Contrasting Discrete and Continuous Time Methods for Bayesian System Identification

Talay M Cheema [1] **Carl Edward Rasmussen** [1]

Abstract

In recent years, there has been considerable interest in embedding continuous time methods in machine learning algorithms. In system identification, the task is to learn a dynamical model from incomplete observation data, and when prior knowledge is in continuous time – for example, mechanistic differential equation models – it seems natural to use continuous time models for learning. Yet when learning flexible, nonlinear, probabilistic dynamics models, most previous work has focused on discrete time models to avoid computational, numerical, and mathematical difficulties. In this work we show, with the aid of small-scale examples, that this mismatch between model and data generating process can be consequential under certain circumstances, and we discuss possible modifications to discrete time models which may better suit them to handling data generated by continuous time processes.

1. Introduction

State space models (SSMs), such as stochastic differential equation (SDE) models, are widely used in engineering and scientific application to describe the behaviour of systems. The idea is to model the evaluation of a state $x_t \in \mathbb{R}^D$ over time $t \in \mathbb{R}$ in a Markovian way, and model observations $\{y_i \in \mathbb{R}^\Delta\}_{i=1}^n$ as a measurement of the state at some particular times $\{t_i\}_{i=1}^n$.

$$x_{t+\delta} = f_{DT}(x_t) + L_{DT}\kappa_t \quad \text{or} \quad dx_t = f_{CT}(x_t)\,dt + L_{CT}\,d\beta_t \tag{1}$$

$$y_i = g(x_{t_i}) + \rho_i \tag{2}$$

Each $\kappa_t \sim \mathcal{N}(0, I)$ independently, each $\rho_i \sim \mathcal{N}(0, R)$ independently, and β_t is standard D-dimensional Brownian motion. The process noise terms κ_t, β_t are included for regularisation: the process noise variance can absorb, for example, model mismatch, and regularise the learning problem (Leander et al., 2014).

For a number of important tasks, such as model predictive control or experiment design, a key requirement is to be able to propagate uncertainty about the dynamics forward through time, motivating a Bayesian approach. If the the system is very well understood, this can be done in a parametric fashion (Särkkä, 2013; Särkkä & Solin, 2019). But for more flexible models, a powerful option is to use Gaussian process (GP) models, yielding a non-parametric method for learning dynamics which can give meaningful uncertainty estimates in the low data regime, yet scale up as further measurements are collected. This is the approach we focus on here. To limit non-identifiability, we consider g fixed.

GPSSMs, in common with many machine learning methods, have mainly been used in discrete time (DT), whether using a pre-trained GP (Deisenroth & Rasmussen, 2011) or for full system identification (Frigola-Alcalde, 2014; Eleftheriadis et al., 2017; Bui, 2017; Ialongo et al., 2019; Doerr et al., 2018; Curi et al., 2020). Yet usually the data considered is generated by a mechanical or chemical system, traditionally modelled in continuous time (CT). Of course, there is no physical reason for observation times to be constrained to a regular grid, and in some applications, particularly in biological systems, data is highly irregularly sampled. This has motivated recent work on CT GPSSMs by Duncker et al., 2019, based on work by Archambeau et al., 2007. This comes with several challenges: the number of trainable parameters becomes very large, the learning algorithm takes much longer to run, and more care is needed to avoid catastrophic numerical errors.

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

Workshop on Continuous Time Methods in Machine Learning at the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, 2022.

This is typical of the disadvantages of CT methods in machine learning – there are numerous challenges to overcome in the implementation, but they can bring compelling advantages. For instance, in the case of neural ordinary differential equations (ODEs), there are substantial memory savings (Chen et al., 2018). Intuitively, when modelling using Bayesian approach, it should be an advantage to select prior distributions to match our prior knowledge of the process we are trying to model, but it remains an open question whether this brings a significant advantage in system identification. We explore this as follows.

Firstly, in Section 2, we examine the qualitative differences in placing a GP prior on CT dynamics compared to DT dynamics by considering the types of trajectories $x(t)$ sampled form such systems.

Secondly, in Section 3, we examine the impact of this in a full, toy, learning problem. We see empirically that there are only small gains for the CT model in low noise, but much more substantial gains in the face of high noise.

2. Contrasting priors

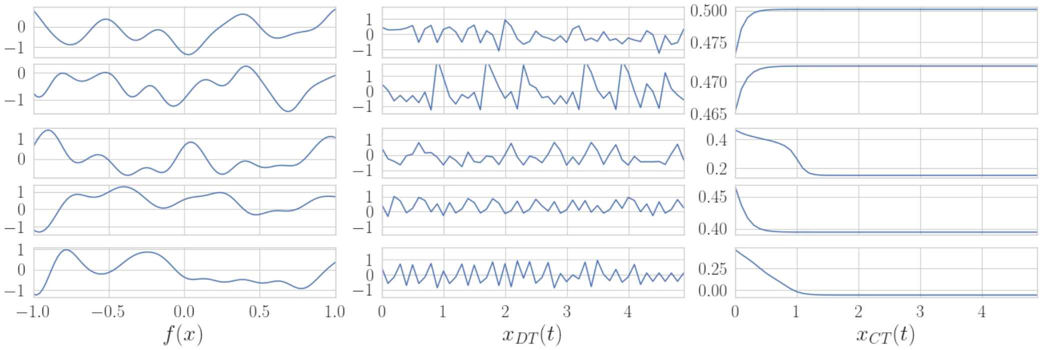

Figure 1: Prior samples from GPSSMs (without noise). Each row is one sample; f is on the left, and $x(t)$ for the DT (centre) and CT (right) cases are shown.

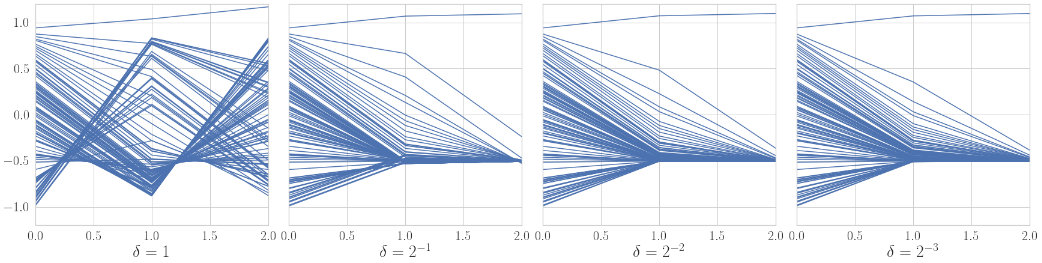

Figure 2: Trajectories generated by a discrete time system with transition function $x + \delta^{-1} f_{CT}(x)$ for decreasing δ. As it converges, the trajectories exhibit the characteristic qualities of CT systems (non-intersecting trajectories, all points converging to equilibria in 1D).

We sample both f_{DT} and f_{CT} identically from a GP, and note that there are several qualitative differences between the sampled trajectories.

Firstly, we plot sampled trajectories for a one-dimensional system in Figure 1; in 1D, CT systems can only be simple systems which are drawn into equilibria. But in DT, the dynamics can be mcuh richer, with stable quasi-periodic behaviour.

Secondly, more generally, the trajectories from CT and DT systems look different. Ignoring the process noise, CT dynamics denerate trajectories which do not cross over, whereas in DT they may do so (Hirsch et al., 2004). Note that if the SDE is interpreted in Itô sense, then it can be shown that if we set $f_{DT}(x) = x + \delta f_{CT}(x)$, and $L_{DT} = \delta L_{CT}$ then the limit as $\delta \to 0$ of the DT system is the CT system, and we see that the trajectories stop intersecting as we make δ smaller in Figure 2. But this does not follow if we merely iterate $f_{DT}(x_t)$; DT systems often exhibit more complex and challenging to analyse behaviour than CT dynamics. It is well-known in the dynamical systems literature that these more challenging behaviours arise when f_{DT} is not diffeomorphic (that is, smooth and bijective with smooth inverse) (Losert & Akin, 1983). In our case, f_{DT} is a GP, and will typically not be guaranteed to be diffeomorphic.

There are two common approaches to construct diffeomorphisms for machine learning purposes.

- Construct the function as the solution to an ODE, for example (Walder & Schölkopf, 2008).
- Construct the function to by composing simpler diffeomorphisms, for example (Papamakarios et al., 2021).

If we were to do the former, we may as well work with a CT model to begin with. The latter is not immediately applicable here, since we would need a fundamentally different probabilistic model for f, but is an interesting direction.

Thirdly, consider the van der Pol oscillator ($D = 2$, with parameter μ), which we use in Section 3.

$$\dot{x}_1 = x_2 \quad \dot{x}_2 = \mu(1 - x_1^2)x_2 - x_1 \tag{3}$$

The true dynamics have a sparse structure in the sense that $\dot{x}_1$ depends only on x_2, but when integrated over a step of δ to generate the DT updates, both update equations will depend on both variables. In higher dimensional problems, this could cause issues, for example if the downstream task involves inferring causal structures in the system.

3. Experimental comparison

We now consider tackling the full learning problem. We optimise the variational lower bound on the log marginal likelihood by gradient ascent. The approximate posterior $q(f)q(x) \approx p(x, f|y)$ follows previously used constructions (Frigola-Alcalde, 2014; Duncker et al., 2019). For $q(f)$ we use a sparse approximation to the GP (Titsias, 2009; Hensman et al., 2015), which is available in closed form given $q(x)$ and certain kernel expectations over $q(x)$. For $q(x)$, we use affine Gaussian transition densities whose parameters are optimised directly with respect to the objective. That is, in DT and CT respectively

$$q(x_{t+\delta}|x_t) = \mathcal{N}(x_{t+\delta}|A_t x_t + b_t, Q_t) \quad \text{or} \quad dx_t = (A_t x_t + b_t)dt + Q_t^{1/2} d\tilde{\beta}_t \tag{4}$$

where $\tilde{\beta}$ is a standard Brownian motion independent of β. We use a squared exponential kernel, for which all the required kernel expectations are available in closed form. For details of the objective function and parameterisation of the variational distribution, we refer the reader to the references.

Intuitively, since the model f is quite flexible, both models should do well under benign circumstance – where there is adequate, high quality data. When the data is lower quality – for example, noisier – then the priors play a more important role, and we might expect to see the DT model perform worse.

We test this hypothesis by training both models on a van der Pol oscillator (Equation (3)) with $\mu = 0.5$. This is a minimal working example in the sense that we require $D \geq 2$ to exhibit interesting nonlinear phenomena in CT, such as the stable oscillations of this system. We train the models in both lower ($R = 10^{-2}$) and higher ($R = 10^{-1}$) observation noise settings and examine the error in the posterior dynamics (Figure 3) and the latent means of $q(x)$ (Figure 4). We allow full state observations ($g(x) = x$) and keep the process noise fixed to a low value ($L_{DT} = \sqrt{0.001}, L_{CT} = \sqrt{0.001/\delta}$). As well as the plots, we estimate the root mean squared error (RMSE) between the learnt mean function and the ground truth dynamics.[1]

In the low noise case, both models perform well, tracking the state of the system closely, learning dynamics qualitatively similar to the groundtruth, and with similar RMSEs (in CT 0.51; in DT, 0.46). Although the DT model has a slightly lower RMSE, the CT model produces a qualitatively better fit: the states track the generating trajectory more closely, and the error is lower where there is data and in the central region. In high noise, both models struggle, but the RMSE is much worse for the DT model (in CT 0.84, in DT 1.37). Qualitatively, the CT model performs much better, with similar behaviour at the edges of the plots to low noise. However, the CT model is more computationally demanding, requiring around five times as long to train, and care to avoid numerical errors.

4. Conclusions

We have shown empirical evidence that using CT models for system identification when the data is generated by a CT process leads to better performance, and discussed the major qualitative differences between CT and DT priors. Worthwhile future work would be to compare CT and DT methods on real-world tasks, and investigate possible methods for flexible priors over diffeomorphisms for more competetive DT models.

[1] For a fair comparison, we convert the CT dynamics functions to DT by integrating: $\int_0^\delta f(x(t))dt$.

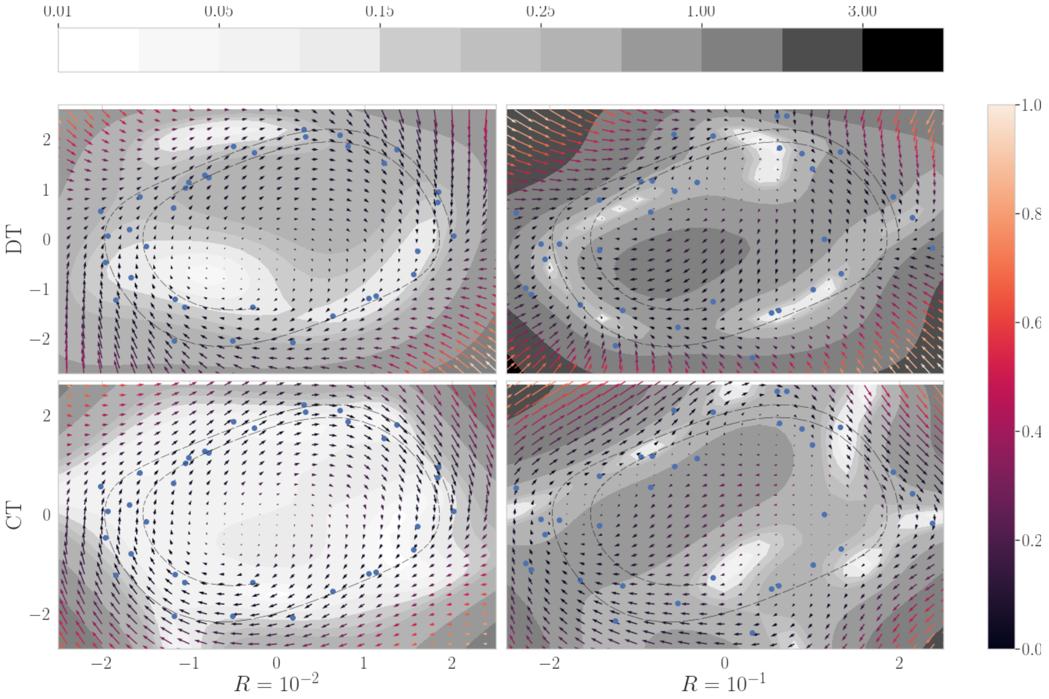

Figure 3: RMSE error (grey contours) and posterior predictive f (arrows, shaded by 2σ confidence) for the van der Pol fit. The blue dots are the training data, and the faint dashed line is the generating process. The CT model produces a visually better result in both noise settings.

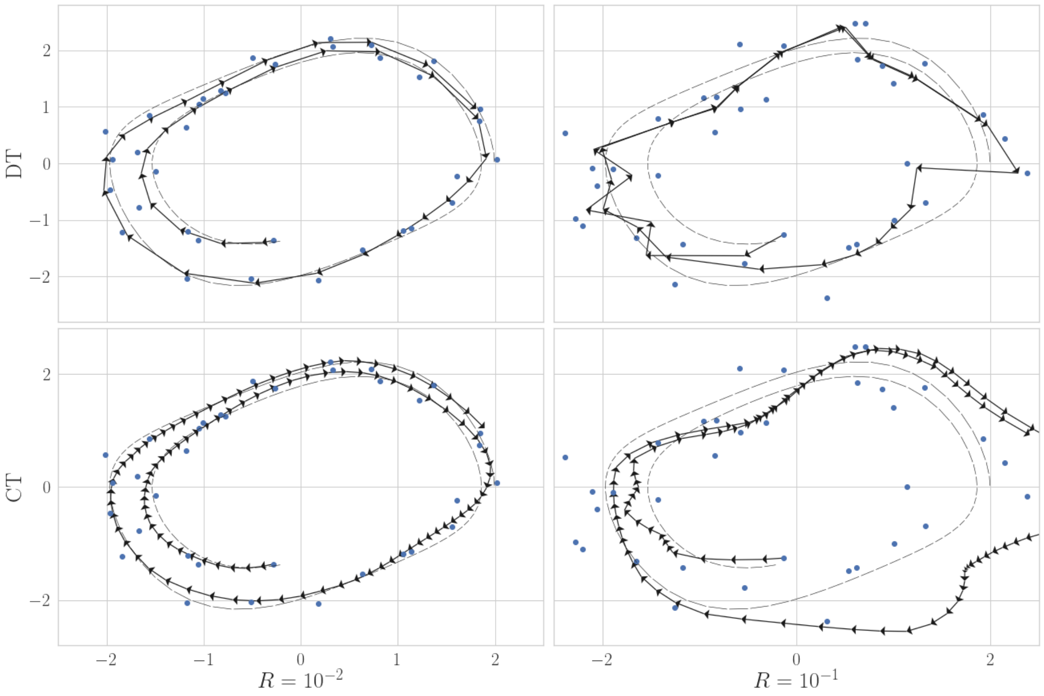

Figure 4: Mean of the approximate posterior trajectory $q(x)$ in the van der Pol fits as black arrows; rest as in Figure 3. Note that the DT model picks up the smooth, non-self-interesecting propoerties of the generating process in low noise, but violated these under high noise.

References

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice Workshop*, 2007.

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

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Curi, S., Melchior, S., Berkenkamp, F., and Krause, A. Structured variational inference in partially observable unstable Gaussian process state space models. In *2nd Conference on Learning for Dynamics and Control (L4DC)*, 2020.

Deisenroth, M. P. and Rasmussen, C. E. PILCO: A model-based and data-efficient approach to policy search. In *28th International Conference on Machine Learning (ICML)*, 2011.

Doerr, A., Daniel, C., Schiegg, M., Duy, N., Schaal, S., Toussaint, M., and Sebastian, T. Probabilistic recurrent state-space models. In *35th International Conference on Machine Learning (ICML)*, 2018.

Duncker, L., Bohner, G., Boussard, J., and Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *36th International Conference on Machine Learning (ICML)*, 2019.

Eleftheriadis, S., Nicholson, T., Deisenroth, M., and Hensman, J. Identification of Gaussian process state space models. In *30th Conference on Neural Information Processing Systems (NeurIPS)*. 2017.

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

Hensman, J., Matthews, A. G. d. G., and Ghahramani, Z. Scalable variational Gaussian process classification. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Hirsch, M. W., Smale, S., and Devaney, R. L. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 2nd edition, 2004. ISBN 978-0-12-349703-1.

Ialongo, A. D., van der Wilk, M., Hensman, J., and Rasmussen, C. E. Overcoming mean-field approximations in recurrent Gaussian process models. In *36th International Conference on Machine Learning (ICML)*, 2019.

Leander, J., Lundh, T., and Jirstrand, M. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences*, 2014. doi: https://doi.org/10.1016/j.mbs.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S0025556414000510.

Losert, V. and Akin, E. Dynamics of games and genes: Discrete versus continuous time. *Journal of Mathematical Biology*, 17(2):241–251, 1983.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research (JMLR)*, 2021. URL http://jmlr.org/papers/v22/19-1028.html.

Särkkä, S. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013. ISBN 978-1-107-61928-9.

Särkkä, S. and Solin, A. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019. ISBN 978-1-316-51008-7.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

Walder, C. and Schölkopf, B. Diffeomorphic dimensionality reduction. In *21st Conference on Neural Information Processing Systems (NeurIPS)*, 2008. URL https://proceedings.neurips.cc/paper/2008/file/647bba344396e7c8170902bcf2e15551-Paper.pdf.