
Fast Gaussian process based gradient matching for parameter identification in systems of nonlinear ODEs

Philippe Wenk^{1,2}
Nico S. Gorbach¹
¹ETH Zürich

Alkis Gotovos¹
Andreas Krause¹
²Max Planck ETH CLS

Stefan Bauer^{2,3}
Joachim M. Buhmann¹
³ MPI-IS Tübingen

Abstract

Parameter identification and comparison of dynamical systems is a challenging task in many fields. Bayesian approaches based on Gaussian process regression over time-series data have been successfully applied to infer the parameters of a dynamical system without explicitly solving it. While the benefits in computational cost are well established, the theoretical foundation has been criticized in the past. We offer a novel interpretation which leads to a better understanding and improvements in state-of-the-art performance in terms of accuracy, robustness and a decrease in run time due to a more efficient setup for general nonlinear dynamical systems.

1 INTRODUCTION

The underlying mechanism of many processes in science and engineering can often be described by ordinary differential equations (ODE). While the form of dynamical systems, the ODEs, can often be derived using expert knowledge, the parameters are usually unknown, can not be directly measured and have to be estimated from empirical time series data. Since nonlinear ODEs typically do not have a closed form solution, standard methods for statistical inference require the computationally expensive numerical integration of the ODEs every time the parameters are changed (Calderhead et al., 2008).

To circumvent the high computational cost of numerical integration, gradient matching techniques have been proposed (e.g. Ramsay et al., 2007; Dondelinger

et al., 2013; Niu et al., 2016; Gorbach et al., 2017a). Gradient matching is based on minimizing the difference between a model interpolating the dynamics of the state variables and the time derivatives provided by the ODEs. The first steps of this approach go back to spline-based methods, with an overview given by Ramsay et al. (2007). Calderhead et al. (2008) then proposed a fully probabilistic treatment by using Gaussian process models, increasing data efficiency and providing final uncertainty estimates. To match the gradients of the GP model to the gradients provided by the ODEs, Calderhead et al. (2008) introduce a product of experts heuristics (PoE). This heuristic has since become the state-of-the-art explanation, being reused, e.g., in Dondelinger et al. (2013) and Gorbach et al. (2017a). The advantages of the method have been well established, with applications in biomedical domains e.g., systems biology or neuroscience (Babtie et al., 2014; Macdonald and Husmeier, 2015; Pfister et al., 2018). Given these applications, a solid understanding and theoretical framework of these methods is of critical importance.

However, there are two important open issues. Regarding the theoretical framework, there has been some criticism regarding the use of the PoE in this context, e.g., by Wang and Barber (2014), who introduced a different modeling paradigm. Regarding empirical performance, Gorbach et al. (2017a) recently introduced a mean field variational scheme to decrease the run time. However, despite the expected run time savings, Gorbach et al. (2017a) also reported a significantly increased accuracy of the variational approach compared to the MCMC sampling based counterparts.

While the criticism of the product of experts approach (Wang and Barber, 2014) lead to the controversy of mechanistic modeling with Gaussian processes (Macdonald et al., 2015) where the theoretical inconsistencies of the modeling proposal of Wang and Barber (2014) have been outlined, a similar analysis investigating the theoretical foundation of the state of the art approach underlying all current modeling

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

proposals (Calderhead et al., 2008; Dondelinger et al., 2013; Gorbach et al., 2017a) has been missing.

Our contributions. In this work, we

1. analyze the Product of Experts (PoE) heuristic, discovering and explaining theoretical inconsistencies of the state of the art approach,
2. provide a theoretical framework, replacing the criticized product of experts heuristic,
3. identify the cause of the performance gains of the variational approach of Gorbach et al. (2017a) over sampling-based methods,
4. combine these insights to create a novel algorithm improving on state-of-the-art performance for nonlinear systems in terms of accuracy and robustness, while providing a more computationally efficient sampling scheme reducing its run time by roughly 35%.¹

2 PRELIMINARIES

2.1 Problem Formulation

In this work, we analyze an arbitrary system of parametric ordinary differential equations, whose state derivatives can be parametrized by a time independent parameter vector $\boldsymbol{\theta}$. In this setting, the true evolution of the dynamical system is given by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) \quad (1)$$

where \mathbf{x} are the time dependent states of the system and \mathbf{f} is an arbitrary, potentially nonlinear vector valued function.

While Equation (1) is meant to represent the dynamics of the system at all time points, \mathbf{y} , \mathbf{x} and $\dot{\mathbf{x}}$ will be used throughout this paper to denote the vector containing the time evolution of one state or one state derivative at the observation times \mathbf{t} , i.e. $\mathbf{x} = [x_0(t_0), \dots, x_0(t_N)]$ etc.

At N discrete time points \mathbf{t} , the state trajectories are observed with additive, i.i.d. Gaussian noise $\boldsymbol{\epsilon}(t_i) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, i.e.,

$$\mathbf{y}(t_i) = \mathbf{x}(t_i) + \boldsymbol{\epsilon}(t_i), \quad i = 1 \dots N, \quad (2)$$

or equivalently

$$p(\mathbf{y}|\mathbf{x}, \sigma) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2 \mathbf{I}). \quad (3)$$

Given these noisy observations \mathbf{y} and the functional form of \mathbf{f} , the goal is to infer the true parameters $\boldsymbol{\theta}$.

¹Code publicly available at <https://github.com/wenkph/FGPGM/>

For the sake of clarity, we present the theory for a one-dimensional system only and assume that all observations were created using one realization of the experiment. The extension to multidimensional systems (as done in the experiments, Section 5) or repeated experiments is straightforward but omitted to simplify the notation.

2.2 Modeling

The key idea of GP-based gradient matching is to build a GP regression model mapping the time points to the corresponding state values. For this, one needs to choose an appropriate kernel function $k_\phi(t_i, t_j)$, which is parametrized by the hyperparameters ϕ . Both, the choice of kernels as well as how to fit its hyperparameters is discussed in Section 7.3.

Once the kernel and its hyperparameters are fixed, the covariance matrix \mathbf{C}_ϕ , whose elements are given by $\mathbf{C}_\phi(i, j) = k(t_i, t_j)$, can be constructed and used to define a standard zero mean Gaussian process prior on the states:

$$p(\mathbf{x}|\phi) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{C}_\phi). \quad (4)$$

As differentiation is a linear operation, the derivative of a Gaussian process is again a Gaussian process. Using probabilistic calculus (see supplementary material, Section 7.1 for details), this fact leads to a distribution over the derivatives conditioned on the states at the observation points:

$$p(\dot{\mathbf{x}}|\mathbf{x}, \phi) = \mathcal{N}(\dot{\mathbf{x}}|\mathbf{D}\mathbf{x}, \mathbf{A}). \quad (5)$$

Lastly, the information provided by the differential equation is used as well. For known states and parameters, one can calculate the derivatives using equation (1). A potential modeling mismatch between the output of the ODEs and the derivatives of the GP model is accounted for by introducing isotropic Gaussian noise with standard deviation γ , leading to the following Gaussian distribution over the derivatives:

$$p(\dot{\mathbf{x}}|\mathbf{x}, \boldsymbol{\theta}, \gamma) = \mathcal{N}(\dot{\mathbf{x}}|f(\mathbf{x}, \boldsymbol{\theta}), \gamma \mathbf{I}). \quad (6)$$

The modeling assumptions are summarized in the graphical models shown in Figure 1.

2.3 Inference

As stated in Section 2.1, the main goal of the inference process is to learn the parameters $\boldsymbol{\theta}$ using the noisy observations \mathbf{y} . Thus, it is necessary to connect the two graphical models shown in Figure 1. As shown in the previous section, \mathbf{x} and the $\dot{\mathbf{x}}$ represent the same variables in both models. However, it is not straightforward to use this fact to combine the two. While

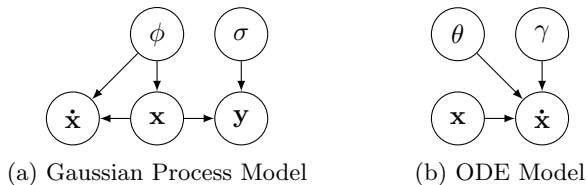


Figure 1: Modeling assumptions of Gaussian Process based gradient matching.

it is intuitive to use the probability density over \mathbf{x} of the Gaussian process model directly as the prior for \mathbf{x} in the ODE response model, handling $\dot{\mathbf{x}}$ is more challenging. In both models, $\dot{\mathbf{x}}$ is a dependent variable. Thus, some heuristic is needed to combine the two conditional distributions $p(\dot{\mathbf{x}}|\mathbf{x}, \phi)$ and $p(\dot{\mathbf{x}}|\mathbf{x}, \theta, \gamma)$.

2.3.1 Product of experts heuristic

The main idea of the product of experts, originally introduced by Hinton (2002), is to infer the probability density of a variable by normalizing the product of multiple expert densities. Calderhead et al. (2008) use this to connect the two distributions over $\dot{\mathbf{x}}$, leading to

$$p(\dot{\mathbf{x}}|\mathbf{x}, \phi, \theta, \gamma) \propto p(\dot{\mathbf{x}}|\mathbf{x}, \phi)p(\dot{\mathbf{x}}|\mathbf{x}, \theta, \gamma) \quad (7)$$

The idea of this approach is that the resulting density only assigns high probability if both experts assign high probabilities. Hence, it considers only cases in which both experts agree. It is thus based on the intuition that the true θ should correspond to a model that agrees both with the ODE model and the observed data. While this is intuitively well-motivated, we will show that the product of experts heuristic leads to theoretical difficulties and offer an alternative.

2.3.2 Markov Chain Monte Carlo based methods

Calderhead et al. (2008) combine the product of experts with equations (3), (4) and (5) and some suitable prior over θ to obtain a joint distribution $p(\mathbf{x}, \dot{\mathbf{x}}, \theta, \phi, \sigma|\mathbf{y})$. After integrating out $\dot{\mathbf{x}}$, which can be done analytically since Gaussian processes are closed under linear operators (and using some proportionality arguments), a sampling scheme was derived that consists of two MCMC steps. First, the hyperparameters of the GP, ϕ and σ , are sampled from the conditional distribution $p(\phi, \sigma|\mathbf{y})$. Then, a second MCMC scheme is deployed to infer the parameters of the ODE model, θ and γ , by sampling from the conditional distribution $p(\theta, \gamma|\mathbf{x}, \phi, \sigma)$.

Dondelinger et al. (2013) then reformulated the approach by directly calculating the joint distribution

$$\begin{aligned} p(\mathbf{y}, \mathbf{x}, \theta, \phi, \gamma, \sigma) &\propto p(\theta) \\ &\times \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{C}_\phi) \\ &\times \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \\ &\times \mathcal{N}(\mathbf{f}(\mathbf{x}, \theta)|\mathbf{D}\mathbf{x}, \mathbf{A} + \gamma\mathbf{I}), \end{aligned} \quad (8)$$

where the proportionality is meant to be taken w.r.t. the latent states \mathbf{x} and the ODE parameters θ . Here $p(\theta)$ denotes some prior on the ODE parameters. This approach was named Adaptive Gradient Matching (AGM).

2.3.3 Variational inference

The main idea of Variational Gradient Matching (VGM), introduced by Gorbach et al. (2017a), is to substitute the MCMC inference scheme of AGM with a mean field variational inference approach, approximating the density in Equation (8) with a fully factorized Gaussian over the states \mathbf{x} and the parameters θ . To obtain analytical solutions, the functional form of the ODEs is restricted to locally linear functions that could be written as

$$f(\mathbf{x}, \theta) = \sum_i \theta_i \prod_{j \in \mathcal{M}_i} x_j \quad \text{where } \mathcal{M}_i \subseteq \{1, \dots, K\}. \quad (9)$$

As perhaps expected, VGM is magnitudes faster than the previous sampling approaches. However, despite being a variational approach, VGM was also able to provide significantly more accurate parameter estimates than both sampling-based approaches of Calderhead et al. (2008) and Dondelinger et al. (2013). In Section 4, we provide justification for these surprising performance differences.

3 THEORY

3.1 Analysis Of The Product Of Experts Approach

As previously stated, Calderhead et al. (2008), Dondelinger et al. (2013) and Gorbach et al. (2017a) all use a product of experts to obtain $p(\dot{\mathbf{x}}|\mathbf{x}, \phi, \theta, \gamma)$ as stated in Equation (7).

In this section, we will first provide an argument based on graphical models and then an argument based on the original mathematical derivation to illustrate challenges arising from this heuristic.

Figure 2 depicts what is happening if the product of experts approach is applied in the gradient matching framework. Figure 2a depicts the graphical model after the two models have been merged using the product



Figure 2: The product of experts approach as a graphical model. After marginalization of $\dot{\mathbf{x}}$, the parameters θ we would like to infer are independent of the observations \mathbf{y} .

of experts heuristic of Equation (7). Using the distribution over \mathbf{x} of the Gaussian process model 1a as a prior for the \mathbf{x} in the ODE response model 1b, effectively leads to merging the two nodes representing \mathbf{x} . Furthermore, the product of experts heuristic implies by its definition that after applying Equation (7), $\dot{\mathbf{x}}$ is only depending on \mathbf{x} , ϕ , θ and γ .

In the graphical model in Figure 2a, the problem is already visible. The ultimate goal of merging the two graphical models is to create a probabilistic link between the observations \mathbf{y} and the ODE parameters θ . However, the newly created connection between these two variables is given via $\dot{\mathbf{x}}$, which has no outgoing edges and of which no observations are available. Marginalizing out $\dot{\mathbf{x}}$ as proposed in the traditional approaches consequently leads to the graphical model in Figure 2b. As there is no directed path connecting other variables via $\dot{\mathbf{x}}$, all the different components are now independent. Consequently, the posterior over θ is now given by the prior we put on θ in the first place.

This problem can further be illustrated by the mathematical derivations in the original paper of Calderhead et al. (2008). After calculating the necessary normalization constants, the last equation in the third chapter is equivalent to stating

$$p(\theta, \gamma | \mathbf{x}, \phi, \sigma) = \int p(\theta) p(\gamma) p(\dot{\mathbf{x}} | \mathbf{x}, \theta, \gamma, \phi, \sigma) d\dot{\mathbf{x}}. \quad (10)$$

It is clear that this equation should simplify to

$$p(\theta, \gamma | \mathbf{x}, \phi, \sigma) = p(\theta) p(\gamma). \quad (11)$$

Thus, one could argue that any links that are not present in the graphical model of Figure 2b but found by Calderhead et al. (2008) and reused in Dondelinger et al. (2013) and Gorbach et al. (2017a) were created by improper normalization of the density $p(\dot{\mathbf{x}} | \mathbf{x}, \theta, \gamma, \phi, \sigma)$.

3.2 Adapting The Original Graphical Model

Despite these technical difficulties arising from the PoE heuristic, the approaches provide good empirical results and have been used in practice, e.g., by Babbie

et al. (2014). In what follows, we derive an alternative model and mathematical justification for Equation (8) to provide a theoretical framework explaining the good empirical performance of Gaussian process-based gradient matching approaches, especially from Gorbach et al. (2017a), which uses only weak or nonexistent priors.

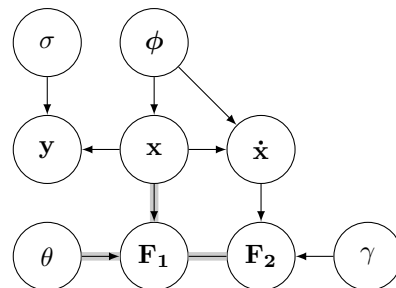


Figure 3: Alternative probabilistic model without PoE heuristic. Gray shaded connections are used to indicate a deterministic relationship.

The graphical model shown in Figure 3 offers an alternative approach to the product of experts heuristic. The top two layers are equivalent to a GP prior on the states, the induced GP on the derivatives and the observation model, as shown in Figure 1a.

The interesting part of the new graphical model is the bottom layer. Instead of adding a second graphical model like in Figure 1b to account for the ODE response, two additional random variables are introduced.

\mathbf{F}_1 is the deterministic output of the ODEs, assuming the values of \mathbf{x} and θ are given, i.e. $\mathbf{F}_1 = \mathbf{f}(\mathbf{x}, \theta)$. The deterministic nature of this equation is represented as a Dirac delta function:

$$p(\mathbf{F}_1 | \mathbf{x}, \theta) = \delta(\mathbf{F}_1 - \mathbf{f}(\theta, \mathbf{x})) \quad (12)$$

If the GP model were able to capture the true states and true derivatives perfectly, this new random variable should be equivalent to the derivatives of the GP, i.e., $\mathbf{F}_1 = \dot{\mathbf{x}}$. However, to compensate for a potential model mismatch and slight errors of both GP states

and GP derivatives, this condition is relaxed to

$$\mathbf{F}_1 = \dot{\mathbf{x}} + \boldsymbol{\epsilon} =: \mathbf{F}_2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \gamma \mathbf{I}) \quad (13)$$

In the graphical model, this intuitive argument is encoded via the random variable \mathbf{F}_2 . Given the $\dot{\mathbf{x}}$ provided by the GP model, Gaussian noise with standard deviation γ is added to create \mathbf{F}_2 , whose probability density can thus be described as

$$p(\mathbf{F}_2 | \dot{\mathbf{x}}, \gamma) = \mathcal{N}(\mathbf{F}_2 | \dot{\mathbf{x}}, \gamma \mathbf{I}). \quad (14)$$

The equality constraint given by equation (13) is represented in the graphical model by the undirected edge between \mathbf{F}_1 and \mathbf{F}_2 . When doing inference, this undirected edge is incorporated in the joint density via a Dirac delta function $\delta(\mathbf{F}_2 - \mathbf{F}_1)$. Thus, the joint density of the graphical model represented in Figure 3 can be written as

$$\begin{aligned} p(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{y}, \mathbf{F}_1, \mathbf{F}_2, \boldsymbol{\theta} | \phi, \sigma, \gamma) &= p(\boldsymbol{\theta}) \\ &\times p(\mathbf{x} | \phi) p(\dot{\mathbf{x}} | \mathbf{x}, \phi) p(\mathbf{y} | \mathbf{x}, \sigma) \\ &\times p(\mathbf{F}_1 | \boldsymbol{\theta}, \mathbf{x}) p(\mathbf{F}_2 | \dot{\mathbf{x}}, \gamma \mathbf{I}) \delta(\mathbf{F}_1 - \mathbf{F}_2). \end{aligned} \quad (15)$$

3.3 Inference In The New Model

Given all the definitions in the previous section, inference can now be directly performed without the need for additional heuristics. The result is a theoretically sound justification of the main result of Calderhead et al. (2008) and Dondelinger et al. (2013):

Theorem 1 *Given the modeling assumptions summarized in the graphical model in Figure 3,*

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}, \phi, \gamma, \sigma) &\propto p(\boldsymbol{\theta}) \\ &\times \mathcal{N}(\mathbf{x} | \mathbf{0}, \mathbf{C}_\phi) \\ &\times \mathcal{N}(\mathbf{y} | \mathbf{x}, \sigma^2 \mathbf{I}) \\ &\times \mathcal{N}(\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) | \mathbf{D}\mathbf{x}, \mathbf{A} + \gamma \mathbf{I}). \end{aligned} \quad (16)$$

The proof can be found in the supplementary material, section 7.2.

4 FAST GAUSSIAN PROCESS GRADIENT MATCHING

4.1 Hyperparameters

Using the theoretical framework of the previous section, it is clear that the performance of any algorithm based on equation 16 will heavily rely on the quality of the hyperparameters ϕ , as \mathbf{C}_ϕ , \mathbf{D} and \mathbf{A} are all depending on ϕ . In GP regression, it is common to fit the hyperparameters to the observations using a maximum likelihood scheme (see supplementary material,

section (7.3)). However, neither VGM nor AGM actually do this.

In AGM, the hyperparameters are inferred concurrently to the states \mathbf{x} and the ODE parameters $\boldsymbol{\theta}$ in one big MCMC scheme. Besides the need for clear hyperpriors on the hyperparameters and obvious drawbacks regarding running time, e.g. at each MCMC step \mathbf{C}_θ , \mathbf{D} and \mathbf{A} have to be recalculated and potentially inverted, this leads to a rough probability landscape requiring a complicated multi-chain setup (Dondelinger et al., 2013).

In VGM, this problem was sidestepped by treating the hyperparameters as tuning parameters that had to be set by hand. As these hyperparameters are not learned from data, it is obviously not a fair comparison. In our experiments, we thus used a maximum likelihood scheme to infer the hyperparameters before using the implementation of Gorbach et al. (2017a). To indicate this modification, the modified VGM algorithm was called MVGM.

4.2 FGPGM

However, MVGM is still outperforming AGM significantly as shown in the experiments in section 5. This suggests that the concurrent optimization of ϕ , \mathbf{x} and $\boldsymbol{\theta}$ suggested by Dondelinger et al. (2013) is not helpful for the performance of the system. Based on this insight, we propose the sequential approach shown in Algorithm 1. In a first step, the Gaussian process model is fit to the standardized data by calculating the hyperparameters via equation (36). Then, the states \mathbf{x} and ODE parameters $\boldsymbol{\theta}$ are inferred using a one chain MCMC scheme on the density given by equation (8).

5 EXPERIMENTS

For all experiments involving AGM, the R toolbox deGradInfer (Macdonald and Dondelinger, 2017) published alongside (Macdonald, 2017) was used. The toolbox was provided with code to run two experiments, namely Lotka Volterra and Protein Transduction. Both of these systems are used in this paper, as they are the two standard benchmark systems used in all previous publications. It should be noted however that the applicability of FGPGM is not restricted to these systems. Unlike AGM, FGPGM refrains from using hard to motivate hyperpriors and our publicly available implementation can easily be adapted to new settings.

For details about implementation and additional plots, refer to the supplementary material in section 7.6. It should be noted however that all algorithms were provided with one hyperparameter. While the toolbox of

Algorithm 1 FGPGM

```

1: Input:  $\mathbf{y}, \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \gamma, N_{MCMC}, N_{burnin}, \mathbf{t}, \sigma_s, \sigma_p$ 
2: Step 1: Fit GP model to data
3: for all  $k \in K$  do
4:    $\boldsymbol{\mu}_{y,k} \leftarrow \text{mean}(\mathbf{y}_k)$ 
5:    $\sigma_{y,k} \leftarrow \text{std}(\mathbf{y}_k)$ 
6:    $\tilde{\mathbf{y}} \leftarrow (\mathbf{y}_k - \boldsymbol{\mu}_k) / \sigma_{y,k}$ 
7:   Find  $\phi_k$  and  $\sigma_k$  by maximizing
      $p(\tilde{\mathbf{y}} | \mathbf{t}, \phi_k, \sigma_k)$  given by equation (36)
8: end for
9: Step 2: Infer  $\mathbf{x}$  and  $\boldsymbol{\theta}$  using MCMC
10:  $\mathcal{S} \leftarrow \emptyset$ 
11: for  $i = 1 \rightarrow N_{MCMC} + N_{burnin}$  do
12:   for each state and parameter do
13:      $\mathcal{T} \leftarrow \emptyset$ 
     Propose a new state or parameter value
     adding a zero mean Gaussian increment
     with standard deviation  $\sigma_s$  or  $\sigma_p$ .
14:     Accept proposed value based on the
     density given by equation (8)
15:     Add current value to  $\mathcal{T}$ 
16:   end for
17:   Add the mean of  $\mathcal{T}$  to  $\mathcal{S}$ 
18: end for
19: Discard the first  $N_{burnin}$  samples of  $\mathcal{S}$ 
20: Return the mean of  $\mathcal{S}$ 
21: Return:  $\mathbf{x}, \boldsymbol{\theta}$ 

```

AGM had to be provided with the true standard deviation of the observation noise, MVGM and AGM were provided with γ . The γ was determined by testing eight different values logarithmically spaced between 1 and 10^{-4} and comparing the results based on observation fit. For more details regarding the experimental setup, we refer to the appendix of Wenk et al. (2019).

5.1 Lotka Volterra

The first system in question is the Lotka Volterra system originally proposed in Lotka (1978). It describes a two dimensional system whose dynamics are given by

$$\begin{aligned} \dot{x}_1(t) &= \theta_1 x_1(t) - \theta_2 x_1(t) x_2(t) \\ \dot{x}_2(t) &= -\theta_3 x_2(t) + \theta_4 x_1(t) x_2(t) \end{aligned}$$

We reproduce the experimental setup of Gorbach et al. (2017a), i.e., the system was observed in the time interval $[0, 2]$ at 20 evenly spaced observation times. The true states were initialized with $[5, 3]$ and Gaussian noise with standard deviation 0.1 (low noise) and standard deviation 0.5 (high noise) was added to create the observations.

5.2 Protein Transduction

The second system to be analyzed is called Protein Transduction and was originally proposed in Vyshemirsky and Girolami (2008). It is known to be notoriously difficult to fit with unidentifiable parameters (Dondelinger et al., 2013). The dynamics are given by:

$$\begin{aligned} \dot{S} &= -\theta_1 S - \theta_2 SR + \theta_3 R_S \\ \dot{dS} &= \theta_1 S \\ \dot{R} &= -\theta_2 SR + \theta_3 R_S + \theta_5 \frac{R_{pp}}{\theta_6 + R_{pp}} \\ \dot{R}_S &= \theta_2 SR - \theta_3 R_S - \theta_4 R_S \\ \dot{R}_{pp} &= \theta_4 R_S - \theta_5 \frac{R_{pp}}{\theta_6 + R_{pp}} \end{aligned} \quad (17)$$

It should be noted that these dynamics contain nonlinear terms violating the functional form assumption given in equation (9) of VGM inherited by MVGM. Nevertheless, both FGPGM and AGM can still be applied. For FGPGM, γ was set to 10^{-4} , while AGM was provided with the true observation noise standard deviations. The experimental setup of Dondelinger et al. (2013) was copied, i.e. the system was observed in the time interval $[0, 100]$ at the discrete observation times $\mathbf{t} = [0, 1, 2, 4, 5, 7, 10, 15, 20, 30, 40, 50, 60, 80, 100]$, the states were initialized with $\mathbf{x}(0) = [1, 0, 1, 0, 0]$ and the parameters were set to $\boldsymbol{\theta} = [0.07, 0.6, 0.05, 0.3, 0.017, 0.3]$. As in Dondelinger et al. (2013), Gaussian noise was added with standard deviation 0.001 (low noise) and 0.01 (high noise). As in the previous papers, a sigmoid kernel was used to deal with the logarithmically spaced observation times and the typically spiky form of the dynamics.

5.3 Evaluation

As the parameters of the Protein Transduction system are unidentifiable, comparing parameter values is not a good metric to rank approaches. Instead, after running the algorithms, we used a numerical integrator to obtain the trajectories corresponding to the dynamical system whose dynamics is given by the the inferred parameters. Then, the RMSE of these trajectories compared to the ground truth at the observation times were evaluated. For each experimental setting, this procedure was repeated for 100 different noise realizations. The results are shown in Figure 4. In Figure 5, we show median plots for the high noise setting of Lotka Volterra, while the running time between the state of the art AGM and FGPGM is shown in Figure 6.

While achieving run time savings of 35%, FGPGM

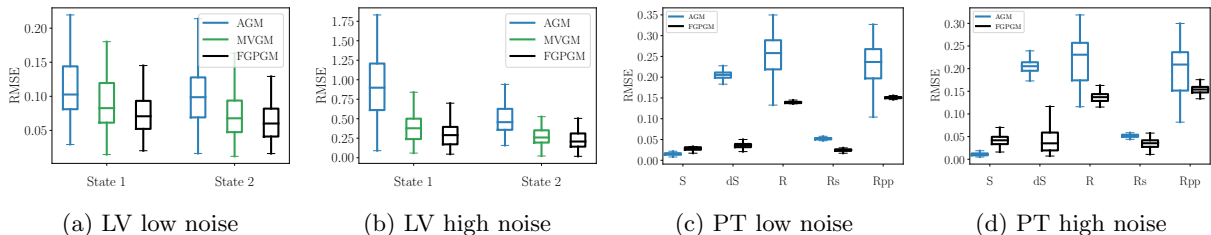


Figure 4: RMSE of the trajectories obtained by numerical integration compared to the ground truth. Boxplot with median (line), 50% (box) and 75% (whisker) quantiles over 100 independent noise realizations. Lotka Volterra is shown on the left, Protein Transduction on the right.

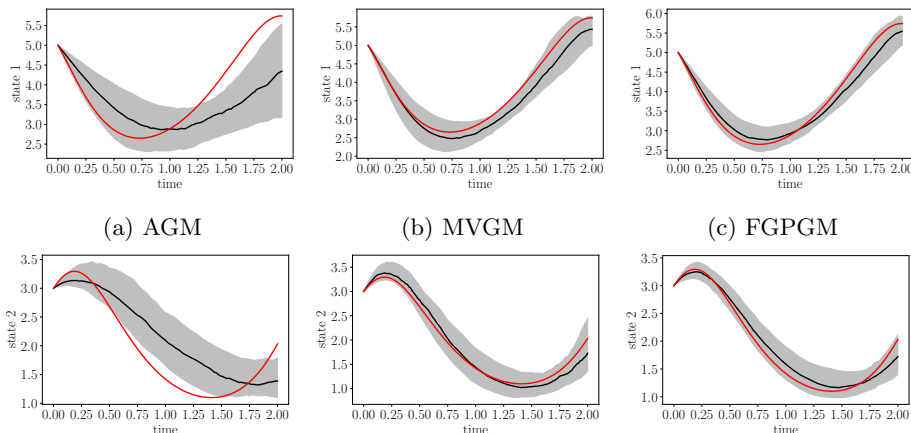


Figure 5: States after numerical integration of the inferred parameters in the high noise case of Lotka Volterra. Ground truth (red), median (black) and 75% quantiles (gray) over 100 independent noise realizations.

Table 1: Median reduction of state RMSE of FGPGM compared to AGM and VGM as baseline.

	AGM
LV low	35%
LV high	62%
PT low	50%
PT high	43%
	MVGM
LV low	13%
LV high	31%

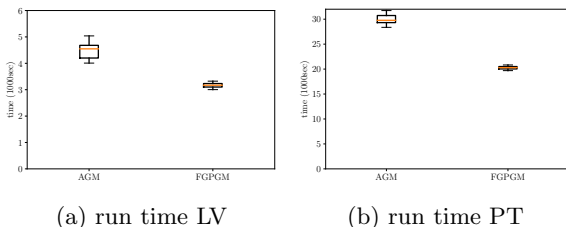


Figure 6: FGPGM shows a clearly reduced run time compared to AGM, saving roughly 35% of computation time.

shows an increased accuracy, reducing the state RMSE up to 62 % as shown in Table 1 and shows much lower variability across noise realizations. The effect is especially striking for the nonlinear system Protein Transduction, where the variance of the states R and R_{pp} has been reduced by at least one order of magnitude, while reducing the bias by more than 30%, see Figures 4c and 4d.

5.4 Spiky Dynamics

As can be seen in Figure 5, all GP based gradient matching algorithms converge to parameter settings where the trajectories are smoother than the ground truth. While learning the hyperparameters in a pre-processing step clearly reduces this effect for FGPGM and MVGM, there is still some bias. If only few observations are available, the GP prior on the states tends to smooth out part of the system dynamics. This "smoothing bias" is then passed on to the ODE parameters in the gradient matching scheme.

To investigate the practical importance of this effect, we evaluate FGPGM on a third system proposed by FitzHugh (1961) and Nagumo et al. (1962) for modeling giant squid neurons. Abbreviating the name of its inventors, we will refer to it as the FHN system. Its dynamics are given by the two ODEs

$$\dot{V} = \theta_1 \left(V - \frac{V^3}{3} + R \right) \quad (18)$$

$$\dot{R} = \frac{1}{\theta_1} (V - \theta_2 + \theta_3 R) \quad (19)$$

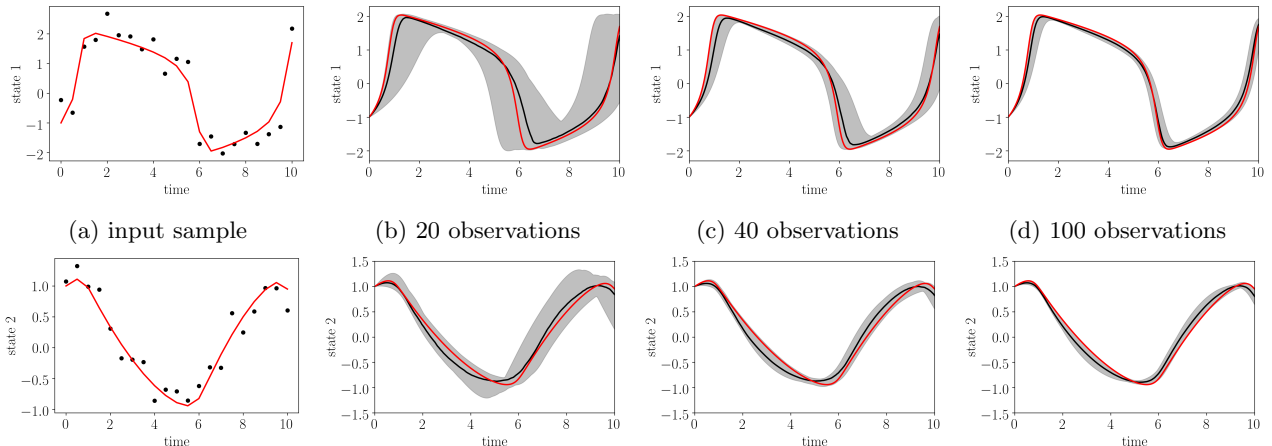


Figure 7: One input sample and median plots of the numerically integrated states after parameter inference for the FHN system with SNR 10. Ground truth (red), median (black) and 75% quantiles (gray) over 100 independent noise realizations. In the most left plots, the black dots represent the noisy observations.

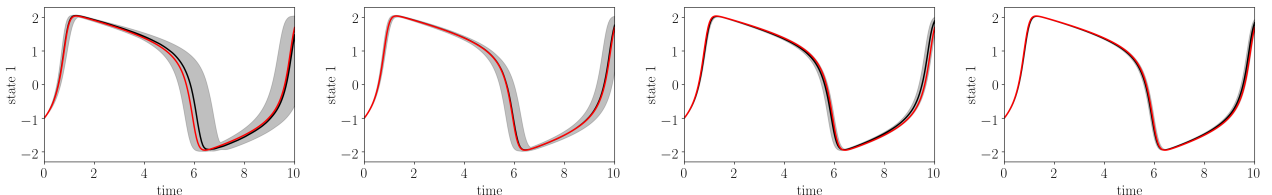


Figure 8: Median plots of the first state of FHN with SNR 100 in case of 10, 25, 50 and 100 observations. Same colors as in Figure 7. Plots of state 2 can be found in the supplementary material.

Due to its highly nonlinear terms, the FHN system has notoriously fast changing dynamics. One example realization including noisy observations is shown in Figure 7. To account for the spikier behavior of the system, we used a Matérn52 kernel. γ was set to $3 * 10^{-4}$.

While the bias towards smoother trajectories is clearly visible, FGPGM finds parameters that tightly hug the ground truth. As to be expected, the smoothing bias gets smaller if we add more observations. Furthermore, increasing the SNR to 100 as shown in Figure 8 leads to an impressive accuracy, even if we reduce the amount of observations to just 10, especially as FGPGM is a hyper-prior free, statistical method

6 DISCUSSION

Gradient matching is a successful tool to circumvent the computational cost of numerical integration for Bayesian parameter identification in dynamical systems, especially if the dynamics, like in most real world systems, are reasonably smooth. Previous Gaussian process-based approaches used a criticized product of experts heuristics, which leads to technical difficulties. We illustrated these theoretical problems and provided

a novel, sound formulation that does not rely on a PoE. We furthermore explained the surprising performance gains of variational over sampling-based approaches and then combined these insights to propose a new algorithm called FGPGM, which jointly learns states and parameters with improved state-of-the-art performance in terms of accuracy, run time, robustness and reduction of "smoothing bias" for general nonlinear dynamical systems.

Unlike the previous MCMC approaches, FGPGM uses a one-chain Metropolis Hastings scheme, which is much easier to tune than the previously used, complicated multi chain setups. Due to the sequential fitting of the hyperparameters, the wild behavior of the probability density motivating the setup in Dondelinger et al. (2013, section 3) and the need for hard to motivate hyperpriors is successfully avoided. Consequently, the inference is significantly more accurate, faster, more robust and significantly decreases the smoothing bias of GP based gradient matching schemes.

Acknowledgements

This research was partially supported by the Max Planck ETH Center for Learning Systems and SNSF grant 200020_159557.

References

- Ann C Babbie, Paul Kirk, and Michael PH Stumpf. Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences*, 111(52):18507–18512, 2014.
- Stefan Bauer, Nico S Gorbach, Djordje Miladinovic, and Joachim M. Buhmann. Efficient and flexible inference for stochastic systems. *Neural Information Processing Systems (NIPS)*, 2017.
- Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.
- Ben Calderhead, Mark Girolami and Neil D. Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *Neural Information Processing Systems (NIPS)*, 2008.
- Taeryon Choi and Mark J Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007.
- Frank Dondelinger, Maurizio Filippone, Simon Rogers and Dirk Husmeier. Ode parameter inference using adaptive gradient matching with gaussian processes. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.
- David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.
- David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.
- Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445–466, 1961.
- Javier González, Ivan Vujčić, and Ernst Wit. Reproducing kernel hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45:26–32, 2014.
- Nico S Gorbach, Stefan Bauer, and Joachim M. Buhmann. Scalable variational inference for dynamical systems. *Neural Information Processing Systems (NIPS)*, 2017a.
- Nico S Gorbach, Andrew An Bian, Benjamin Fischer, Stefan Bauer, and Joachim M Buhmann. Model selection for gaussian process regression. In *German Conference on Pattern Recognition*, pages 306–318. Springer, 2017b.
- Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.
- Alfred J Lotka. The growth of mixed populations: two species competing for a common food supply. In *The Golden Age of Theoretical Ecology: 1923–1940*, pages 274–286. Springer, 1978.
- Benn Macdonald. *Statistical inference for ordinary differential equations using gradient matching*. PhD thesis, University of Glasgow, 2017.
- Benn Macdonald and Frank Dondelinger. deGradInfer: Parameter inference for systems of differential equation. <https://CRAN.R-project.org/package=deGradInfer>, 2017.
- Benn Macdonald and Dirk Husmeier. Gradient matching methods for computational inference in mechanistic models for systems biology: a review and comparative analysis. *Frontiers in bioengineering and biotechnology*, 3, 2015.
- Benn Macdonald, Catherine F. Higham and Dirk Husmeier. Controversy in mechanistic model with gaussian processes. *International Conference on Machine Learning (ICML)*, 2015.
- Junichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.
- Mu Niu, Simon Rogers, Maurizio Filippone, and Dirk Husmeier. Fast inference in nonlinear dynamical systems using gradient matching. *International Conference on Machine Learning (ICML)*, 2016.
- Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.
- Niklas Pfister, Stefan Bauer, and Jonas Peters. Identifying causal structure in large-scale kinetic systems. *arXiv preprint arXiv:1810.11776*, 2018.
- Jim O Ramsay, Giles Hooker, David Campbell, and Jiguo Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.
- Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

- Vladislav Vyshemirsky and Mark A Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008.
- Yali Wang and David Barber. Gaussian processes for bayesian estimation in ordinary differential equations. *International Conference on Machine Learning (ICML)*, 2014.
- Philippe Wenk, Gabriele Abbati, Stefan Bauer, Michael A Osborne, Andreas Krause, and Bernhard Schölkopf. Odin: Ode-informed regression for parameter and state inference in time-continuous dynamical systems. *arXiv preprint arXiv:1902.06278*, 2019.
- Xinxin Zhang, Ruirui Ji, and Yingmin Yi. Inference of transcriptional regulation using multi-objective optimization with gaussian process. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on*, pages 1820–1825. IEEE, 2016.