# Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations

**Yali Wang** YALI.WANG.1@ULAVAL.CA
Department of Computer Science, Laval University, Canada

**David Barber** DAVID.BARBER@UCL.AC.UK
Department of Computer Science, University College London, U.K.

## Abstract

Bayesian parameter estimation in coupled ordinary differential equations (ODEs) is challenging due to the high computational cost of numerical integration. In gradient matching a separate data model is introduced with the property that its gradient may be calculated easily. Parameter estimation is then achieved by requiring consistency between the gradients computed from the data model and those specified by the ODE. We propose a Gaussian process model that directly links state derivative information with system observations, simplifying previous approaches and improving estimation accuracy.

## 1. Introduction

Ordinary differential equations (ODEs) are continuous time models with the interaction between variables described by $\dot{x}(t) = f(x(t), \theta)$, for vector $x$ and vector output function $f$. The task is to estimate any unknown parameters $\theta$ of the ODEs by fitting them to observed data collected at a set of discrete observation times, $t_1, \ldots, t_T$. A principled approach to this problem is to first numerically integrate the ODEs for a given value of $\theta$ and initial value $x(0)$ to obtain a vector of values $X \equiv x_{t_1}, \ldots, x_{t_T}$. Parameter estimation is achieved by finding $\theta$ such that $X$ closely matches the observed data. However, numerical integration is computationally demanding, rendering this otherwise ideal scheme impractical in all but the smallest systems, see for example (Vyshemirsky & Girolami, 2008).

In gradient matching we avoid explicit numerical integration by considering an alternative model of the data, $x(t) = g(t, \phi)$. Given this fit to the data, one can compute the gradients of the fitted function at the observed timepoints, $\dot{x}(t) = \dot{g}(t, \phi)$. Gradient matching estimates parameters $\theta$ of the ODE and parameters $\phi$ of the fitted function $g$ by requiring that the gradients in both models are consistent at the observed timepoints. A review of this class of approaches can be found in (Ramsay et al., 2007).

As described in (Calderhead et al., 2008), previous gradient matching approaches provided only limited point-parameter estimates or can prove numerically inconsistent. Recently, Gaussian Processes (GPs) have been considered as data models within the gradient matching framework (Calderhead et al., 2008; Dondelinger et al., 2013) and for the solution of linear operator equations (Graepel, 2003). GPs provide a distribution over fitted functions and associated gradients. Using priors on the parameters of the GP model and the ODE, this gives a flexible Bayesian parameter estimation procedure. More concretely, in (Calderhead et al., 2008) GP parameters $\phi$ are fitted first to the data, and subsequently the parameters of the ODE $\theta$ are estimated. The estimation accuracy is however limited by the lack of feedback from ODE parameter inference to GP parameter inference. To address this Dondelinger et al. (2013) introduced bidirectional interaction between ODE and GP parameters, demonstrating improved parameter estimation.

These GP approaches have similar computational complexity and can run up to two orders of magnitude faster than numerical integration. The benefits of a Bayesian approach to parameter estimation in ODEs are now well-established and we propose to improve on previous approaches by introducing a simpler generative model that directly links state derivatives to system observations using a GP. This plays a similar role to numerical integration but without the corresponding high computational cost.

### 1.1. ODE System Description

We consider continuous time dynamical systems in which the motions of $K$ states $\mathbf{x}(t) \equiv [x_1(t), x_2(t), \ldots, x_K(t)]^T$ are represented by a set of $K$ ODEs

$$\dot{\mathbf{x}}(t) \equiv \frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \theta)$$

*Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

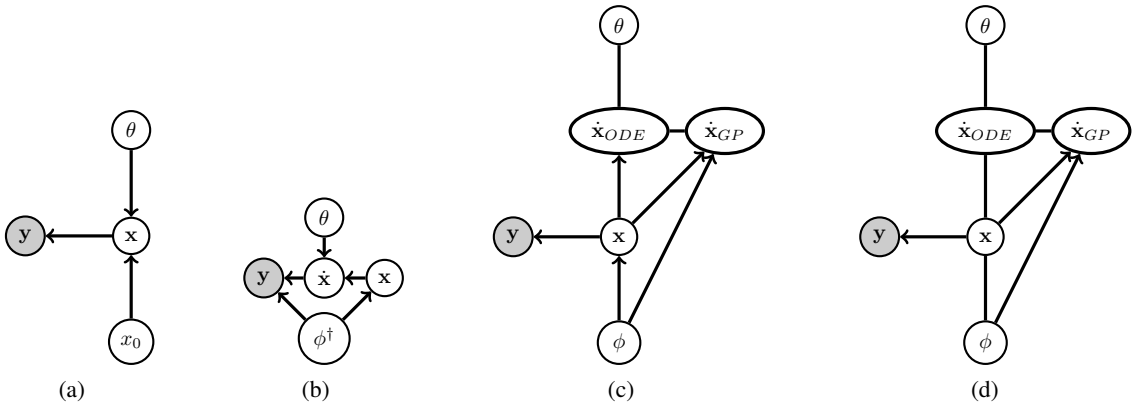



*Figure 1.* (a) Numerical Integration with an initial term $x_0$. (b) Our GP-ODE approach corresponds to a generative belief network. (c) Calderhead et al. (2008) approach, which is based on a form of compatibility function, expressed as a chain graph. (d) The chain graph of the Dondelinger et al. (2013) approach uses a modified compatibility function. Note that the difference between (c) and (d) is that in (d) the links $\mathbf{x} - \dot{\mathbf{x}}_{ODE}$ and $\phi - \mathbf{x}$ are undirected, reflecting the different normalisation requirement.

where $\theta$ is a vector of parameters of the ODE. For notational convenience, we additionally define the state matrix $\mathbf{X} \equiv [\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_T)]$ and $k$-th state sequence $\mathsf{x}_k \equiv [x_k(t_1), x_k(t_2), \ldots, x_k(t_T)]^T$. Given potentially noisy observations of $\mathbf{X}$ (see below), the task is to infer a posterior distribution over the parameters $\theta$.

### 1.2. Observation Model

The $T$ observations $\mathbf{Y} = [\mathbf{y}(t_1), \mathbf{y}(t_2), \ldots, \mathbf{y}(t_T)]$ are obtained from the states according to independent additive noise $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$ where the noise for the $k$-th state, $k \in \{1, 2, \ldots, K\}$, is Gaussian, $\epsilon_k(t) \sim \mathcal{N}(0, \sigma_k^2)$. This gives then an observation model

$$p_{OBS}(\mathbf{Y}|\mathbf{X}) = \prod_t p_{OBS}(\mathbf{y}(t)|\mathbf{x}(t))$$

with $p_{OBS}(\mathbf{y}(t)|\mathbf{x}(t)) = \mathcal{N}(\mathbf{x}(t), \sigma^2\mathbf{I})$.

If unknown, the parameters of the observation model ($\sigma$ in this case) form part of the parameters that need to be estimated. This is achieved by placing a prior over their values and incorporating these parameters into the model in the standard way. This step is unproblematic and, to avoid notational clutter, we drop these observation parameters as variables in the model descriptions below (they will however be included in the experiments).

### 1.3. Bayesian Numerical Integration

Given the ODE and an assumed initial value $x_0$, we can then (in principle) numerically integrate the system. For example[1], for $K = 1$, using a simple approach based on discretising time in small intervals of $\delta$, a numerical value $x'$ for the integrated path is given by $x'_{n+1} = x'_n + \delta f(x'_n, \theta)$, with $x'_0 = x_0$. This is iterated until the desired end time. This can be considered as a procedure that, for a given initial value, produces (in this case a deterministic) distribution $p(\mathsf{x}|x_0, \theta) = \delta(\mathsf{x} - \mathsf{x}'(x_0))$ over the values of the state at the observation times. Here $\delta(\cdot)$ is the Dirac delta function. Given then a prior on $\theta$ and the integration constant $x_0$, this defines a joint distribution

$$p(\mathsf{y}, \mathsf{x}, x_0, \theta) = p_{OBS}(\mathsf{y}|\mathsf{x})p(\mathsf{x}|x_0, \theta)p(x_0)p(\theta)$$

from which samples $p(\theta, x_0|\mathsf{y})$ can be drawn. This ideal procedure can produce excellent results (Vyshemirsky & Girolami, 2008); however the computational expense is prohibitive in larger models with the bottleneck being the explicit numerical integration that needs to be carried out for every value of $\theta, x_0$ of interest (Calderhead et al., 2008).

## 2. The GP-ODE generative model

As an alternative to explicit Bayesian numerical integration, we propose the following generative model over states $\mathbf{X}$, their derivatives $\dot{\mathbf{X}}$, observations $\mathbf{Y}$ and remaining parameters using a simple belief network, fig(1b),

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X}, \dot{\mathbf{X}}, \phi^\dagger, \theta) &= p(\theta)p(\phi^\dagger) \\ &\times p_{GP}(\mathbf{Y}|\dot{\mathbf{X}}, \phi^\dagger)p_{ODE}(\dot{\mathbf{X}}|\mathbf{X}, \theta)p_{GP}(\mathbf{X}|\phi^\dagger) \quad (1) \end{aligned}$$

where $\phi^\dagger \equiv (\mathbf{x}_0, \phi)$. To generate data from this model we first sample parameters $\phi^\dagger$, $\theta$ from their priors and then a state $\mathbf{X}$ from the GP prior $p_{GP}(\mathbf{X}|\phi^\dagger)$. A state derivative is subsequently obtained by sampling from $p_{ODE}(\dot{\mathbf{X}}|\mathbf{X}, \theta)$. Finally, given these state derivatives $\dot{\mathbf{X}}$, observations $\mathbf{Y}$ are generated by sampling from the GP $p_{GP}(\mathbf{Y}|\dot{\mathbf{X}}, \phi^\dagger)$. In this way we combine a smoothness prior assumption on the

[1] In practice we use the Runge-Kutta method.

state $\mathbf{X}$ together with derivative information obtained from the ODE in a single generative model[2].

ENCODING THE ODE: $p_{ODE}(\dot{\mathbf{X}}|\mathbf{X},\theta)$

The temporal evolution of the ODE is encoded in the distribution $p_{ODE}(\dot{\mathbf{X}}|\mathbf{X},\theta)$. In the deterministic ODE case (which we assume throughout) this will simply be a delta function distribution $\delta(\dot{\mathbf{X}} - \mathbf{f}(\mathbf{X},\theta)) \equiv \prod_t \delta\left(\dot{\mathsf{x}}(t) - \mathbf{f}(\mathsf{x}(t),\theta)\right)$, though Gaussian additive noise would be straightforward to incorporate for the case of Gaussian SDEs.

PRIOR ON LATENT STATE: $p_{GP}(\mathbf{X}|\phi^\dagger)$

The GP prior assumes that each state dimension is a priori independent $p_{GP}(\mathbf{X}|\phi^\dagger) = \prod_k p_{GP}(\mathsf{x}_k|\phi_k^\dagger)$, with $p_{GP}(\mathsf{x}_k|\phi_k^\dagger)$ formed from a GP with mean function[3] $\mu_{\phi_k}(t)$ and covariance function $c_{\phi_k}(t,t')$.

IMPLICIT INTEGRATION: $p_{GP}(\mathbf{Y}|\dot{\mathbf{X}},\phi^\dagger)$

The term $p_{GP}(\mathbf{Y}|\dot{\mathbf{X}}) = \prod_k p_{GP}(\mathsf{y}_k|\dot{\mathsf{x}}_k)$ (dropping parameter dependencies on the r.h.s for compactness of notation)

$$p_{GP}(\mathsf{y}_k|\dot{\mathsf{x}}_k) = \int p_{OBS}(\mathsf{y}_k|\mathsf{x}_k)p_{GP}(\mathsf{x}_k|\dot{\mathsf{x}}_k)d\mathsf{x}_k$$

plays a key role in our model and specifies how to implicitly integrate a given state derivative curve to arrive at a distribution over observations. Since differentiation is a linear operation, the derivative of a GP is also a GP – see for example (Solak et al., 2002). Hence the joint distribution $p_{GP}(\mathsf{y}_k,\mathsf{x}_k,\dot{\mathsf{x}}_k)$ is Gaussian distributed. Using the prior $p_{GP}(\mathbf{X}|\phi^\dagger)$ and observation model $p_{OBS}(\mathbf{Y}|\mathbf{X})$ we obtain the covariance functions

$$\begin{aligned}
cov(\dot{x}_k(t),\dot{x}_k(t')) &= \frac{\partial^2 c_{\phi_k}(t,t')}{\partial t\partial t'} - \frac{\partial \mu_{\phi_k}(t)}{\partial t}\frac{\partial \mu_{\phi_k}(t')}{\partial t'}\\
cov(\dot{x}_k(t),x_k(t')) &= \frac{\partial c_{\phi_k}(t,t')}{\partial t} - \mu_{\phi_k}(t')\frac{\partial \mu_{\phi_k}(t)}{\partial t}\\
cov(y_k(t),\dot{x}_k(t')) &= cov(x_k(t),\dot{x}_k(t'))\\
cov(y_k(t),y_k(t')) &= c_{\phi_k}(t,t') + \sigma^2\delta(t-t')
\end{aligned}$$

and mean functions

$$\bar{y}_k(t) = \bar{x}_k(t) = \mu_{\phi_k}(t), \quad \dot{\bar{x}}_k(t) = \partial\mu_{\phi_k}(t)/\partial t$$

Given the state derivatives, the observations are then Gaussian distributed[4]

$$p_{GP}(\mathsf{y}_k|\dot{\mathsf{x}}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k^{y|\dot{x}}, \boldsymbol{\Sigma}_k^{y|\dot{x}})$$

where

$$\begin{aligned}
\boldsymbol{\mu}_k^{y|\dot{x}} &= \boldsymbol{\mu}_{\phi_k} + \mathbf{C}_{\phi_k}^{x\dot{x}}(\mathbf{C}_{\phi_k}^{\dot{x}\dot{x}})^{-1}\left(\dot{\mathsf{x}}_k - \dot{\bar{\mathsf{x}}}_k\right)\\
\boldsymbol{\Sigma}_k^{y|\dot{x}} &= \mathbf{C}_{\phi_k} + \sigma_k^2\mathbf{I} - \mathbf{C}_{\phi_k}^{x\dot{x}}(\mathbf{C}_{\phi_k}^{\dot{x}\dot{x}})^{-1}\mathbf{C}_{\phi_k}^{\dot{x}x}
\end{aligned}$$

$\mathbf{C}_{\phi_k}^{\dot{x}\dot{x}}$, $\mathbf{C}_{\phi_k}^{\dot{x}x}$ and $\mathbf{C}_{\phi_k}^{x\dot{x}}$ are constructed using the results above evaluated at the observation times $t_1, t_2, \ldots, t_T$.

### 2.1. Parameter Estimation

From (1), the conditional marginal distribution over observations, latent states and parameters is given by

$$\begin{aligned}
&p(\mathbf{Y},\mathbf{X}|\phi^\dagger,\theta)\\
&= p_{GP}(\mathbf{X}|\phi^\dagger)\int p_{GP}(\mathbf{Y}|\dot{\mathbf{X}},\phi^\dagger)p_{ODE}(\dot{\mathbf{X}}|\mathbf{X},\theta)d\dot{\mathbf{X}}
\end{aligned}$$

This integral can be analytically evaluated in the case of Gaussian additive noise in the ODE. In the deterministic case, this reduces to simply

$$p(\mathbf{Y},\mathbf{X}|\phi^\dagger,\theta) = p_{GP}(\mathbf{X}|\phi^\dagger)p_{GP}(\mathbf{Y}|\dot{\mathbf{X}} = \mathbf{f}(\mathbf{X},\theta),\phi^\dagger)$$

The distribution over observations, latent states and parameters is then given by

$$p(\mathbf{Y},\mathbf{X},\phi^\dagger,\theta) = p(\mathbf{Y},\mathbf{X}|\phi^\dagger,\theta)p(\phi^\dagger)p(\theta) \qquad (2)$$

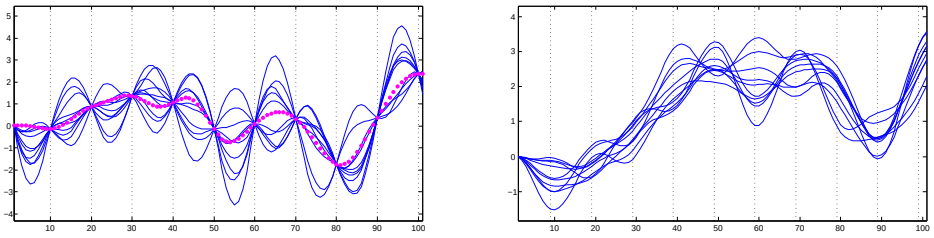
Estimation of the parameters and latent state $\mathbf{X}$ can then be carried out for example by sampling from the posterior $p(\mathbf{X},\phi^\dagger,\theta|\mathbf{Y})$, see section(3). Note that, in contrast to (Calderhead et al., 2008; Dondelinger et al., 2013), the normalisation constant of the joint distribution (2) is known which facilitates sampling.

Previous approaches (Calderhead et al., 2008; Dondelinger et al., 2013) used a GP to compute $p(\mathbf{X},\dot{\mathbf{X}}|\mathbf{Y})$; parameter estimation is achieved by requiring gradients from this to match the desired gradient $\mathbf{f}(\mathbf{X},\theta)$. The key difference between this and our approach is our direct link from the latent gradient $\dot{\mathbf{X}}$ to the observation $\mathbf{Y}$. This term can be expressed as $p(\mathbf{Y}|\dot{\mathbf{X}}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}|\dot{\mathbf{X}})d\mathbf{X}$ where $p(\mathbf{X}|\dot{\mathbf{X}})$ implicitly performs numerical integration, as we describe below. Since $\dot{\mathbf{X}}$, $\mathbf{X}$ and $\mathbf{Y}$ are defined only at the measurement times, no fine time discretization is required in our model.

[2] It is natural to consider forming the joint $p(y,x,\dot{x})$ as $p_{OBS}(y|x)p_{ODE}(\dot{x}|x)p_{GP}(x)$. However, the marginal $p(y,x) = p_{OBS}(y|x)p_{GP}(x)$ is then vacuous, containing no contribution from the ODE. All models, including fig(1b,c,d), are 'incorrect' compared to the true model fig(1a); the challenge is to combine aspects of numerical integration with the GP and observation model that achieves coherent parameter estimation with reduced computational cost over explicit numerical integration.

[3] There are different ways to define $p(\mathbf{x},x_0)$. One approach is to express this as $p_{GP}(\mathbf{x}|x_0)p(x_0)$, which allows one to use the same prior $p(x_0)$ as for the BNI model, section(1.3). In the experiments we more simply defined a joint GP $p_{GP}(\mathbf{x},x_0)$, for each $k$, with mean $\mu_{\phi_k}(t)$ equal to the mean of the observed data, for all $t$. This is equivalent to defining a Gaussian prior on $\mathbf{x}_0$ with mean that of the observed data.

[4] For a Gaussian defined on joint variables $\mathbf{z} = (\mathbf{x},\mathbf{y})$ with $p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_{z,z})$, the conditional is Gaussian with mean and covariance given from the block mean and covariances, $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{x,y}\boldsymbol{\Sigma}_{y,y}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_y\right), \boldsymbol{\Sigma}_{x,x} - \boldsymbol{\Sigma}_{x,y}\boldsymbol{\Sigma}_{y,y}^{-1}\boldsymbol{\Sigma}_{y,x}\right)$, see *e.g.* (Barber, 2012).



(a) (b)

*Figure 2.* (a) The dotted curve is the true (but unknown) derivative curve $\dot{\mathbf{x}}$, which is only observed every 10th timepoint, giving observation $\dot{\mathbf{x}}_{\mathcal{M}}$ at these measurement times. From this we calculate the GP posterior $p_{GP}(\mathbf{x}|\dot{\mathbf{x}}_{\mathcal{M}})$ (assuming $x_0 = 0$). (b) Samples from $p_{GP}(\mathbf{x}|\dot{\mathbf{x}}_{\mathcal{M}})$. The derivative of each sample is plotted in (a). The GP constrains the samples from $p_{GP}(\mathbf{x}|\dot{\mathbf{x}}_{\mathcal{M}})$ such that their derivatives match the observed derivatives $\dot{\mathbf{x}}_{\mathcal{M}}$. The model in section(2) describes the distribution only on the marginal quantities $p_{GP}(\mathbf{x}_{\mathcal{M}}|\dot{\mathbf{x}}_{\mathcal{M}})$ and thus avoids working with the fine time discretization required in explicit numerical integration.

### 2.2. Informal Justification

To minimise notational issues, we consider a univariate system with $K = 1$. Furthermore we discretize time so that $t = n\delta$, for integer time-index $n \in \{1, \ldots, N\}$ and real discretization interval $\delta$. Note that in the model in section(2) the timepoints are defined only at the observation times; however in this section we need to notationally distinguish between times that the data are observed and a finer discretisation of time that could be used to carry out numerical integration. The times at which data will be observed are therefore described by a subset $m \in \mathcal{M}$ of the fine time discretization; for example we measure $y_m$ at time indices $\mathcal{M} = \{1, 10, 20, \ldots\}$. To emphasise that the measurements only occur at a subset of all discrete times, we write $\mathbf{y}_{\mathcal{M}}$ for the observed measurements. Given a curve $x_n$, the numerical derivative is given by $\dot{x}_n = (x_n - x_{n-1})/\delta$. We assume that $\dot{x}_1 = x_1 - x_0$, where $x_0$ is the constant of integration. For a vector $\mathbf{x}$, the derivative vector is then given by the difference equation $\dot{\mathbf{x}} = \mathbf{D}\mathbf{x} - \mathbf{b}$, where the square invertible matrix $\mathbf{D}$ has zero elements, except for $D_{n,n-1} = -1/\delta, D_{n,n} = 1/\delta$, $D_{1,1} = 1$ and $\mathbf{b}$ is the zero vector except for $b_1 = x_0$. To explain the fundamental mechanism, we fix the parameters of the GP and observation model. The posterior over the discretized state is

$$p(\mathbf{y}_{\mathcal{M}}, \mathbf{x}, \dot{\mathbf{x}}) = p_{GP}(\mathbf{y}_{\mathcal{M}}|\dot{\mathbf{x}})p_{ODE}(\dot{\mathbf{x}}|\mathbf{x})p_{GP}(\mathbf{x}) \quad (3)$$

Writing $\mathbf{x}'$ for the numerically integrated curve, we have

$$p_{GP}(\mathbf{y}_{\mathcal{M}}|\dot{\mathbf{x}}) = \int p_{OBS}(\mathbf{y}_{\mathcal{M}}|\mathbf{x}'_{\mathcal{M}})p_{GP}(\mathbf{x}'|\dot{\mathbf{x}})d\mathbf{x}'$$

Assuming that the derivative curve is obtained by differencing, we can invert this relation using Bayes' rule

$$p_{GP}(\mathbf{x}'|\dot{\mathbf{x}}) \propto \delta(\dot{\mathbf{x}} - \mathbf{D}\mathbf{x}' + \mathbf{b})p_{GP}(\mathbf{x}')$$

Since $\mathbf{D}$ is invertible, the GP plays no role, to give

$$p_{GP}(\mathbf{x}'|\dot{\mathbf{x}}) \propto \delta(\dot{\mathbf{x}} - \mathbf{D}\mathbf{x}' + \mathbf{b})$$

Using $p_{ODE}(\dot{\mathbf{x}}|\mathbf{x}) = \delta(\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}, \theta))$, and integrating (3) over $\dot{\mathbf{x}}$, we obtain the joint distribution over the observations $\mathbf{y}_{\mathcal{M}}$ and latent curve $\mathbf{x}$, $p(\mathbf{y}_{\mathcal{M}}, \mathbf{x}) = p(\mathbf{y}_{\mathcal{M}}|\mathbf{x})p_{GP}(\mathbf{x})$, where

$$p(\mathbf{y}_{\mathcal{M}}|\mathbf{x}) \propto$$
$$\int p_{OBS}(\mathbf{y}_{\mathcal{M}}|\mathbf{x}'_{\mathcal{M}})\delta(\dot{\mathbf{x}} - \mathbf{D}\mathbf{x}' + \mathbf{b})\delta(\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}, \theta))d\mathbf{x}'d\dot{\mathbf{x}}$$
$$= p_{OBS}(\mathbf{y}_{\mathcal{M}}|\mathbf{x}'_{\mathcal{M}} = \left[\mathbf{D}^{-1}(\mathbf{f}(\mathbf{x}, \theta) + \mathbf{b})\right]_{\mathcal{M}})$$

This can be interpreted as taking a set of gradients $\mathbf{f}(\mathbf{x}, \theta)$, integrating them numerically (via the inversion $\mathbf{D}^{-1}$ which performs summation of the components of $\mathbf{f}$) and taking the measurement indices of this vector. The likelihood of the observations $p(\mathbf{y}_{\mathcal{M}}) = \int p(\mathbf{y}_{\mathcal{M}}|\mathbf{x})p_{GP}(\mathbf{x})d\mathbf{x}$ is equivalent to the mass of GP curves $\mathbf{x}$ whose numerical derivatives match $\mathbf{f}(\mathbf{x}, \theta)$ weighted by how well they fit the observed data $\mathbf{y}_{\mathcal{M}}$ at the observation times $\mathcal{M}$. Taking the limit $\delta \to 0$, the above difference equation becomes the differential equation (1.1) and the Gaussian over $\mathbf{x}$ becomes a GP, with $\dot{\mathbf{x}}$ the associated GP derivative, as specified by the model (1), see figure(2).

## 3. Inference in the GP-ODE model

There are a number of approaches one could take to draw samples from the GP-ODE posterior $p(\mathbf{X}, \phi^{\dagger}, \theta|\mathbf{Y})$ and our philosophy was to choose the simplest that provides good results. Writing $\Phi = \{\mathbf{x}_0, \phi, \sigma\}$ for all the parameters of the GP and observation model, we sample from $p(\mathbf{X}, \phi^{\dagger}, \theta|\mathbf{Y})$ using a Gibbs procedure to produce a set of samples $\Phi^i, \theta^i, \mathbf{X}^i$. We initialize $\Phi^0$, $\theta^0$ at random and draw $\mathbf{X}^0 \sim p_{GP}(\mathbf{X}|\mathbf{Y}, \Phi^0)$. We subsequently draw samples, indexed by $i = 1 : L$ by alternately drawing from

1. $\theta^i, \Phi^i \sim p(\theta, \Phi|\mathbf{X}^{i-1}, \mathbf{Y})$
2. $\mathbf{X}^i \sim p(\mathbf{X}|\theta^i, \Phi^i, \mathbf{Y})$

We present a naive approach for drawing from these conditionals below[5].

### 3.1. Parameter sampling

We draw from $p(\theta^i, \Phi^i|\mathbf{X}^{i-1}, \mathbf{Y})$ using Gibbs sampling:

1. Set $\theta^{i,0} = \theta^{i-1}$,$\Phi^{i,0} = \Phi^{i-1}$
2. For $j = 1 : L_p$
   (a) $\Phi^{i,j} \sim p(\Phi|\mathbf{X}^{i-1}, \theta^{i,j-1}, \mathbf{Y})$
   (b) $\theta^{i,j} \sim p(\theta|\mathbf{X}^{i-1}, \Phi^{i,j}, \mathbf{Y})$
3. Set $\theta^i = \theta^{i,L_p}$, $\Phi^i = \Phi^{i,L_p}$

where these conditional distributions can be obtained from the joint (2). Where there are multiple components of a parameter, we again use Gibbs sampling to obtain a univariate sample of a component conditioned on the remaining components. In the experiments we assume that the parameters take values from known discrete sets (the priors are discrete), in which case sampling from these conditionals is particularly straightforward.

### 3.2. State sampling

It is natural to consider drawing samples from $p(\mathbf{X}|\theta, \Phi, \mathbf{Y})$ using Metropolis-Hastings (similar to (Dondelinger et al., 2013)) with $p_{GP}(\mathbf{X}|\Phi, \mathbf{Y})$ as the proposal. However, in our experience, this results in poor mixing. We therefore use Gibbs sampling in which we draw a state from $p(\mathbf{x}(t)|\mathbf{X}_{\setminus t}, \theta, \Phi, \mathbf{Y})$, where $\mathbf{X}_{\setminus t}$ are the states except for $\mathbf{x}(t)$, drawing samples in sequence from times $t \in \{t_1, \ldots, t_T\}$. To draw from $p(\mathbf{x}(t)|\mathbf{X}_{\setminus t}, \theta, \Phi, \mathbf{Y})$ we use either Metropolis-Hastings with proposal $p_{GP}(\mathbf{x}(t)|\mathbf{X}_{\setminus t}, \Phi, \mathbf{Y})$ or Gibbs sampling for each component of the vector $\mathbf{x}(t)$ based on discrete values[6]. After $L_x$ sweeps through all timepoints, we obtain the new sample $\mathbf{X}^i$.

## 4. Relation to previous approaches

### 4.1. Gradient Matching

(Calderhead et al., 2008) is based on matching gradients via what could be termed a 'compatability' function (for the case $K = 1$ and fixed $\sigma$ for notational simplicity)

$$\omega(\dot{\mathbf{x}}, \mathbf{x}|\theta, \phi) \equiv p_{GP}(\dot{\mathbf{x}}|\mathbf{x}, \phi)p_{ODE}(\dot{\mathbf{x}}|\mathbf{x}, \theta)$$

This is used to define

$$p(\theta|\mathbf{x}, \phi) \propto p(\theta)\omega'(\mathbf{x}|\theta, \phi)$$

using the marginal compatibility

$$\omega'(\mathbf{x}|\theta, \phi) \equiv \int \omega(\dot{\mathbf{x}}, \mathbf{x}|\theta, \phi)d\dot{\mathbf{x}}$$

with presumably the intention that this has high value when the gradient distributions overlap[7]. The marginal compatibility $\omega'(\mathbf{x}|\theta, \phi)$ is analytically computed since the terms under the integral are Gaussian. The authors modify the deterministic ODE by the addition of fictitious noise to give a Gaussian distribution for $p_{ODE}(\dot{\mathbf{x}}|\mathbf{x}, \theta)$ with mean $\mathbf{f}(\mathbf{x}, \theta)$. To ease comparison with our approach (the extension to the stochastic ODE case is trivial), we take the deterministic ODE case, for which the above reduces to

$$p(\theta|\mathbf{x}, \phi) \propto p(\theta)p_{GP}(\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \theta)|\mathbf{x}, \phi) \quad (4)$$

The joint distribution over observations, latent states and parameters is then defined as

$$p(\mathbf{y}, \mathbf{x}, \theta, \phi) = p_{OBS}(\mathbf{y}|\mathbf{x})p(\theta|\mathbf{x}, \phi)p_{GP}(\mathbf{x}|\phi)p(\phi) \quad (5)$$

Inference is then achieved by sampling, conditioned on the observed sequence $\mathbf{y}$. The unknown normalisation term of (4) is a function of $\phi$ and thus makes direct Gibbs sampling from this posterior problematic. The approach taken in (Calderhead et al., 2008) is to first refactor the joint distribution in the form[8]

$$p(\mathbf{y}, \mathbf{x}, \theta, \phi) = p(\theta|\mathbf{x}, \phi)p(\mathbf{x}|\phi, \mathbf{y})p(\phi|\mathbf{y})p(\mathbf{y})$$

Conditioned on $\mathbf{y}$, ancestral sampling is then performed:

$$\begin{aligned}
&\phi \sim p(\phi|\mathbf{y}) \\
&\mathbf{x} \sim p(\mathbf{x}|\phi, \mathbf{y}) \propto p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi) \\
&\theta \sim p(\theta|\mathbf{x}, \phi)
\end{aligned}$$

Here, $p(\phi|\mathbf{y}) \propto p(\phi) \int p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)d\mathbf{x}$ for which the integral can be evaluated analytically. A disadvantage of this model is that the posterior $p(\phi|\mathbf{y})$ does not take the ODE system dynamics into consideration. Effectively, a GP is fitted to the data first (without knowledge of the system dynamics) and the parameters $\theta$ of the ODE are subsequently adjusted to best match the fitted GP.

The gradient matching approach can be defined as a graphical model chain graph (see for example (Koller & Friedman, 2009)) distribution[9] , fig(1c), with factors

$$\begin{aligned}
&p_{OBS}(\mathbf{y}|\mathbf{x})p_{ODE}(\dot{\mathbf{x}}_{ODE}|\mathbf{x}, \theta)p_{GP}(\dot{\mathbf{x}}_{GP}|\mathbf{x}, \phi) \\
&\times \delta\left(\dot{\mathbf{x}}_{ODE} - \dot{\mathbf{x}}_{GP}\right) p_{GP}(\mathbf{x})p(\theta)p(\phi)
\end{aligned}$$

[5]More sophisticated sampling strategies could be considered. However for the benchmark experiments, these approaches have proved adequate.

[6]For the experiments, the Gibbs approach proved adequate.

[7]The mathematical motivation for this is less clear. Given distributions $p$ and $q$, their 'overlap' $\int p(x)q(x)dx$ is maximal when $q(x)$ is a delta distribution placing all its mass on the most likely state of $p(x)$; this is not necessarily the same as matching $q$ to $p$.

[8]Whilst this can be interpreted as generative model, this is unnatural since the term $p(\theta|\mathbf{x}, \phi)$ means that the parameters of the ODE depend on the generated state $\mathbf{x}$.

[9]This chain graph structure is the same for the trivial extension to the stochastic ODE case.

The undirected link between $\theta$ and $\dot{\mathbf{x}}_{ODE}$ is necessary to ensure that the variables $\dot{\mathbf{x}}_{ODE}$, $\dot{\mathbf{x}}_{GP}$, $\theta$ form a component of the chain graph. Marginalising this chain distribution over $\dot{\mathbf{x}}_{GP}$ and $\dot{\mathbf{x}}_{ODE}$ gives the marginal distribution

$$p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)p(\phi)\frac{p(\theta)\omega'(\mathbf{x}|\theta,\phi)}{\int p(\theta)\omega'(\mathbf{x}|\theta,\phi)d\theta} \quad (6)$$

which matches (5). We can also write this as

$$p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)p(\phi)p(\theta)m_{GM}(\mathbf{x}|\theta,\phi) \quad (7)$$

where we define the gradient matching function

$$m_{GM}(\mathbf{x}|\theta,\phi) \equiv \frac{\omega'(\mathbf{x}|\theta,\phi)}{\int p(\theta)\omega'(\mathbf{x}|\theta,\phi)d\theta} \quad (8)$$

### 4.2. Adaptive Gradient Matching

Dondelinger et al. (2013) considered a modified gradient matching approach with joint distribution

$$p(\mathbf{y},\mathbf{x},\dot{\mathbf{x}},\theta,\phi) \propto p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)\omega(\dot{\mathbf{x}},\mathbf{x}|\theta,\phi)p(\theta)p(\phi)$$

and marginal

$$p(\mathbf{y},\mathbf{x},\theta,\phi) \propto p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)\omega'(\mathbf{x}|\theta,\phi)p(\theta)p(\phi)$$

A benefit of this approach is that the marginal

$$p(\mathbf{y}|\phi) \propto \int p_{OBS}(\mathbf{y}|\mathbf{x})p_{GP}(\mathbf{x}|\phi)\int \omega'(\mathbf{x}|\theta,\phi)p(\theta)d\theta d\mathbf{x}$$

does depend on the ODE; in contrast to (Calderhead et al., 2008) the parameters of the GP are influenced by the ODE (and vice versa). The marginal $p(\mathbf{y},\mathbf{x},\theta,\phi)$ can also be written in the same form as expression (7) but with the adaptive gradient matching function

$$m_{AGM}(\mathbf{x}|\theta,\phi) \equiv \frac{\omega'(\mathbf{x}|\theta,\phi)}{\int p(\theta)p(\phi)p_{GP}(\mathbf{x}|\phi)\omega'(\mathbf{x}|\theta,\phi)d\theta d\mathbf{x}d\phi}$$

The improved performance of AGM over GM (Dondelinger et al., 2013) may be attributed to the fact that $m_{AGM}$ is proportional to the marginal compatibility $\omega'$ and therefore always encourages matching between the GP and the ODE, whereas $m_{GM}$ less strongly encourages matching due to the partial cancellation of $\omega'$ in both the numerator and denominator of (8). No such issues arise in the GP-ODE approach in which the coupling between the GP and ODE parameters occurs through the implicit numerical integration mechanism, as described in section(2.2), which ensures agreement between the ODE and GP curves.

The factors in the corresponding chain graph, fig(1d), are the same as for the gradient matching method of section(4.1). However, all variables except $\mathbf{y}$ form a component in the chain graph, giving the correct form for the marginal distribution on $p(\mathbf{y},\mathbf{x},\theta,\phi)$. As for the gradient matching method, this has no natural interpretation as a generative model of the data.

## 5. Experiments

We illustrate our framework on two benchmark systems, Lotka-Volterra and Signal Transduction Cascade in (Dondelinger et al., 2013). To aid comparison, wherever possible, we have chosen the same parameter settings and priors as the original authors. Our main interest is to study the implications of the different joint distributions specified by the competing approaches. As such we wish to make as similar as possible the sampling approaches for the three competing models in order to minimize differences due to different sampling strategies. To facilitate comparison we therefore used the same discretized sampling strategy for all methods. For the AGM and our GP-ODE approach we used Gibbs sampling for a discretized set of values, analogous to section(3.1) and the Gibbs scheme of section(3.2) for state samples. The cost of drawing a single sample in all competing approaches is similar, scaling $O(KT^3)$. We stopped each sampling scheme (all written in Matlab) after a similar CPU time. Following (Dondelinger et al., 2013), we set $p(\theta)$ to a Gamma prior $Ga(4, 0.5)$, $p(\phi)$ to a uniform prior $U(0, 100)$ and $p(\sigma)$ to a Gamma prior $Ga(1, 1)$. For the sampling process, the standard deviations of the observation noise $\sigma$ in both models are initialized as the ground truth. For comparison we ran the Bayesian Numerical Integration approach using the same discretized parameter values wherever possible.

### 5.1. Lotka-Volterra

The Lotka-Volterra model is an ecological system that is used to describe the periodical interaction between a prey species $[S]$ and a predator species $[W]$:

$$\frac{d[S]}{dt} = [S](\alpha - \beta[W]), \qquad \frac{d[W]}{dt} = -[W](\gamma - \delta[S])$$

where $\theta = [\alpha, \beta, \gamma, \delta]^T$ and $\mathbf{x}(t) = [[S], [W]]$. The ground truth data are generated using numerical integration over the interval [0, 2] with $\theta = [2, 1, 4, 1]$ and initial state values $[S] = 5$, $[W] = 3$. The clean data are then sampled with the sampling interval 0.2. Finally clean data are corrupted with additive Gaussian noise $\mathcal{N}(\mathbf{0}, 0.5^2\mathbf{I})$ to form the observations $\mathbf{Y}$. We chose the squared-exponential covariance function $c_{\phi_k}(t,t') = \sigma_k^x \exp(-l_k(t-t')^2)$ where $\phi_k = [\sigma_k^x, l_k]$. Assuming a common parameter across observation dimensions, the parameter vector $\phi$ is simplified to $[\sigma^x, l]$; we initialize it as $[1, 10]$. The parameters are initialized as $\theta = [1.5, 0.5, 3.5, 0.5]$. We discretized the ODE parameters $\alpha$, $\beta$, $\gamma$, $\delta$ over [1.5, 2.5], [0.5, 1.5], [3.5, 4.5], [0.5, 1.5] all with the interval 0.1; the parameter $\sigma^x$ is discretized over the range [0.1,1] with interval 0.1; the lengthscale $l$ is discretized over [5, 50] with interval 5; the standard deviation of the noise $\sigma$ was discretized over [0.1, 1] with interval 0.1. The parameter $x_0$ was, for each state dimension $k$, discretized in the range $[1, 10]$ us-

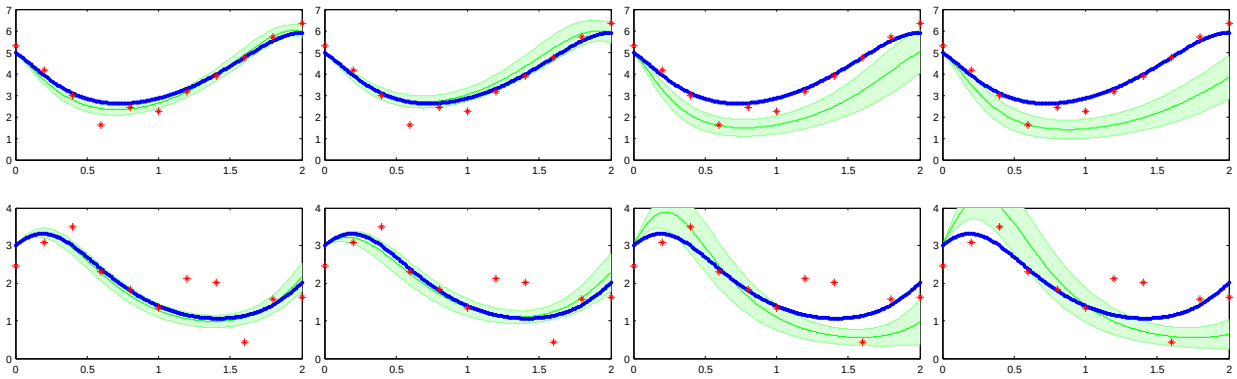


*Figure 3.* Bayesian Inference for Lotka-Volterra. The results for prey and predator are respectively shown in the first and second row. In all the plots, observations are red stars and the ground truth is the blue curve. Plotted in green are the reconstructions using the posterior samples of $\theta$. To aid comparison between the approaches we numerically integrated the ODE starting from the true initial state $[5, 3]$ with each curve generated by numerical integration using a parameter sample $\theta$. The green plots are the mean and one standard deviation of these resulting curves. First column: Bayesian Numerical Integration. This represents the solution that we wish to approximate. Second column: Our GP-ODE method. Third column: The Adaptive Gradient Matching method. Fourth column: The Gradient Matching method. All competing methods were run for approximately 300s of CPU time.

ing 20 uniformly spaced bins. After drawing ODE parameters $\theta$ from the posterior (see table(5.1)), we plot the numerically integrated curves (setting $x_0$ to the true value to aid visual comparison), see figure(3). The 'best' method is that which most closely approximates the Bayesian Numerical Integration method of section(1.3). For small noise levels (not shown), all three competing methods produce similar results; however as the noise increases, the Adaptive Gradient Matching and Gradient Matching approaches diverge markedly from the Bayesian Numerical Integration approach, whilst the GP-ODE approach fairs well.

### 5.2. Signal Transduction Cascade

The Signal Transduction Cascade model is described by a 5-dimensional ODE system

$$\begin{aligned}
\frac{d[S]}{dt} &= -k_1[S] - k_2[S][R] + k_3[RS] \\
\frac{d[S_d]}{dt} &= k_1[S] \\
\frac{d[R]}{dt} &= -k_2[S][R] + k_3[RS] + \frac{V[Rpp]}{Km + [Rpp]} \\
\frac{d[RS]}{dt} &= k_2[S][R] - k_3[RS] - k_4[RS] \\
\frac{d[Rpp]}{dt} &= k_4[RS] - \frac{V[Rpp]}{Km + [Rpp]}
\end{aligned}$$

where $\theta = [k_1, k_2, k_3, k_4, V, Km]$ and $\mathbf{x}(t) = [[S], [S_d], [R], [RS], [Rpp]]^T$. The ground truth data are generated over the interval [0, 100] with $\theta = [0.07, 0.6, 0.05, 0.3, 0.017, 0.3]$ and initial state $[S] = 1$, $[S_d] = 0$, $[R] = 1$, $[RS] = 0$, $[Rpp] = 0$. Then the data are sampled at $t$= [0, 1, 2, 4, 5, 7, 10, 15, 20, 30, 40, 50, 60, 80, 100]. Finally the drawn samples are corrupted with additive Gaussian noise $\mathcal{N}(\mathbf{0}, 0.1^2\mathbf{I})$ to construct the noisy observations. The non-stationarity is captured by the covariance function[10]

$$c_{\phi_k}(t, t') = \sigma_k^x \arcsin\left(\frac{a_k + b_k t t'}{\sqrt{(a_k + b_k t^2 + 1)(a_k + b_k t'^2 + 1)}}\right)$$

where $\phi_k = [\sigma_k^x, a_k, b_k]$. The ODE parameters are initialized as $\theta = [0.05, 0.4, 0.03, 0.1, 0.015, 0.1]$. We discretized the ODE parameters $k_1$, $k_2$, $k_3$, $k_4$, $V$, $Km$ over [0.05, 0.09], [0.4, 0.8], [0.03, 0.07], [0.1, 0.5], [0.015, 0.019], [0.1, 0.5] with the respective intervals 0.01, 0.1, 0.01, 0.1, 0.001, 0.1; the parameters $\sigma^x$, $a$, $b$ over [0.1,0.9], [0.5, 2.5], [0.5, 2.5] with the respective intervals 0.2, 0.5, 0.5; the standard deviations of the noise $\sigma$ over [0.06, 0.14] with the interval 0.02. The 5 components of $x_0$ were discretized in the intervals [0.5 1.5], [-0.1 0.1], [0.5 1.5], [-0.1 0.1], [-0.1 0.1] using 50 uniformly spaced bins. All three competing approaches were run for approximately 30mins CPU time. All three methods produce reasonable solutions and the reconstructions using numerical integration with parameters $\theta$ sampled from the respective posteriors are similar. As such we show only the ideal Bayesian Numerical Integration procedure and our GP-ODE method in figure(4). In table(5.1) the GP-ODE method closely matches the Bayesian Numerical Integration approach, with the Gradient Matching and Adaptive Gradient Matching approaches producing broadly similar parameter estimates.

[10] In contrast to the Lotka Volterra model, here we use a GP with separate hyperparameters for each state dimension due to the different length scales in each dimension.

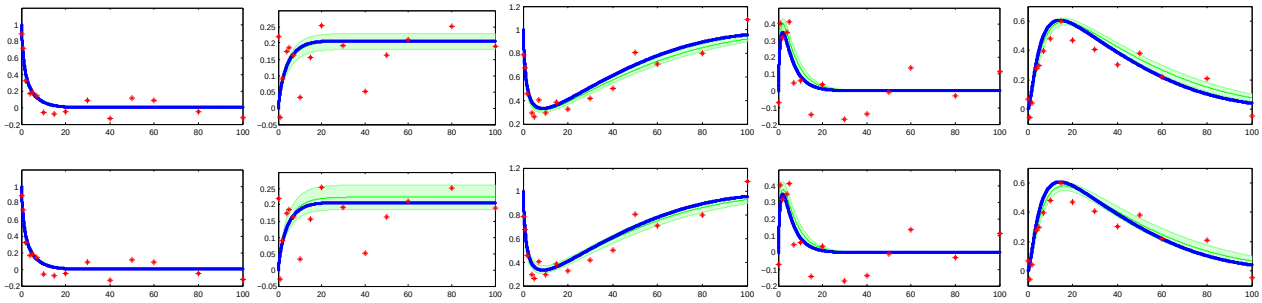


*Figure 4.* Bayesian Numerical Integration (top row) and GP-ODE results (bottom row) for the Signal Transduction Cascade. The results for $[R]$, $[S_d]$, $[R]$, $[RS]$, $[Rpp]$ (from left to right). In all the plots, observations are red stars and the ground truth is the blue curve. The green curves show reconstructions using sampled parameters $\theta$ from each posterior. Each curve is obtained by numerical integration using the sampled parameter, starting from the same initial point $x_0 = [1, 0, 1, 0, 0]$.

| PARAMETER | TRUE VALUE | NUM-BAYES | GP-ODE | AGM | GM |
|---|---|---|---|---|---|
| $\alpha$ | 2 | 2.2680±0.1853 | 2.2380±0.1953 | 1.9480±0.2819 | 1.6429±0.2488 |
| $\beta$ | 1 | 1.2070±0.1249 | 1.1490±0.1910 | 1.1750±0.1909 | 0.9242±0.2256 |
| $\gamma$ | 4 | 3.8330±0.2640 | 3.9590±0.3002 | 3.8390±0.2947 | 3.6449±0.2223 |
| $\delta$ | 1 | 0.9850±0.0857 | 0.9860±0.0995 | 1.2320±0.1638 | 1.1737±0.1803 |
| $k_1$ | 0.07 | 0.0683±0.0122 | 0.0747±0.0130 | 0.0771±0.0130 | 0.0762±0.0130 |
| $k_2$ | 0.6 | 0.6800±0.0876 | 0.6230±0.1246 | 0.5460±0.1259 | 0.5632±0.1256 |
| $k_3$ | 0.05 | 0.0548±0.0125 | 0.0530±0.0135 | 0.0593±0.0111 | 0.0594±0.0115 |
| $k_4$ | 0.3 | 0.2290±0.0498 | 0.2960±0.0281 | 0.3750±0.0999 | 0.3754±0.1051 |
| $V$ | 0.017 | 0.0177±0.0012 | 0.0177±0.0014 | 0.0172±0.0015 | 0.0173±0.0014 |
| $Km$ | 0.3 | 0.3860±0.0792 | 0.4220±0.0690 | 0.4090±0.0911 | 0.4186±0.0953 |

*Table 1.* ODE Parameter Estimation for both the Lotka Volterra (upper) and Signal Transduction Cascade (lower) ODEs. In each we plot the mean and standard deviation of the $\theta$ parameter samples from the respective competing approaches, namely the 'ideal' Bayesian Numerical Integration procedure, our GP-ODE approach, the Adaptive Gradient Matching approach (Dondelinger et al., 2013) and the Gradient Matching approach (Calderhead et al., 2008).

## 6. Discussion

Bayesian parameter estimation in ODEs using numerical integration is an ideal but computationally prohibitive method for all but the smallest systems due to the high cost of explicit numerical integration required to evaluate a single point in the posterior. Any other model will necessarily make assumptions that are formally inconsistent with this ideal approach and the aim is therefore to trade the accuracy of matching the ideal Bayesian numerical integration approach for computational speed.

Whilst previous alternatives based on using Gaussian Processes have demonstrated some success in circumventing the high computational cost, these are not natural generative models of the data. Despite the improvements in (Dondelinger et al., 2013), previous GP approaches use a heuristic compatibility function that leads to a complex chain graph with unknown normalisation constant.

In contrast, our method has a natural link to numerical integration and is we believe conceptually the closest match amongst competing GP approaches to the ideal numerical integration mechanism. Our GP-ODE approach is a simple generative model of data and as such is amenable to alternative approximation techniques, other than Monte Carlo. For example, variational approximations are in principle possible to apply directly to the posterior.

In our experience on toy benchmark problems, our GP-ODE approach performs at least as well as alternative GP approaches and sometimes significantly better, particularly in the case of observations with appreciable noise. Code is available from github.com/odegp/code.

### Acknowledgements

We would like to thank Dirk Husmeier and Benn Macdonald for helpful discussions and provision of their code.

## References

Barber, D. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

Calderhead, B., Girolami, M., and Lawrence, N. D. Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes. In *NIPS*, 2008.

Dondelinger, F., Filippone, M., Rogers, S., and Husmeier, D. ODE parameter inference using adaptive gradient matching with Gaussian processes. In *AISTATS*, 2013.

Graepel, T. Solving noisy linear operator equations by Gaussian processes: application to ordinary and partial differential equations. In *ICML*, 2003.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Technique*. MIT Press, 2009.

Ramsay, J., Hooker, G., Campbell, D., and Cao, J. Parameter Estimation for Differential Equations: A Generalized Smoothing Approach. *Journal of the Royal Statistical Society: Series B*, 69(5):741–796, 2007.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. Derivative observations in Gaussian Process models of dynamic systems. In *NIPS*, 2002.

Vyshemirsky, V. and Girolami, M. Bayesian ranking of biochemical system models. *Bioinformatics*, 24:833–839, 2008.