Understanding Local Linearisation in Variational Gaussian Process State Space Models

Talay M Cheema [1]

Abstract

We describe variational inference approaches in Gaussian process state space models in terms of local linearisations of the approximate posterior function. Most previous approaches have either assumed independence between the posterior dynamics and latent states (the mean-field (MF) approximation), or optimised free parameters for both, leading to limited scalability. We use our framework to prove that (i) there is a theoretical imperative to use non-MF approaches, to avoid excessive bias in the process noise hyperparameter estimate, and (ii) we can parameterise only the posterior dynamics without any less of performance. Our approach suggests further approximations, based on the existing rich literature on filtering and smoothing for nonlinear systems, and unifies approaches for discrete and continuous time models.

1. Introduction

Much time series data of engineering or scientific interest can be well-described by dynamical systems models; in particular, many physical and biological systems are described by mechanistic differential equation models. For tasks such as predictive control or experiment design, good quality uncertainty estimates over the dynamics or predictions are needed, which suggests a Bayesian approach.

Where the system dynamics are fairly well known, there exists a wealth of methods for predictions with uncertainty. However, in some important application domains, such as model based reinforcement learning or synthetic biology, the system dynamics are not well determined a priori and must be determined from noisy measurements. Using Gaussian process (GP) priors for the dynamics provides a non-parametric approach which can give meaningful uncertainty estimates in the low data regime, yet scale up as we make collect further measurements.

There are two major classes of dynamical models: autoregressive (AR) models, which model transitions directly in the space of observations, and state space models (SSMs), which model transitions in a latent space. We focus on GPSSMs since they offer advantages in the common cases of partial observations and redundant (i.e. high dimensional) measurements.

Most work for deterministic approximate inference in GPSSMs in the past has enforced models in discrete time with factorised variational distributions, and has the number of parameters scaling with the length of the time series. In this work we give a common treatment to both continuous and discrete time cases. We show that there is a theoretical imperative to use correlated variational distributions, echoing the empirical evidence of (Ialongo et al., 2019), but we show the free parameters of that model are superfluous. We propose a family of methods inspired by a relaxation of the optimal method (in the sense of the variational objective), which we categorise by their implicit linearisation of the approximate posterior function.

The rest of the paper is organised as follows. In Section 2, we review relevant background material on GPSSMs, the filtering and smoothing theory, and related work. In Section 3 we detail the theoretical results and the inference methods which follow. We give an initial experimental investigation of these methods in Section 4.

2. Background

2.1. State space models

We consider principally discrete time state space models of the following form

$$x_{t+1} = f(x_t, u_t) + \kappa_t$$
$$y_t = g(x_t) + \rho_t$$
$$\kappa_t \sim \mathcal{N}(0, Q), \rho_t \sim \mathcal{N}(0, R) x_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$$

with the continuous time analogue for the latents

$$dx_t = f(x_t, u_t)dt + d\beta_t$$

[1]Department of Engineering University of Cambridge, UK. Correspondence to: Talay M Cheema <tmc49@cam.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

where β is Brownian motion with diffusion matrix Q, and the rest of the model is unchanged from the discrete time case. We seek to model the functions f and g, with each $x_t \in \mathbb{R}^D, y \in \mathbb{R}^\Delta$, $u_t \in \mathbb{R}^{D_c}$. We have observations for $y_{t_i}, i \in \{1:T\}$; x is not observed, and u is a sequence of deterministic control inputs. We hereafter suppress reference to u in the text without loss of generality.

Autoregressive models (in discrete time) assume that the states are observed directly, and may allow f to depend on multiple preceding states. However, we are often in the regime of partial observations ($\Delta < D$), for which we may need many preceding states to form a good prediction, or in the regime of redundant measurements ($\Delta >> D$). In either case, autoregressive models' transition functions will tend to require much higher dimensional input spaces than state space models.

2.2. Filtering and smoothing

In the special case where f and g are both affine, the exact state posterior is Guass-Markov distributed, and can be efficiently calculated using the Kalman filter (states given current and preceding measurements) or Kalman smoother (full state posterior). The linear transition and emission functions can be estimated in a maximum likelihood or maximum a posteriori fashion using the expectation maximisation (EM) algorithm.

In the general nonlinear case, the exact posterior is no longer Gaussian. Kalman filter extensions are a popular family of methods for nonlinear state estimation which use the Kalman filter/smoother on an approximate time-varying linear system constructed by explicitly linearising the dynamics. Some examples are given in Table 1 along with the moment matching (MM) or assumed density filtering (ADF) approximation; see (Särkkä, 2013; Särkkä & Solin, 2019) for further details. Approximate EM methods can be extended to this case (Ghahramani & Roweis, 1999).

Table 1. Approximations for Kalman-like filter/smoothers. Abbreviations: EKF – Extended Kalman Filter, SLF – Statistically Linearised Filter, ADF – Assumed Density Filter, $E_t[\cdot] = \int \cdot q(x_t = x)dx$.

METHOD	A_t	b_t	$\tilde{Q}_t$
EKF	$\frac{\partial f}{\partial x}\Big\vert \mu_t$	$f(\mu_t) - A_t\mu_t$	Q
SLF	$\mathbb{E}_t\left[\frac{\partial f}{\partial x}\right]$	$\mathbb{E}_t[f(x)] - A_t\mu_t$	Q
ADF	0	$\mathbb{E}_t[f(x)]$	$\mathrm{Cov}_t[f(x)]$

2.3. Bayesian non-parametrics

In some applications, we require uncertainty estimates for downstream tasks, and gather data incrementally, such as in active learning or model-based reinforcement learning. Placing a Gaussian process prior on f and carrying out Bayesian inference fulfills the desiderata of these applications, insofar as the model complexity scales with the size of the dataset.

In this case, to avoid difficulties with non-identifiability, we enforce linear measurements, $g(x) = Cx$. We lose no generality in the sense that any nonlinearity in g can be transferred to f in exchange for a possible increase in D (Frigola-Alcalde, 2014). To be precise, usually f is chosen to have a prior independent across output dimensions, i.e.

$$p(f) = \prod_{d=1}^{D} p(f_d) \qquad p(f_d) = \mathcal{GP}(m_d(\cdot), k_d(\cdot,\cdot))$$

where $m_d : \mathbb{R}^D \to \mathbb{R}$ and $k_d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ are the mean and covariance function of the dth GP.

2.4. Variational approximations

Variational approximations for Gaussian processes often rely on a pseudo-point approximation, in which we introduce inducing inputs $z \in \mathbb{R}^D$ and corresponding inducing outputs $v \in \mathbb{R}^D$ (Titsias, 2009). We treat the z as deterministic variational parameters, and include v in our variational distribution. Conceptually, the pair (z, v) stand in for input-output pairs of the GP, allowing us to reframe approximate inference as surrogate GP regression on the inducing points. This is particularly important when the inputs are latent variables (Damianou et al., 2016), which includes our case. The variational approximation for f is constructed from the inducing outputs using the prior conditional:

$$q(f) = q(f, v) = q(v)p(f|v)$$

where the first equality follows since $v = f(z)$ is a finite index subset of f.

The earliest Bayesian treatments of GPSSMs used computationally intensive Markov chain Monte Carlo (MCMC) schemes (Frigola-Alcalde, 2014). In order to avoid this, subsequent efforts have focused on deterministic approximations, principally variational inference (VI), within which the objective function for training is

$$\begin{aligned}\mathcal{F} &= \int q(x, f) \log \frac{p(y, x, f)}{q(x, f)} dxdf \\ &= \int q(x) \log p(y|x) dx - D_{KL}(q(x, f)||p(x, f)) \leq \log p(y|x)\end{aligned}$$

with equality iff $D_{KL}(q(x, f)||p(x, f)) = 0$, where D_{KL} is the KL divergence. We maximise the variational objective with respect to the hyperparameters and the parameters of the variational distribution, and each upward step does some combination of (1) increasing the likelihood of the hyperparameters and (2) bringing the approximate posterior closer to the true posterior. See (Bui, 2017) for an alternative treatment by power expectation propagation.

Since computation of the exact posterior is intractable, we limit q to an approximating family which is computationally convenient; however, this comes at a cost: the maximum

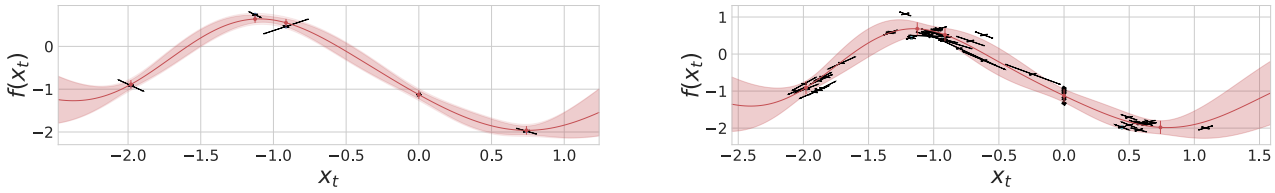

Figure 1. Schematic representation of the two methods. *(Left)* MF approach, where $q(f)$ has been set optimally according to $q(x)$. *(Right)* Locally linearised approach, where $q(x|f)$ has been generated from some samples of v according to a statistically linearised filter. In red is the mean of $q(f)$ with one standard deviation uncertainty shaded; red dots and error bars depict inducing outputs. In black are the pairs (μ_t, μ_{t+1}) from $q(x|f)$, with arrows showing one standard deviation either side in the directions of the eigenvectors. In the MF case, $q(f)$ is approximately passes through the points, aligned with one principal direction, and we have T means and covariances to optimise. In the linearised case, the principal directions are set according to the function, and we have M inducing outputs to optimise.

of the objective may be biased away from the maximum of the likelihood. This bias has been shown to be particularly severe when strong correlations present in the true posterior which are enforced absent in the approximation (Turner & Sahani, 2011). We will now consider this in more detail.

3. Theoretical results and proposed method

3.1. Optimal process noise

Most of the previous work on GPSSMs has used the MF assumption; that is, that the states and dynamics are uncorrelated in the variational distribution.

$$q(x, f) = q(x)q(f) = q(x)q(v)p(f|v)$$

If we take $q(x)$ to be Gauss-Markov, then $q(v)$ is optimally Gaussian and available in closed form (Frigola-Alcalde, 2014). However, it is clear from the prior that correlations between x and f are significant (x is generated by f). Recent applications of correlated approximations (Doerr et al., 2018; Ialongo et al., 2019) have shown empirically that the effect is signicant. We now show this from the effect on the process noise hyperparameter.

Consider the general Gauss-Markov approximations

$$q(x_{t+1}|x_t, f) = \mathcal{N}(x_{t+1}|h_t[f](x_t), \tilde{Q}_t[f])$$
$$\text{or} \qquad dx_t = h_t[f](x_t)dt + d\tilde{\beta}$$

where $\tilde{\beta}$ is Brownian motion with diffusion matrix $\tilde{Q}$, h_t is an affine function of x which depends on t and f, and $\tilde{Q}_t$ is a positive definite matrix which depends on t and f.

$$\mathcal{F} = \int \log p(y|x) dq(x) - D_{KL}(q(v|z)||p(v|z)) - \int \underbrace{\int D_{KL}(q(x)||p(x|f))\, p(f|v,z)df}_{-\mathcal{F}_{x|f}}\, q(v)dv$$

Since Q appears only in the middle term, we optimise to get in the discrete time case

$$Q = \frac{1}{T}\sum_{t=1}^{T}(\tilde{Q}_{t-1} + \mathbb{E}[(h_{t-1}[f](x) - f(x))(h_{t-1}[f](x) - f(x))^\top])$$

where the expectation is over $q(x, f)$. In continuous time we get an analagous integral in place of the sum.

The first term is the average process noise of the variational approximation; relative to this the process noise is biased larger by the approximation error covariance of the linearisation. The residual contribution of $\mathcal{F}_{x|f}$ when Q is set optimally is $-\frac{1}{2}(\sum_t \log|\tilde{Q}| - T\log|Q|)$, thus this term pushes h to approximate f. On the other hand, the observation term pushes h to yield states which are close to the observations. Thus this extra term allows the approximate the process noise to grow to absorb error due to model mismatch, which can be desirable for behaviour far from the prior (e.g., much faster than the dynamics of interest), or to allow flexibility during training.

Clearly, in the MF case, where h has no dependence on f, the error term will be larger in general, leading to inflated process noise, as seen empirically in (Ialongo et al., 2019) (though note there, h is also nonlinear, which gives more flexibility, but $q(x)$ is no longer Gaussian, and must be sampled also).

Constructing a meaningful linearisation with full dependence on f is not straightforward, so we settle for dependence only on the inducing outputs v.

3.2. Optimal smoothing

Most previous approaches have required full parameterisation of $q(x|f)$ ($O(TD^2)$ parameters), although (Eleftheriadis et al., 2017) mitigated this by using a bi-directional RNN recognition network in the MF case. We now show that there is a more natural, non-parametric way to amortise inference in the correlated case.

We can see from the results of (Archambeau et al., 2007; Duncker et al., 2019) that $\mathcal{F}_{x|v} = \int \mathcal{F}_{x|f} p(f|v)df$ can be maximised by an iterative smoothing algorithm for any v. We can mimic this result for the discrete time case as follows. Augment the variational objective for each v with Lagrange terms to enforce the marginals and transition parameters are

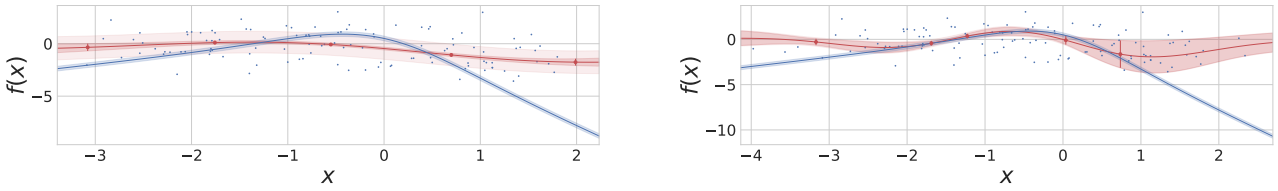

Figure 2. Results from the kink experiment. *(Left)* The fit for the MF case. *(Right)* The fit for the correlated case, with $q(x|f)$ constructed using statistically linearised smoothing. In blue is the ground truth (line) with one standard deviation of groundtruth process noise (shaded). In red is $q(f)$ (dark line) with one standard deviation of function uncertainty (shaded, dark), and of function uncertainty and process noise estimate (shaded, light); dots and error bars show inducing points. The faint blue dots show the observed pairs (y_t, y_{t+1}). The correlated approach has much better calibrated posterior uncertainty.

consistent, i.e.

$$\mathcal{L}_v = \mathcal{F}_v + \sum_{t=1}^{T} \lambda_t^\top (\mathbb{E}_{t+1}[x] - \mathbb{E}_t[A_t x + b_t])$$
$$+ \sum_{t=1}^{T} \mathrm{tr}(L_t(\mathbb{E}_{t+1}[xx^\top] - \mathbb{E}_t[(A_t x + +b_t)(A_t x + b_t)^\top] - \tilde{Q}_t))$$

wherein $E_t[\cdot] = \int \cdot q(x_t = x)dx$. Then forming an explicit recursion for the augmented parameter set, we find

$$\tilde{Q}_t^{-1} = (2L_t + Q^{-1}), A_t = \tilde{Q}_t Q^{-1} \mathbb{E}_t\left[\frac{\partial \mu_f(x)}{\partial x}\right]$$
$$b_t = \tilde{Q}_t Q^{-1} \mathbb{E}_t[\mu_f(x)] - A_t \mu_t - \tilde{Q}_t \lambda_t$$
$$L_{t-1} = A_t L_t A_t^\top - \frac{\partial \mathcal{F}_v}{\partial \Sigma_t}, \lambda_{t-1} = A_t^\top(\lambda_t + 2L_t b_t) + 2\frac{\partial \mathcal{F}_v}{\partial \mu_t}$$

which should be initialised with $L_t = 0, \lambda_t = 0$. Note that the part of this which depends only on preceding values is identical to the SLF.

This shows that no additional free parameters are needed, other than the initial state prior and $q(v)$'s parameters. In discrete time, this reduces the number of parameters from $O(TD^2)$ to $O(M^2D)$. The importance of the reduction of parameters is greater in continuous time, where the number of latent states may be arbitrarily larger than the number of observations T.

However, this method comes with substantial computational cost (due to the derivatives required) and may be difficult to use in practice, due to issues such as local optima. A sensible relaxation is to replace this optimal smoothing algorithm with some other, e.g. from Table 1.

4. Experiments

One-dimensional illustration We consider discrete time dynamics generated by a 1D 'kink' function (see Figure 2), which generates oscillating trajectories with high observation noise and modest process noise ($Q = 0.05^2, R = 0.8$).

$$f(x) = 0.8 + (x + 0.2)\left(1 - \frac{5}{1 + \exp(-2x)}\right)$$

Learning is hard here, due to the large observation noise. We fix the observation model to the groundtruth, and see that the MF model attributes all the uncertainty to process noise, whereas the correlated model attributes most of the uncertainty to f.

Ball-beam We use the ball-beam dataset from (De Moor, 2006).[1] We train on the first half of the observations using $D = 4, M = 100$, and evaluate the (VI-optimal) analytic one-step prediction. The observation noise is fixed to focus on the latent part of the model, where the differences lie. The SLF approach shows promising performance, with comparable RMSE but better NLPP.

Table 2. Training objective and one step predictive test evaluation for the two methods. Lower is better.

METHOD	TRAIN NVFE	RMSE	NLPP
MF	-0.2381	0.1341	-0.8676
SLF	-4.6209	0.1355	-1.0795

5. Conclusions and further work

We have shown that ignoring key correlations in the variational approximation biases the process noise larger, reducing predictive power, and that we can introduce correlation whilst also improving the scalability. We show empirical evidence that such non-parametric state inference methods can perform comparable to a fully parameterised MF method. Future work would include a more thorough empirical investigation, and exploring non-parametric methods which break the linearisation constraint, e.g. the particle filter/smoother.

[1] For reproducibility, the dataset is 'Data of the ball-and-beam setup in STADIUS', section 'Mechanical Systems', code [96-004]. Two states are initialised equal to y and u, and the others as their first difference. We use 10 posterior samples for the SLF method, and optimise with Adam.

References

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. Gaussian process approximations of stochastic differential equations. In Lawrence, N. D., Schwaighofer, A., and Candela, J. Q. (eds.), *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pp. 1–16, Bletchley Park, UK, June 2007. PMLR.

Bui, T. D. *Efficient Deterministic Approximate Bayesian Inference for Gaussian Process Models*. PhD thesis, University of Cambridge, 2017.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17(42):1–62, 2016.

De Moor, B. L. R. *DaISy: Database for the Identification of Systems*. http://homes.esat.kuleuven.be/˜smc/daisy/. Department of Electrical Engineering, ESAT/STADIUS, KU Leuven, Belgium, 2006.

Doerr, A., Daniel, C., Schiegg, M., Duy, N.-T., Schaal, S., Toussaint, M., and Sebastian, T. Probabilistic Recurrent State-Space Models. In *International Conference on Machine Learning*, pp. 1280–1289. PMLR, July 2018.

Duncker, L., Bohner, G., Boussard, J., and Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1726–1734. PMLR, June 2019.

Eleftheriadis, S., Nicholson, T., Deisenroth, M., and Hensman, J. Identification of Gaussian Process State Space Models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5309–5319. Curran Associates, Inc., 2017.

Frigola-Alcalde, R. *Bayesian Time Series Learning with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

Ghahramani, Z. and Roweis, S. T. Learning nonlinear dynamical systems using an EM algorithm. In *Advances in Neural Information Processing Systems 12*, 1999.

Ialongo, A. D., Van Der Wilk, M., Hensman, J., and Rasmussen, C. E. Overcoming mean-field approximations in recurrent Gaussian process models. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2931–2940. PMLR, June 2019.

Särkkä, S. *Bayesian Filtering and Smoothing*. Cambridge University Press, Cambridge, United Kingdom, 2013. ISBN 978-1-139-34420-3 978-1-107-61928-9 978-1-107-03065-7.

Särkkä, S. and Solin, A. *Applied Stochastic Differential Equations*. Number 10 in Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge ; New York, NY, 2019. ISBN 978-1-316-51008-7 978-1-316-64946-6.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, April 2009. PMLR.

Turner, R. E. and Sahani, M. Two problems with variational expectation maximisation for time series models. In Barber, D., Cemgil, A. T., and Chiappa, S. (eds.), *Bayesian Time Series Models*. Cambridge University Press, Cambridge, UK ; New York, 2011. ISBN 978-0-521-19676-5.