}_N \otimes \mathbf{A}_{t'}] \, \mathbb{V}\left[\bar{\mathbf{f}}_{t'-1}, \bar{\mathbf{f}}_t\right]. && (18) \end{aligned}$$

Assuming the inductive hypothesis

$$\mathbb{V}\left[\bar{\mathbf{f}}_{t'}, \bar{\mathbf{f}}_t\right] = \mathbf{C}^{\mathbf{r}} \otimes \mathbf{C}^{\tau}_{t't} \tag{19}$$

and inducting on t', the base case $(t' = t)$ follows immediately from the result in Eq. (17), and the induction step shows that

$$\mathbb{V}\left[\bar{\mathbf{f}}_{t'}, \bar{\mathbf{f}}_t\right] = \mathbf{C}^{\mathbf{r}} \otimes \underbrace{\left[\mathbf{A}_{t'} \mathbf{C}^{\tau}_{t'-1,t}\right]}_{\mathbf{C}^{\tau}_{t't}}. \tag{20}$$

A.2. Conditional Independence given Current Pseudo-Points

Fortunately, under the assumption of separability it is very straightforward to compute the approximate posterior marginal distribution over $\bar{\mathbf{f}}_{t,n}$ given the approximate posterior marginal distribution over $\bar{\mathbf{u}}_t$. This is a consequence of property derived by O'Hagan (1998), which we generalise here to vector-valued RVs.

First consider the prior joint distribution over $\bar{\mathbf{f}}_{t,n}$, $\bar{\mathbf{u}}_t$, and $\bar{\mathbf{u}}_{t'}$:

$$\begin{bmatrix} \bar{\mathbf{f}}_{t,n} \\ \bar{\mathbf{u}}_t \\ \bar{\mathbf{u}}_{t'} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}^{\bar{\mathbf{f}}}_{t,n} \\ \mathbf{m}^{\bar{\mathbf{u}}}_{t} \\ \mathbf{m}^{\bar{\mathbf{u}}}_{t'} \end{bmatrix}, \begin{bmatrix} \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t} & & \\ \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t} & \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t} & \\ \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} & \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t't} & \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t'} \end{bmatrix} \right). \tag{21}$$

The upper triangle of the covariance matrix is given by symmetry. Note that this covariance matrix follows from the results in the preceding section with

$$\bar{\mathbf{f}}_t := \begin{bmatrix} \bar{\mathbf{f}}_{t,n} \\ \bar{\mathbf{u}}_t \end{bmatrix}. \tag{22}$$

From the usual rules of conditioning for Gaussians, we deduce that the covariance matrix of $\bar{\mathbf{f}}_{t,n}, \bar{\mathbf{u}}_{t'} \mid \bar{\mathbf{u}}_t$ is

$$\begin{bmatrix} \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t} & \\ \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} & \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t'} \end{bmatrix} - \begin{bmatrix} \alpha & \\ \beta & \gamma \end{bmatrix}, \tag{23}$$

where

$$\alpha := \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{f}}_n\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t}\right] \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t}\right]^{-1} \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t}\right] \tag{24}$$

$$\beta := \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t't}\right] \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t}\right]^{-1} \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t}\right] \tag{25}$$

$$\gamma := \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t't}\right] \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{t}\right]^{-1} \left[\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \otimes \mathbf{C}^{\tau}_{tt'}\right]. \tag{26}$$

Observe that the off-diagonal block is

$$\begin{aligned} \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} - \beta &= \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} - \left(\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}} \left(\mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}}\right)^{-1} \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n}\right) \otimes \left(\mathbf{C}^{\tau}_{t't} \left(\mathbf{C}^{\tau}_{t}\right)^{-1} \mathbf{C}^{\tau}_{t}\right) \\ &= \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} - \mathbf{C}^{\mathbf{r}}_{\bar{\mathbf{u}}\bar{\mathbf{f}}_n} \otimes \mathbf{C}^{\tau}_{t't} \\ &= 0, \end{aligned}$$

from which it follows that

$$\bar{\mathbf{f}}_{t,n} \perp\!\!\!\perp \bar{\mathbf{u}}_{t'} \mid \bar{\mathbf{u}}_t \tag{27}$$

for any t'. Moreover

$$\bar{\mathbf{f}}_{t,n} \perp\!\!\!\perp \bar{\mathbf{u}}_{1:T\setminus t} \mid \bar{\mathbf{u}}_t \tag{28}$$

follows from this result and the marginalisation property of Gaussians. Consider the covariance of the conditional distribution $\bar{\mathbf{f}}_{t,n}, \bar{\mathbf{u}}_{1:T\setminus t} \mid \bar{\mathbf{u}}_t$, assuming w.l.o.g. that $t = T$:

$$\begin{bmatrix} \mathbf{C}_{\bar{\mathbf{f}}_{T,n}|\bar{\mathbf{u}}} & \mathbf{C}_{\bar{\mathbf{f}}_{T,n}\bar{\mathbf{u}}_1|\bar{\mathbf{u}}_T} & \mathbf{C}_{\bar{\mathbf{f}}_{T,n}\bar{\mathbf{u}}_2|\bar{\mathbf{u}}_T} & \\ \mathbf{C}_{\bar{\mathbf{u}}_1\bar{\mathbf{f}}_{T,n}|\bar{\mathbf{u}}_T} & \mathbf{C}_{\bar{\mathbf{u}}_1|\bar{\mathbf{u}}_T} & \mathbf{C}_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_2|\bar{\mathbf{u}}_T} & \cdots \\ & \vdots & & \end{bmatrix}. \tag{29}$$

Observe that by the marginalisation property of Gaussians, any of $\mathbf{C}_{\bar{\mathbf{f}}_{T,n}\bar{\mathbf{u}}_1|\bar{\mathbf{u}}_T}$, $\mathbf{C}_{\bar{\mathbf{f}}_{T,n}\bar{\mathbf{u}}_2|\bar{\mathbf{u}}_T}$, etc being non-zero would contradict property Eq. (27). It must therefore hold that they are indeed zero, so property Eq. (28) must hold if Eq. (27) does.

A.3. Block-Diagonal Structure

Furthermore, this conditional independence property implies that

$$\mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}}^{-1} = \begin{bmatrix} \mathbf{0} & \dots & \mathbf{C}_{\bar{\mathbf{f}}_t \bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t}^{-1} & \dots & \mathbf{0} \end{bmatrix}. \tag{30}$$

This is easily proven by considering that, were it not the case, then

$$\mathbb{E}[\mathbf{f}_t \mid \mathbf{u}] \neq \mathbb{E}[\mathbf{f}_t \mid \mathbf{u}_t], \tag{31}$$

contradicting property Eq. (28). From this it follows from repeated application of Eq. (30) that the larger matrix $\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}}^{-1}$ is block-diagonal, and is given by

$$\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}}^{-1} = \begin{bmatrix} \mathbf{C}_{\mathbf{f}_1 \bar{\mathbf{u}}_1} \mathbf{C}_{\bar{\mathbf{u}}_1}^{-1} & & & \\ & \mathbf{C}_{\mathbf{f}_2 \bar{\mathbf{u}}_2} \mathbf{C}_{\bar{\mathbf{u}}_2}^{-1} & & \\ & & \ddots & \\ & & & \mathbf{C}_{\mathbf{f}_T \bar{\mathbf{u}}_T} \mathbf{C}_{\bar{\mathbf{u}}_T}^{-1}. \end{bmatrix} \tag{32}$$

A.4. A Further Conditional Independence Property

Note the Kronecker structure present in Eq. (17): this is the finite-dimensional manifestation of the conditional independence property of separable GPs. It follows immediately from this that

$$\mathbf{f}_t \perp\!\!\!\perp (\bar{\mathbf{u}}_t \backslash \mathbf{u}_t) \mid \mathbf{u}_t, \tag{33}$$

where $\bar{\mathbf{u}}_t \backslash \mathbf{u}_t$ comprises the latent variables in $\bar{\mathbf{u}}_t$. Note this means that the conditional density

$$p(\mathbf{f}_t \,|\, \mathbf{u}_t) := \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} (\mathbf{u}_t - \mathbf{m}_{\mathbf{u}_t}), \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t \mathbf{f}_t}) \tag{34}$$

must be equal to

$$p(\mathbf{f}_t \,|\, \bar{\mathbf{u}}_t) := \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} (\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}), \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t \mathbf{f}_t}). \tag{35}$$

Recalling that $\mathbf{u}_t = \mathbf{H}_{\mathbf{u}_t} \bar{\mathbf{u}}_t$, it follows that

$$\mathbf{m}_{\mathbf{u}_t} = \mathbf{H}_{\mathbf{u}_t} \mathbf{m}_{\bar{\mathbf{u}}_t}, \quad \mathbf{C}_{\mathbf{u}_t} = \mathbf{H}_{\mathbf{u}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \mathbf{H}_{\mathbf{u}_t}^\top. \tag{36}$$

Substituting this into the expression for the mean in Eq. (34) yields the following:

$$\mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} (\mathbf{H}_{\mathbf{u}_t} \bar{\mathbf{u}}_t - \mathbf{H}_{\mathbf{u}} \mathbf{m}_{\bar{\mathbf{u}}_t}) = \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} (\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}),$$

which must hold for *any* value of $\bar{\mathbf{u}}_t$. We deduce that

$$\mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t} = \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t}. \tag{37}$$

Substituting this result into the expressions for the covariance in Eq. (35) shows it to be consistent with the covariance in Eq. (34):

$$\begin{aligned} \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t \mathbf{f}_t} &= \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t \mathbf{f}_t} \\ &= \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \mathbf{H}_{\mathbf{u}_t}^\top \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t \mathbf{f}_t} \\ &= \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t \mathbf{f}_t} \\ &= \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t \mathbf{f}_t}. \end{aligned} \tag{38}$$

Appendix B. Conditional Independence in Sum-Separable GPs

A sum-separable kernel

$$\kappa\big((\mathbf{r},\tau),(\mathbf{r}',\tau')\big) = \sum_{p=1}^{P} \kappa_p^{\mathbf{r}}\big(\mathbf{r},\mathbf{r}'\big)\,\kappa_p^{\tau}\big(\tau,\tau'\big) \tag{39}$$

is not itself separable. However, we can interpret a zero-mean GP f with a sum-separable kernel as a sum over a collection of P independent GPs $f_p \sim \mathcal{GP}(0,\kappa_p)$, where each κ_p *is* separable. The covariance between f and any f_p is

$$\begin{aligned}
\operatorname{cov}\big(f((\mathbf{r},\tau))\,,f_p\big((\mathbf{r}',\tau')\big)\big) &= \mathbb{E}\left[(\sum_{q=1}^{P} f_q((\mathbf{r},\tau)))f_p\big((\mathbf{r}',\tau')\big)\right] \\
&= \sum_{q=1}^{P}\mathbb{E}\big[f_q((\mathbf{r},\tau))\,f_p\big((\mathbf{r}',\tau')\big)\big] \\
&= \mathbb{E}\big[f_p((\mathbf{r},\tau))\,f_p\big((\mathbf{r}',\tau')\big)\big] \\
&= \kappa_p^{\mathbf{r}}\big(\mathbf{r},\mathbf{r}'\big)\,\kappa_p^{\tau}\big(\tau,\tau'\big)\,,
\end{aligned} \tag{40}$$

which is separable. Note that it is trivial to extend this to handle a non-zero mean GP f by adding the mean function to the sum over the f_ps.

This structure persists in the state-space formulation. Consider a set of P state space approximations $\bar{f}_p$, one for to each f_p. The finite dimensional marginals form an LGSSM of the form

$$\bar{\mathbf{f}}_t^p = \mathbf{A}_{p,t}\bar{\mathbf{f}}_{t-1}^p + \varepsilon_{p,t}, \varepsilon_{p,t} \sim \mathcal{N}(0,\mathbf{Q}_{p,t}) \tag{41}$$

Let

$$\bar{\mathbf{f}}_{n,t} := \begin{bmatrix}\bar{\mathbf{f}}_t^1 \\ \vdots \\ \bar{\mathbf{f}}_t^P\end{bmatrix}, \quad \mathbf{H}^\top := \operatorname{vcat}\underbrace{\begin{bmatrix}1 & 1 & \dots & 1 \\ 0 & 0 & & 0 \\ & & \vdots & \\ 0 & 0 & & 0\end{bmatrix}}_{D\times P} \tag{42}$$

where vcat transforms a matrix into a column vector by stacking its columns on top of each other, then

$$\mathbf{f}_{n,t} = \mathbf{H}\bar{\mathbf{f}}_{n,t} \tag{43}$$

$$\mathbf{y}_{n,t} = \mathbf{f}_{n,t} + \eta_{n,t}, \qquad \eta_{n,t} \sim \mathcal{N}(0,\mathbf{S}_{n,t})\,. \tag{44}$$

B.1. Extending Conditional Independence Results to Sum-Separable Processes

To extend the results in App. A.1 first note that each $\bar{\mathbf{f}}_{1:T}^p$ is independent of any other $\bar{\mathbf{f}}_{1:T}^q$ $(p \neq q)$ a priori. Consequently, we can apply the analysis in App. A.1 to each $\bar{\mathbf{f}}_{1:T}^p$ separately and arrive at the associated results in precisely the same way. Similarly by introducing pseudo-points $\bar{\mathbf{u}}^p_{1:T}$ for each p with the same structure as $\bar{\mathbf{u}}_{1:T}$ we may apply the analysis in App. A.2 to obtain

$$\bar{\mathbf{f}}^p_{n,t} \perp\!\!\!\perp \bar{\mathbf{u}}^p_{t'} \mid \bar{\mathbf{u}}^p_t \tag{45}$$

for any t', and

$$\bar{\mathbf{f}}^p_{n,t} \perp\!\!\!\perp \bar{\mathbf{u}}^p_{1:T \setminus t} \mid \bar{\mathbf{u}}^p_t. \tag{46}$$

Letting $\bar{\mathbf{u}}_t$ be the concatenation of all $\bar{\mathbf{u}}^p_t$, we further wish to show that

$$\mathbf{f}_{n,t} \perp\!\!\!\perp \bar{\mathbf{u}}_{1:T \setminus t} \mid \bar{\mathbf{u}}_t. \tag{47}$$

To do this, we first introduce some additional notation:

$$\begin{aligned} \alpha &:= \mathbf{f}_{n,t}, && (48)\\ \beta_p &:= \bar{\mathbf{u}}^p_t, && (49)\\ \gamma_p &:= \bar{\mathbf{u}}^p_{1:T \setminus t}. && (50) \end{aligned}$$

By property 46 we know that

$$\mathbb{V}[\alpha, \gamma_p \mid \beta_p] = \mathbf{0}, \tag{51}$$

so it suffices to show that

$$\mathbb{V}[\alpha, \gamma_p \mid \beta_{1:P}] = \mathbb{V}[\alpha, \gamma_p \mid \beta_p]\,, \tag{52}$$

to which the same marginal consistency argument can be made as was used to show property 28. Since $\beta_p \perp\!\!\!\perp \beta_q$ and $\gamma_p \perp\!\!\!\perp \beta_q$, $p \neq q$, we have that

$$\begin{aligned} \mathbb{V}[\alpha, \gamma_p \mid \beta_{1:P}] &= \mathbf{C}_{\alpha,\gamma_p} - \begin{bmatrix} \mathbf{C}_{\alpha,\beta_1} & \dots & \mathbf{C}_{\alpha,\beta_P} \end{bmatrix} \begin{bmatrix} \mathbf{C}^{-1}_{\beta_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{C}^{-1}_{\beta_P} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\beta_1,\gamma_p} \\ \vdots \\ \mathbf{C}_{\beta_P,\gamma_p} \end{bmatrix} \\ &= \mathbf{C}_{\alpha,\gamma_p} - \sum_{q=1}^{P} \mathbf{C}_{\alpha,\beta_q} \mathbf{C}^{-1}_{\beta_q} \mathbf{C}_{\beta_q,\gamma_p} \\ &= \mathbf{C}_{\alpha,\gamma_p} - \mathbf{C}_{\alpha,\beta_p} \mathbf{C}^{-1}_{\beta_p} \mathbf{C}_{\beta_p,\gamma_p} \\ &= \mathbb{V}[\alpha, \gamma_p \mid \beta_p] \end{aligned} \tag{53}$$

as required. The intuition here is that no additional information is gained by additionally conditioning $\mathbf{f}_{n,t}$ on $\bar{\mathbf{u}}^q_{1:T \setminus t}$, $q \neq p$ because they are independent a priori. From property 47 the other results in App. A follow.

Appendix C. The Optimal Approximate Posterior Under Gaussian Likelihoods

We first derive an expression for $\mathbf{G}_t$. While the form of the reconstruction term presented in Eq. (8) is useful in the general case, in the case that

$$p(\mathbf{y}_{n,t} \,|\, \mathbf{f}_{n,t}) = \mathcal{N}(\mathbf{y}_{n,t}; \mathbf{f}_{n,t}, \mathbf{S}_{n,t})\,, \tag{54}$$

it is most straightforwardly expressed in the vectorised form

$$\mathbb{E}_{q(\mathbf{f}_t)}[\log \mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)] \,, \tag{55}$$

where $\mathbf{y}_t$ and $\mathbf{f}_t$ the vectors formed by stacking the individual $\mathbf{y}_{n,t}$ and $\mathbf{f}_{n,t}$, and $\mathbf{S}_t$ is the diagonal matrix comprising the $\mathbf{S}_{n,t}$.

By the result in App. A.4 we have that $p(\mathbf{f}_t \,|\, \bar{\mathbf{u}}_t) = p(\mathbf{f}_t \,|\, \mathbf{u}_t)$, so

$$\begin{aligned}\mathbb{E}_{q(\mathbf{f}_t)}[\mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)] &= \mathbb{E}_{q(\bar{\mathbf{u}}_t)}\left[\mathbb{E}_{p(\mathbf{f}_t \,|\, \bar{\mathbf{u}}_t)}[\log \mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)]\right] \\ &= \mathbb{E}_{q(\bar{\mathbf{u}}_t)}\left[\mathbb{E}_{p(\mathbf{f}_t \,|\, \mathbf{u}_t)}[\log \mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)]\right] . \end{aligned} \tag{56}$$

Recall that $p(\mathbf{f}_t \,|\, \mathbf{u}_t) = \mathcal{N}(\mathbf{f}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}(\mathbf{u}_t - \mathbf{m}_{\mathbf{u}_t}), \mathbf{Q}_t)$ where $\mathbf{Q}_t := \mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}\mathbf{C}_{\mathbf{u}_t\mathbf{f}_t}$, then expanding the inner expectation yields

$$\mathbb{E}_{q(\mathbf{f}_t)}[\mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)] = \mathbb{E}_{q(\bar{\mathbf{u}}_t)}[\log \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}(\mathbf{u}_t - \mathbf{m}_{\mathbf{u}_t}))] - \frac{1}{2}\mathrm{tr}\big(\mathbf{S}_t^{-1}\mathbf{Q}_t\big) \,. \tag{57}$$

Recall that by definition, $\mathbf{u}_t = \mathbf{H}_t\bar{\mathbf{u}}_t$, from which $\mathbf{m}_{\mathbf{u}_t} = \mathbf{H}_{\mathbf{u}_t}\bar{\mathbf{u}}_t$ follows. Let $\mathbf{B}_t := \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}\mathbf{H}_{\mathbf{u}_t}$ then expanding the remaining expectation yields

$$\mathbb{E}_{q(\mathbf{f}_t)}[\mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{S}_t)] = \log \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{B}_t(\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t})) - \frac{1}{2}\mathrm{tr}\Big(\mathbf{S}_t^{-1}\left[\mathbf{Q}_t + \mathbf{B}_t\hat{\mathbf{C}}_{\bar{\mathbf{u}}_t}\mathbf{B}_t^\top\right]\Big) \,. \tag{58}$$

From this form it is clear that

$$\mathbf{G}_t = -2\nabla_{\hat{\mathbf{C}}_{\bar{\mathbf{u}}_t}}\mathbb{E}_{q(\mathbf{f}_t)}[p(\mathbf{y}_t \,|\, \mathbf{f}_t)] = \mathbf{B}_t^\top\mathbf{S}_t^{-1}\mathbf{B}_t \tag{59}$$

as required.

Similar manipulations that also involve $\mathcal{KL}[q(\bar{\mathbf{u}}) \,\|\, p(\bar{\mathbf{u}})]$ produce the expression for the optimal approximate posterior mean $\hat{\mathbf{m}}_{\bar{\mathbf{u}}}^*$.

Appendix D. State-Space DTC

Eq. (11) introduces the following LGSSM:

$$\tilde{p}(\bar{\mathbf{u}}) := p(\bar{\mathbf{u}}) \,, \quad \tilde{p}(\mathbf{y}_t \,|\, \bar{\mathbf{u}}_t) := \prod_{n=1}^{N_t}\mathcal{N}\big(\mathbf{y}_{n,t}; \mathbf{m}_{\mathbf{f}_{n,t}} + \mathbf{C}_{\mathbf{f}_{n,t}\mathbf{u}_t}\Lambda_{\mathbf{u}_t}\mathbf{H}_{\bar{\mathbf{u}}}(\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}), \mathbf{S}_{n,t}\big) \,. \tag{60}$$

To see that $\tilde{p}(\bar{\mathbf{u}} \,|\, \mathbf{y}) = \mathcal{N}\Big(\bar{\mathbf{u}}; \hat{\mathbf{m}}_{\bar{\mathbf{u}}}^*, [\hat{\Lambda}^*]_{\bar{\mathbf{u}}}^{-1}\Big)$ first note that $\tilde{p}(\mathbf{y}_t \,|\, \bar{\mathbf{u}}_t)$ can be written as

$$\tilde{p}(\mathbf{y}_t \,|\, \bar{\mathbf{u}}_t) = \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}\mathbf{H}_{\mathbf{u}_t}(\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}), \mathbf{S}_t) \,. \tag{61}$$

Furthermore, it follows from this that the conditional distribution $\tilde{p}(\mathbf{y} \,|\, \bar{\mathbf{u}})$ can be expressed succinctly as

$$\tilde{p}(\mathbf{y} \,|\, \bar{\mathbf{u}}) = \mathcal{N}(\mathbf{y}; \mathbf{m}_{\mathbf{f}} + \mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{H}_{\mathbf{u}}(\bar{\mathbf{u}} - \mathbf{m}_{\bar{\mathbf{u}}}), \mathbf{S}) \,, \tag{62}$$

which follows from recalling that $\mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}$ is a block-diagonal matrix comprising T blocks, the t^{th} of which is $\mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}$. From here we can simply apply the usual rules for linear-Gaussian systems (Murphy, 2012) to obtain the desired result.

D.1. Computing Approximate Posterior Marginals

Observe that, as with any $\mathbf{f}_t$ from the training data, the marginal distribution over some $\mathbf{f}_{*t}$ under the approximate posterior only involves $\bar{\mathbf{u}}_t$ as $\mathbf{f}_{*t} \perp\!\!\!\perp \bar{\mathbf{u}}_{\backslash t} \mid \bar{\mathbf{u}}_t$:

$$\begin{aligned}
q(\mathbf{f}_{*t}) &= \mathcal{N}\Big(\mathbf{f}_{*t}; \hat{\mathbf{m}}_{\mathbf{f}_{*t}}, \hat{\mathbf{C}}_{\mathbf{f}_{*t}}\Big) \text{ where} \\
\hat{\mathbf{m}}_{\mathbf{f}_{*t}} &:= \mathbf{m}_{\mathbf{f}_{*t}} + \mathbf{C}_{\mathbf{f}_{*t}\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t} (\hat{\mathbf{m}}_{\bar{\mathbf{u}}_t} - \mathbf{m}_{\bar{\mathbf{u}}_t}), \\
\hat{\mathbf{C}}_{\mathbf{f}_{*t}} &:= \mathbf{C}_{\mathbf{f}_{*t}} - \mathbf{C}_{\mathbf{f}_{*t}\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t} \Big[\mathbf{C}_{\bar{\mathbf{u}}_t} - [\hat{\Lambda}^*_{\bar{\mathbf{u}}_t}]^{-1}\Big] \mathbf{H}_{\mathbf{u}_t}^\top \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t\mathbf{f}_{*t}}.
\end{aligned}$$

Performing smoothing in the approximate model provides $\hat{\mathbf{m}}_{\bar{\mathbf{u}}_t}$ and $[\hat{\Lambda}^*_{\bar{\mathbf{u}}_t}]^{-1}$, from which the optimal approximate posterior marginals are straightforwardly obtained via the above.

D.2. Compute the Saturated Bound

Recall the standard saturated bound introduced by Titsias (2009), that is obtained at the optimal approximate posterior:

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y}; \mathbf{m}_\mathbf{f}, \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}} + \mathbf{S}) - \frac{1}{2} \mathrm{tr}\big(\mathbf{S}^{-1} [\mathbf{C}_\mathbf{f} - \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}}]\big) \,. \tag{63}$$

To see how this can be computed efficiently using the approximate model, first note that the approximate model definition in 60 implies that

$$\begin{aligned}
\tilde{p}(\mathbf{y}) &= \mathcal{N}\Big(\mathbf{y}; \mathbf{m}_\mathbf{f}, \mathbf{C}_{\mathbf{f}\mathbf{u}} \Lambda_\mathbf{u} \mathbf{H}_\mathbf{u} \mathbf{C}_{\bar{\mathbf{u}}} \mathbf{H}_\mathbf{u}^\top \Lambda_\mathbf{u} \mathbf{C}_{\mathbf{u}\mathbf{f}} + \mathbf{S}\Big) \\
&= \mathcal{N}(\mathbf{y}; \mathbf{m}_\mathbf{f}, \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}} + \mathbf{S}) &\quad (64) \\
&= \mathcal{N}(\mathbf{y}; \mathbf{m}_\mathbf{f}, \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}} + \mathbf{S}) \,, &\quad (65)
\end{aligned}$$

which is precisely the first term of Eq. (63). The second equality follows by noticing that the action of the linear maps $\mathbf{C}_{\mathbf{f}\mathbf{u}} \Lambda_\mathbf{u} \mathbf{H}_\mathbf{u}$ and $\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}}$ are equal, which follows from the conditional independence property in App. A.4 and the block-diagonal structure of $\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}}$ derived in App. A.3. This means that to compute the first term of the optimal ELBO is simply the log marginal likelihood of the approximate model, which can be computed efficiently via filtering.

By application of the same two properties, the trace term can be written as

$$\begin{aligned}
\frac{1}{2} \sum_{t=1}^T \mathrm{tr}\big(\mathbf{S}_t^{-1} [\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t\mathbf{f}_t}]\big) &= \frac{1}{2} \sum_{t=1}^T \mathrm{tr}\big(\mathbf{S}_t^{-1} [\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t\mathbf{f}_t}\big) \\
&= \frac{1}{2} \mathrm{tr}\big(\mathbf{S}_t^{-1} [\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t\mathbf{f}_t}\big) \\
&= \frac{1}{2} \mathrm{tr}\Big(\mathbf{S}_t^{-1} [\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t} \Lambda_{\mathbf{u}_t} \mathbf{H}_{\mathbf{u}_t} \mathbf{C}_{\bar{\mathbf{u}}_t} \mathbf{H}_{\mathbf{u}_t}^\top \Lambda_{\mathbf{u}_t} \mathbf{C}_{\mathbf{u}_t\mathbf{f}_t}]\Big) \,.
\end{aligned} \tag{66}$$

These quantities are all straightforward to compute by running the approximate model forwards through time and simply computing the marginal statistics.

Appendix E. FITC

Consider the approximate model employed by FITC:

$$\tilde{p}(\mathbf{y} \mid \bar{\mathbf{u}}) := \mathcal{N}(\mathbf{y}; \mathbf{m}_{\mathbf{f}} + \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} (\bar{\mathbf{u}} - \mathbf{m}_{\bar{\mathbf{u}}}), \operatorname{diag}(\mathbf{C}_{\mathbf{f}} - \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}} \mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}}) + \mathbf{S}) \tag{67}$$

We know that in our separable setting, $\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}} \Lambda_{\bar{\mathbf{u}}}$ is block-diagonal from App. A.3. This means that the t^{th} block on the diagonal of the conditional covariance matrix is

$$\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t \mathbf{f}_t}, \tag{68}$$

and the entire conditional distribution factorises as follows:

$$\tilde{p}(\mathbf{y} \mid \bar{\mathbf{u}}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{y}_t; \mathbf{m}_{\mathbf{f}_t} + \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} (\bar{\mathbf{u}}_t - \mathbf{m}_{\bar{\mathbf{u}}_t}), \operatorname{diag}(\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t \bar{\mathbf{u}}_t} \Lambda_{\bar{\mathbf{u}}_t} \mathbf{C}_{\bar{\mathbf{u}}_t \mathbf{f}_t}) + \mathbf{S}_t)\,. \tag{69}$$

By comparing this with equation 5 of (Hartikainen et al., 2011), and letting the likelihood in that equation $p(\mathbf{y}_k \mid \mathbf{x}_k) = \mathcal{N}(\mathbf{y}_k; [\mathbf{I}_N \otimes \mathbf{H}]\mathbf{x}_k, \mathbf{S}_t)$, the correspondence is clear.

Appendix F. Inference Under Non-Gaussian Likelihoods

- While not explored empirically here, the Markov property holds regardless the form of the likelihood.
- Optimal approximate posterior can be obtained by optimising “filtering distributions”
- Employ e.g. CCVI to optimse.

Appendix G. Additional Experimental Details

The kernel of the GP used in all experiments is

$$\kappa\big((\mathbf{r}, \tau), (\mathbf{r}', \tau')\big) = \kappa^{\mathbf{r}}\big(\mathbf{r}, \mathbf{r}'\big)\, \kappa^{\tau}\big(\tau, \tau'\big) \tag{70}$$

where $\kappa^{\mathbf{r}}$ is an Exponentiated Quadratic kernel with length scale 0.9 and amplitude 0.92, and κ^{τ} is a Matern-3/2 kernel with length scale 1.2. The particular values of the length scales / amplitudes is of little importance to the proof-of-concept experiments presented in this work – the motivation for their choice was to avoid 1 for the sake of catching bugs in our code (it generally being a good idea to avoid setting numbers to $\mathbb{R}$'s multiplicative identity element when debugging).

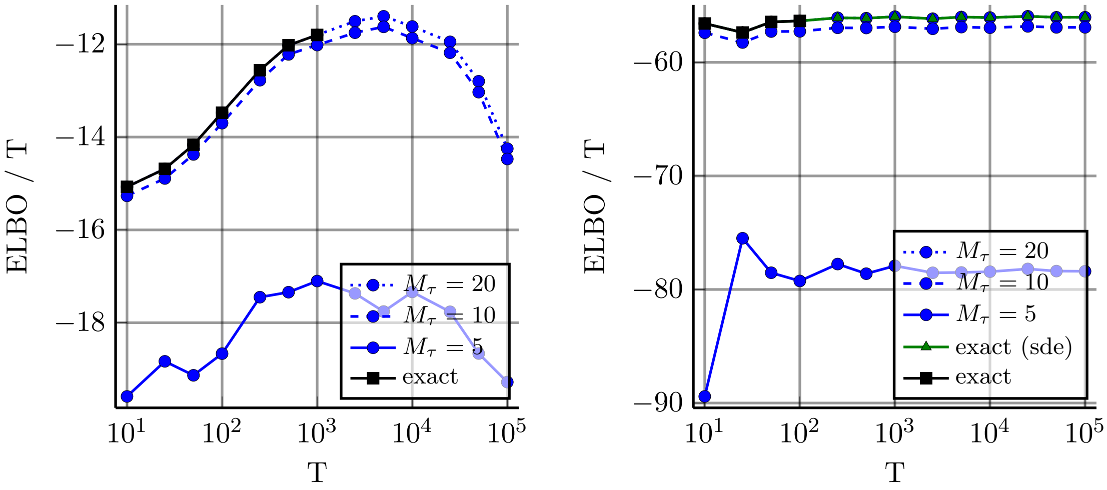

Figure 3: The ELBO obtained vs the exact log marginal likelihood. The bound appears reasonably tight when $M_\tau = 10$ are used per time point, and very tight for $M_\tau = 20$. $M_\tau = 5$ is clearly insufficient.