Multiphase MCMC Sampling for Parameter Inference in Nonlinear Ordinary Differential Equations

Alan Lazarus **Dirk Husmeier** **Theodore Papamarkou**
School of Mathematics and Statistics, University of Glasgow, United Kingdom

Abstract

Traditionally, ODE parameter inference relies on solving the system of ODEs and assessing fit of the estimated signal with the observations. However, nonlinear ODEs often do not permit closed form solutions. Using numerical methods to solve the equations results in prohibitive computational costs, particularly when one adopts a Bayesian approach in sampling parameters from a posterior distribution. With the introduction of gradient matching, we can abandon the need to numerically solve the system of equations. Inherent in these efficient procedures is an introduction of bias to the learning problem as we no longer sample based on the exact likelihood function. This paper presents a multiphase MCMC approach that attempts to close the gap between efficiency and accuracy. By sampling using a surrogate likelihood, we accelerate convergence to the stationary distribution before sampling using the exact likelihood. We demonstrate that this method combines the efficiency of gradient matching and the accuracy of the exact likelihood scheme.

1 Introduction

Nonlinear ordinary differential equations (ODEs) are used in all branches of science to model the evolution and interaction between different dependent variables over some independent variable. Their use ranges from ecology [1] to biophysics [2, 3] and systems biology [4]. Although these systems are so commonly implemented, there is still not a universally accepted method of inferring their parameters, limiting our understanding of the associated systems and our ability to make predictions. Given the high level of uncertainty involved in the early phases of these processes, it makes sense that we propagate this uncertainty through to the parameter estimates. A Bayesian approach allows us to naturally account for the uncertainty in a posterior distribution.

Traditionally, ODE parameter inference methods ignored efficiency and any MCMC parameter inference approaches aimed only to accurately infer parameters of the ODEs [5]. With no upper bound on the run time, ergodicity of MCMC guarantees convergence to the correct posterior distribution. Recently, surrogate methods based on gradient matching have been proposed that allow us to circumvent the numerical integration step by finding a smooth interpolant to the noisy data. This then allows computation of a gradient of our estimated signal, which can be matched with that obtained from the ODEs. Instead of quantifying how well the solution of the ODEs matches the data, we quantify how well the derivatives predicted by the ODEs match the derivatives obtained from an interpolant to the data [6, 7, 8, 9]. Problematically, our improvement in efficiency comes at the expense of accuracy which is limited by the simple fact that we no longer adopt the exact likelihood for the ODEs.

The method proposed in this paper aims to close the gap between accuracy and efficiency, where we hope to sample based on the exact likelihood function after using the surrogate gradient matching likelihood function for accelerated convergence. We begin by giving an outline of the general ODE inference problem in Section 2 as well as the properties of these systems that lead to a more difficult parameter learning process. This provides motivation for the use of Gaussian processes in parameter inference as outlined in Section 3, leading to the introduction of a multiphase sampling technique that will be tested on four different benchmark ODE systems.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. JMLR: W&CP volume 7X.

2 The ODE Inference Problem

2.1 Standard MCMC Parameter Inference

In the standard parameter inference problem, we have observed some noisy data $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$ over a specific timeframe where $\mathbf{x}$ corresponds to the true signal governed by the ODEs:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{X}, \boldsymbol{\theta}, t) \tag{1}$$

Assuming additive iid Gaussian noise, we have the negative log-likelihood given by:

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}, \sigma) = \frac{mn}{2}\ln\sigma^2 + \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{(y_j(t_i) - x_j(\boldsymbol{\theta}, t_i))^2}{2\sigma^2}$$

where m is the number of variables and n is the number of time points. The constant term $\frac{n}{2}\ln 2\pi$ has been suppressed for ease of representation. The lack of closed-form solutions in nonlinear ODEs requires the application of numerical integration methods, like the Runge-Kutta method. Considering Bayes theorem, intractability in the denominator necessitates the use of MCMC sampling techniques that allow us to sample from the distribution $p(\boldsymbol{\theta}|\mathbf{y})$ without evaluating a marginal distribution. To the detriment of this approach is the repeated numerical integration of the ODEs, which is a computationally onerous procedure.

2.2 Difficulties in MCMC Inference

Although underpinned by asymptotic theory of ergodicity, MCMC sampling applied to the exact likelihood function faces substantial computational costs, due to the need for a numerical integration in every step of the iteration. This is aggravated by potential multimodality and entrapment in local optima. Resultantly, the length of time required to achieve convergence to the stationary distribution makes these methods practically not viable. In addition, problems such as stiffness of ODE systems compound this inefficiency, making the inference problem intractable in any realistic timeframe except for simple toy problems.

The solution to our problem is to find an alternative method of sampling from the posterior distribution that minimizes the number of numerical integrations. The most immediate leap in this direction would be consideration of procedures for estimating the underlying signal of the ODEs from the noisy observations.

3 Using Gaussian Processes to Reformulate the Inference Problem

3.1 Gaussian Process Smoothing

Assuming a zero mean GP prior for the latent signal vector $\mathbf{x}$, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ with entries of $\mathbf{K}$ given by the evaluation of some predefined kernel over the space $\mathcal{T} \times \mathcal{T}$, the Gaussian likelihood and prior enable analytic derivation of the GP posterior distribution:

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}, \sigma) \propto p(\mathbf{y}|\mathbf{x}, \sigma)p(\mathbf{x}|\boldsymbol{\phi}) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}) \tag{2}$$

where $\mathbf{x}$ is our latent signal vector, $\boldsymbol{\phi}$ is the vector of GP hyperparameters and σ is the observational noise standard deviation. Using Gaussian and matrix identities, we obtain:

$$\mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}) \propto \mathcal{N}(\mathbf{x}|\mathbf{m}, \Sigma) \tag{3}$$

with (this time making time dependence explicit):

$$\mathbf{m}(t) = \mathbf{k}(t, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \quad \boldsymbol{\Sigma} = \sigma^2\mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1} \tag{4}$$

where $\mathbf{k}(t, \mathcal{T})$ is a GP kernel function evaluated at time point t over the entire training set ($\mathcal{T}$). The posterior mean expression is a linear transformation of the observations (a linear smoother [10]) averaging over the noise to smooth the observed data.

It appears that the Gaussian process smoothing approach could help us bypass the expensive numerical integrations of the traditional method. However, choice of a suitable metric for assesment of different points in parameter space is non-trivial. Assuming differentiability of the kernel function, differentiability of GPs [11] may be exploited in a reformulation of the inference problem from a matching of signals with noisy observations to a matching of the interpolants' gradients with the gradients from the ODEs [12, 6, 13].

3.2 Gradient Matching with Random Interpolants

Implementing an alternative method, such as Gaussian process smoothing [11], allows us to neglect the numerically expensive Runge-Kutta methods. The method of Calderhead et al [6] treats the hyperparameters as random and samples these in each MCMC step. Using a distribution from the GP and the ODEs, the method relies on a product of experts approach to make the following integral tractable:

$$p(\boldsymbol{\theta}, \gamma|\mathbf{X}, \boldsymbol{\phi}) = \int p(\dot{\mathbf{X}}, \boldsymbol{\theta}, \gamma|\mathbf{X}, \boldsymbol{\phi})d\dot{\mathbf{X}} \tag{5}$$

where $\mathbf{X}$ has entries x_{ij} (the smoothed signal of the ith species at time j), $\boldsymbol{\theta}$ is the vector of ODE parameters and γ is a gradient mismatch hyperparameter to be defined in Section 3.3. Wang and Barber [14] observed the lack of motivation for this marginalization as it encourages a matching of modes rather than distributions. Dondelinger et al [7] alleviate this problem by introducing a regularizer—making the interpolant consistent with the ODEs—and sampling the parameters

from a proper posterior distribution. See Macdonald and Husmeier [13] for an overview.

The common problem with these random hyperparameter approaches is the reliance on innefficient Gibbs sampling in combination with GP sampling requiring $\mathcal{O}(T^3)$ computations to sample from a surrogate distribution that introduces bias to our posterior samples. We propose an extension of the gradient matching paradigm to exact likelihood MCMC methods.

3.3 Fixed Interpolant Gradient Matching

Taking the mean of this posterior distribution as defined in eq. (4) as a fixed interpolant of the data in a gradient matching scheme would fail to account for the same level of uncertainty as the methods discussed in Section 3.2. However, rather than accounting for uncertainty we aim for fast convergence to a stationary distribution close to the true posterior. It is in this true likelihood space that we wish to account for our uncertainty about the parameters of the ODEs.

The paradigm of fixed interpolants results in a deviation from the method of Calderhead et al [6] who treat the GP and ODE gradients as random variables. The proposed gradient matching scheme fixes the GP hyperparameters and interpolant based on maximum likelihood estimates. This is computationally cheaper than sampling from the posterior, as in [6, 7], allowing the computational resources to be used for a longer burn-in phase. We assume the derivative of the interpolant $\tilde{\mathbf{x}}$ to be determined by the ODEs subject to a Gaussian-distributed mismatch error with variance γ^2, $\frac{d\tilde{\mathbf{x}}}{dt} \sim \mathcal{N}(f(\tilde{\mathbf{x}}, \boldsymbol{\theta}, t), \gamma^2)$, and we represent this distribution by its conditional mean:

$$\dot{\hat{\mathbf{x}}} = \frac{d\hat{\mathbf{x}}}{dt} = \mathbf{k}'(t, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{6}$$

where $\mathbf{k}'(t, \mathcal{T}) = \frac{\partial \mathbf{k}(t, \mathcal{T})}{\partial t}$. The resultant likelihood function for the MCMC is the negative log-likelihood of a multivariate Gaussian distribution:

$$-\ln \hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}, \gamma) = n \ln \gamma^2 + \frac{1}{2\gamma^2}\left|\left|\frac{d\hat{\mathbf{x}}(t)}{dt} - f(\hat{\mathbf{x}}(t), \boldsymbol{\theta})\right|\right|^2 + C.$$

where $||\cdot||$ corresponds to the Euclidean norm. Introduction of a prior distribution over the parameters, $\boldsymbol{\theta}$ and the mismatch term, γ, will provide the posterior distribution of our MCMC sampler, $p(\boldsymbol{\theta}, \gamma|\hat{\mathbf{X}}, \dot{\hat{\mathbf{X}}})$. For the purposes of this work, the mismatch parameter γ was assumed constant over time and across different states and is sampled from the posterior using MCMC. However, the likelihood function is easily adjusted in order to account for varying γ across these variables.

Inherent in a fixed interpolant gradient matching scheme is an obvious reduction in computational complexity of each MCMC step since we no longer require numerical integrations of the ODEs. However, as a beneficial side-effect, the gradient matching surrogate likelihood also leads to a smoother likelihood surface, as shown in Figure 1 where we observe the ability to smooth over the local optima of the exact likelihood surface, similarly to an annealing approach [15]. Thus, the benefit of the fixed interpolant gradient matching surrogate likelihood is twofold since it reduces computational complexity of each step and avoids entrapment in local optima of the log likelihood surface.

3.4 Multiphase MCMC Sampling

Used alone, the fixed interpolant likelihood function will not permit sufficient coverage of the uncertainty in our parameter values, and will cause bias in the parameter estimates. As such, we propose the use of the fixed interpolant gradient matching surrogate likelihood in a burnin phase of a three-phase scheme. Below, we detail the multiple phases of the algorithm, where a corrective phase aims to correct the bias introduced in the surrogate burnin phase.

Surrogate burnin phase We begin with a surrogate burnin phase initiated from some initial point in parameter space. This phase, corresponding to lines 2-5 of Algorithm 1, is continued until some target potential scale reduction factor (PSRF) (see Chapter 8 of the book by Gilks et al. [16]) is achieved or until some maximum number of MCMC steps have been performed. The numerical efficiency in the gradient matching phase enables fast convergence to a region close to the true stationary distribution, avoiding local optima of the true likelihood function as illustrated in Figure 1.

Corrective phase Beginning from the last point sampled in surrogate space, fix the noise variance parameter and sample the ODE parameters for $N_{precorr}$ steps, initiated at some estimate obtained during the smoothing process. Effectively, we aim to estimate reasonable signals for the time-varying state variables at this stage, avoiding a sub-optimal attractor state that corresponds to a large variance and only learning the global mean of the data. Beginning from the final point of this precorrective phase, sample using the exact likelihood until a target PSRF value $PSRF_2$ is achieved or a preselected number of corrective steps N_{corr}, have been performed. This can be found in lines 6-13 of Algorithm 1.

Sampling phase The MCMC steps performed until now have been used to drive us towards the correct stationary distribution. We now initiate a sampling phase at the last sampled point from the corrective phase. The sampling phase is continued until a target $PSRF_3$ is reached or some number of steps, N_{samp},

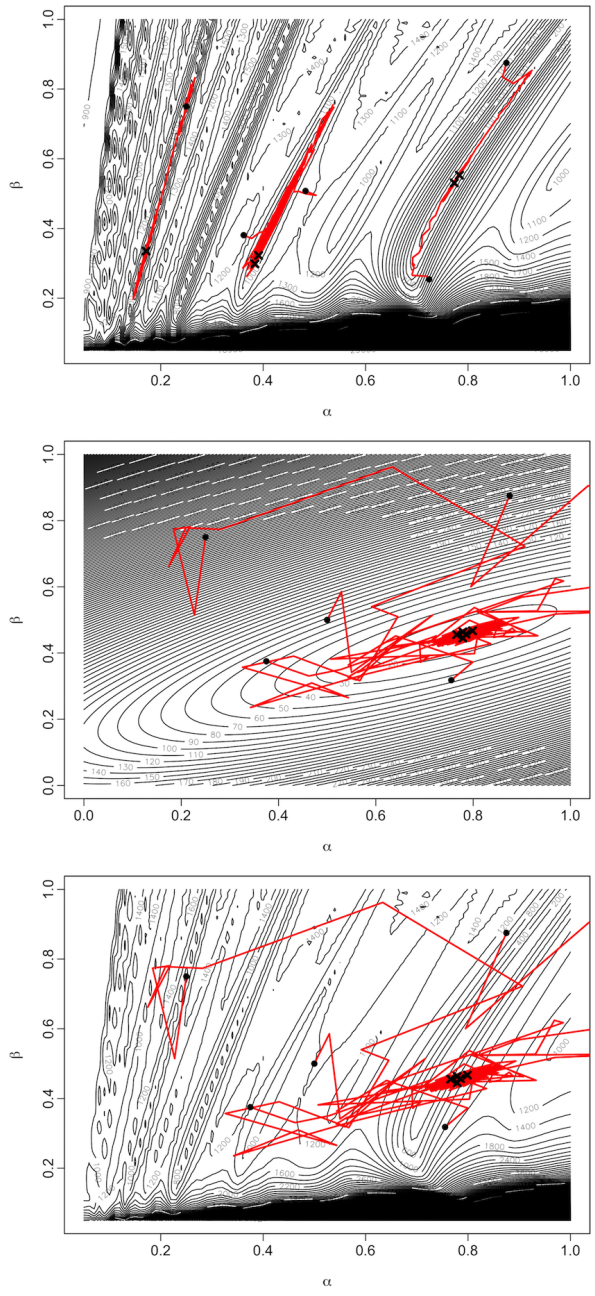

Figure 1: Top: the two-dimensional true likelihood surface of the Lotka-Volterra system. MCMC trajectories obtained with DRAM get trapped in local optima. Middle: unimodal gradient matching surrogate likelihood surface of the Lotka-Volterra system with gradient matching MCMC chains superimposed. Bottom: superimposing the gradient matching MCMC chains onto the true likelihood surface shows the acceleration of convergence to the stationary distribution as we smooth over local optima from the exact likelihood surface. Points denote start points and crosses indicate end points of the MCMC trajectories.

have been completed. This phase, given by lines 14-16 in Algorithm 1, provides samples from the posterior distribution for the parameters of the ODEs.

Algorithm 1 provides pseudocode for the multi phase scheme where the following nomenclature is adopted:

$$A_1(\boldsymbol{\theta},\gamma;\boldsymbol{\theta}',\gamma') = 1 \wedge \frac{\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}',\gamma')\pi(\boldsymbol{\theta}')\pi(\gamma')\hat{q}(\boldsymbol{\theta},\gamma|\boldsymbol{\theta}',\gamma')}{\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta},\gamma)\pi(\boldsymbol{\theta})\pi(\gamma)\hat{q}(\boldsymbol{\theta}',\gamma'|\boldsymbol{\theta},\gamma)}$$

$$A_2(\boldsymbol{\theta},\sigma;\boldsymbol{\theta}',\sigma') = 1 \wedge \frac{p(\mathbf{y}|\boldsymbol{\theta}',\sigma')\pi(\boldsymbol{\theta}')\pi(\sigma^{2\prime})q(\boldsymbol{\theta},\sigma|\boldsymbol{\theta}',\sigma')}{p(\mathbf{y}|\boldsymbol{\theta},\sigma)\pi(\boldsymbol{\theta})\pi(\sigma^2)q(\boldsymbol{\theta}',\sigma'|\boldsymbol{\theta},\sigma)}$$

where $p, \hat{p}$ are the exact and surrogate likelihoods, π is a prior and $q, \hat{q}$ are Gaussian proposal distributions with adaptive covariance.

Algorithm 1 Multi-phase MCMC Sampling

1: Assign initial parameters $\boldsymbol{\theta}_0$ and select cheap surrogate likelihood function $\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta},\gamma)$, an approximation of $p(\mathbf{y}|\boldsymbol{\theta},\sigma)$ (our computationally expensive likelihood function).
2: **repeat**
3: Sample $\boldsymbol{\theta}_t$ and γ_t from $\hat{q}(\boldsymbol{\theta}_t,\gamma_t|\boldsymbol{\theta}_{t-1},\gamma_{t-1})$.
4: Accept proposed point based on $A_1(\boldsymbol{\theta},\gamma;\boldsymbol{\theta}',\gamma')$
5: **until** PSRF_1 or N_{surr} reached
6: **repeat**
7: Beginning from the last point sampled in surrogate space, keep σ^2 fixed at some estimate and sample $\boldsymbol{\theta}_t$ from $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.
8: Accept proposed point based on $A_2(\boldsymbol{\theta},\sigma;\boldsymbol{\theta}',\sigma)$
9: **until** Until N_{pre} reached
10: **repeat**
11: Beginning from the last point sampled in the pre-corrective phase, sample $\boldsymbol{\theta}_t$ and σ_t from $q(\boldsymbol{\theta}_t,\sigma_t|\boldsymbol{\theta}_{t-1},\sigma_{t-1})$.
12: Accept proposed point based on $A_2(\boldsymbol{\theta},\sigma;\boldsymbol{\theta}',\sigma')$
13: **until** PSRF_2 or N_{corr} reached
14: **repeat**
15: Beginning from the last point sampled in corrective phase, sample $\boldsymbol{\theta}_t$ and σ_t from some proposal distribution $q(\boldsymbol{\theta}_t,\sigma_t|\boldsymbol{\theta}_{t-1},\sigma_{t-1})$ (this will differ from those of the previous two phases).
16: Accept proposed point based on $A_2(\boldsymbol{\theta},\sigma;\boldsymbol{\theta}',\sigma')$
17: **until** PSRF_3 or N_{samp} reached

3.5 Delayed Acceptance Metropolis-Hastings

One of the implementations of the gradient matching surrogate likelihood will be in a delayed acceptance Metropolis-Hastings (DAMH) sampling scheme. The general technique was introduced by Sherlock et al. [17]. The idea is to propose new moves in the surrogate space prior to their proposal in the exact likelihood space. This saves computation time by limiting the number of exact likelihood evaluations for rejected

points since, assuming similarity between the two likelihoods, proposals are filtered out by rejection based on the initial acceptance criterion; see eq.7. The method still depends on acceptance via evaluation of the exact likelihood, but the ability to reject based only on a cheap proxy reduces computational complexity.

We implemented DAMH with a gradient matching likelihood as the cheap initial likelihood. This will be used for sampling in a multiphase scheme with a surrogate burnin phase (use Algorithm 1 with steps 10 to 17 replaced by DAMH with gradient matching likelihood function). The choice of acceptance criterion is based on the property of detailed balance where we may take a product of surrogate and true likelihood acceptance expressions $\alpha(x,y) = \alpha_1(x,y)\alpha_2(x,y)$ such that:

$$\alpha_1(x,y) = 1 \wedge \frac{\hat{p}(y)q(x|y)}{\hat{p}(x)q(y|x)}, \alpha_2(x,y) = 1 \wedge \frac{p(y)/\hat{p}(y)}{p(x)/\hat{p}(x)} \quad (7)$$

where $x \wedge y = \min(x,y)$. Banterle et al. [18] show that the generated Markov chain is reversible with respect to p and so, assuming ergodicity, p is the limiting distribution of the Markov chain. An inefficiency is introduced by the fact that $\alpha_{DA}(x,y) \leq \alpha_{MH}(x,y)$ where $\alpha_{MH}(x,y)$ is the acceptance criterion in Metropolis-Hastings and $\alpha_{DA}(x,y)$ the acceptance criterion of delayed acceptance. However, this inefficiency by effective sample size will hopefully be overcome by computational efficiency that will enable more samples from the posterior distribution. To alleviate this problem, we take at each iteration a standard MH step on the expensive likelihood with probability β.

4 Benchmark ODE Systems

This section presents the four systems of ODEs that will be used for comparison of the proposed method with 4 other benchmark MCMC sampling techniques.

Lotka-Volterra (LV) The equations model the interaction between predator (y) and prey (x) variables over time. This system exhibits periodic population evolution resulting from the interactions between the two variables of the system. Parameter values used were $\alpha = 0.76, \beta = 0.5, \gamma = 0.4$ and $\delta = 0.3$. 100 points were simulated between times 0 and 100.

$$\frac{dx}{dt} = \alpha x - \beta xy, \quad \frac{dy}{dt} = -\gamma y + \delta xy \quad (8)$$

FitzHugh-Nagumo (FHN) The ODEs model [2, 3] the movement of signals along excited cells. V denotes the voltage of the signal, which accounts for self-excitation of the membrane, and R is a recovery variable acting as a negative feedback mechanism.

$$\frac{dV}{dt} = \gamma\left(V - \frac{V^3}{3} + R\right), \frac{dR}{dt} = -\frac{V - \alpha + \beta R}{\gamma} \quad (9)$$

Motivated by the literature [19], parameter values used were $\alpha = 0.2, \beta = 0.2$ and $\gamma = 3$. The observed dataset consists of 100 points measured at equal intervals between times 0 and 20.

Goodwin Oscillator (GO) The ODEs, first introduced by Goodwin [4], models the concentration of an enzymatic protein and messenger ribonucleic acid (mRNA) of some species. Production of mRNA p_1 is met by a binding of mRNA with ribosomes, forming a enzymatic protein p_2 with feedback, inhibiting mRNA transcription. Constant decay terms, although biologically inconsistent, encourage periodicity in the signals of the species, a trait that is more difficult to induce under a more accurate system configuration [20].

$$\frac{dp_1}{dt} = \frac{k_1}{36 + k_2 p_2} - k_3, \quad \frac{dp_2}{dt} = k_4 p_1 - k_5 \quad (10)$$

Parameters used were motivated by the literature [21], taking $k_1 = 72, k_2 = 1, k_3 = 2, k_4 = 1, k_5 = 1$. For observations, we simulated 120 data points between times 0 and 60 at equally spaced time intervals. The various local optima of the resultant likelihood surface in two dimensions, as shown in Figure 3 of the supplementary material (SM), lead to a more challenging parameter learning problem.

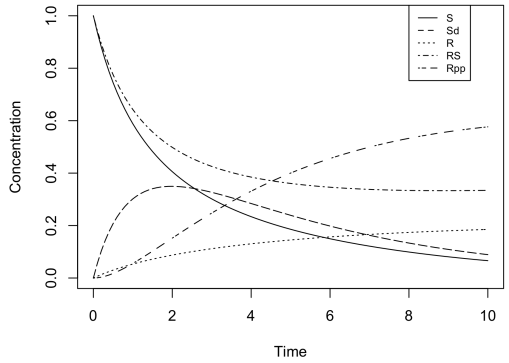

Figure 2: Pre-equilibrium observed data from the STC model. The data measured between times 0 and 100 is given in Figure 1 of the SM where we observe the system entering into equilibrium.

Signal Transduction Cascade (STC) The final system models the signal transduction cascade with input enzymatic protein S. This binds to protein R, forming protein complex RS, inducing phosphorylation of protein R into R_{pp} from which state the protein may be deactivated. Additionally, the system below also describes the degradation of the input S into Sd.

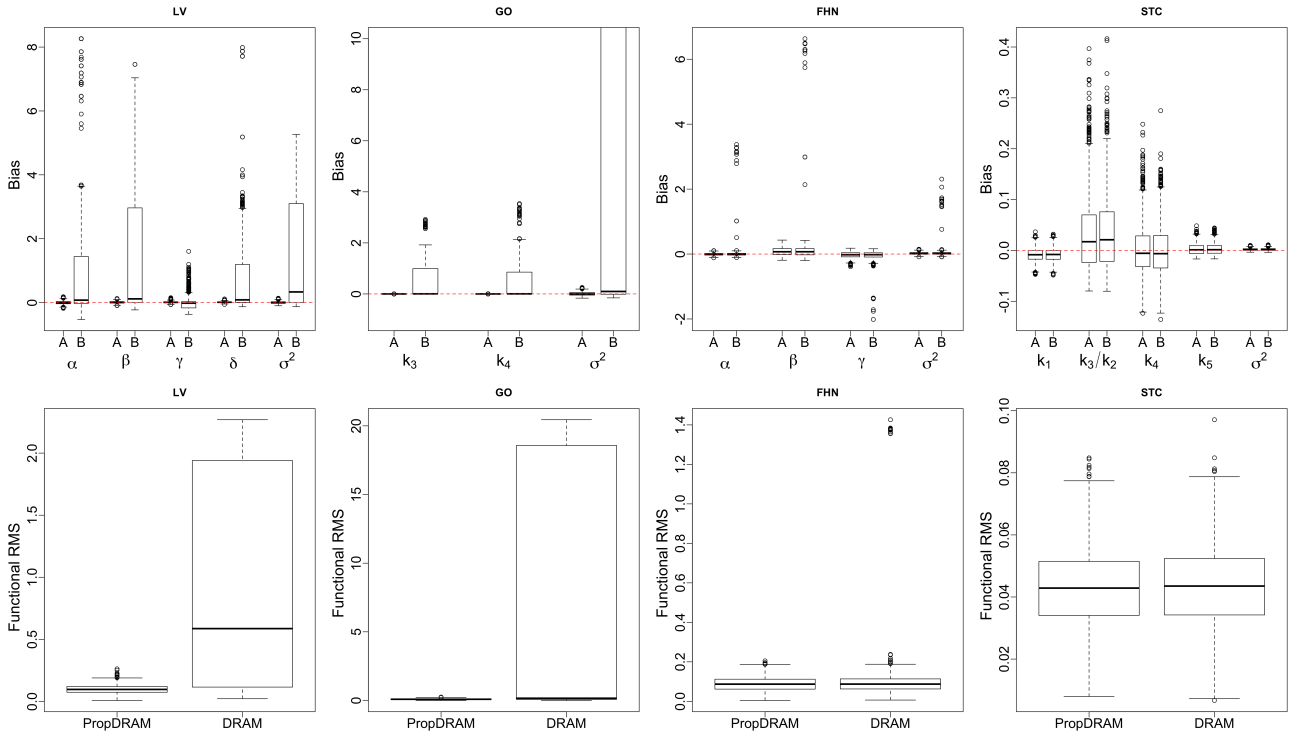

Figure 3: Top: The boxplots show the inferred parameter distributions obtained by combining the posterior distributions from 10 independent data instantiations using propDRAM (A) and traditional DRAM with the exact likelihood function (B) for each of the 4 ODE models. We observe the advantage of the surrogate burnin as this allows us to evade various local optima; compare with Figure 1. Bottom: Boxplots showing functional RMS of posterior samples for each of the four models using the three phase proposed scheme. The proposed method outperforms DRAM on all benchmark data (indicated by the abbreviation on top of the panels).

Variables inside [] correspond to the concentrations of the different species.

$$
\begin{aligned}
\frac{d[S]}{dt} &= -k_1[S] - k_2[S][R] + k_3[RS] \\
\frac{d[S_d]}{dt} &= k_1[S] \\
\frac{d[R]}{dt} &= -k_2[S][R] + k_3[RS] + \frac{V[R_{pp}]}{K_m + [R_{pp}]} \\
\frac{d[RS]}{dt} &= k_2[S][R] - k_3[RS] - k_4[RS] \\
\frac{d[R_{pp}]}{dt} &= k_4[RS] - \frac{V[R_{pp}]}{K_m + [R_{pp}]}
\end{aligned}
\tag{11}
$$

Our parameter value choice is as chosen in the literature [7] with parameter values as follows: $k_1 = 0.07, k_2 = 0.6, k_3 = 0.05, k_4 = 0.3, V = 0.017, K_m = 0.3$. Due to practical nonidentifiability issues, the noisy data consists of 20 data points between times 0 and 10. This corresponds to the period prior to the data reaching equilibrium, as displayed in Figure 2.

5 Results

5.1 Comparison of Methods

We present parameter inference results using four alternative methods on each of the ODE systems above. For compactness, the following nomenclature is adopted: **PropDRAM** is our proposed scheme from Algorithm 1 with gradient matching surrogate likelihood and sampling using DRAM [22] with the exact likelihood function. **PopMCMC** uses the exact likelihood function in population MCMC [23], a competitive MCMC method for inference in multimodal likelihood spaces. **DRAM** corresponds to Delayed Rejection Adaptive Metropolis with the exact likelihood function [22]. **PropDAMH** is our proposed scheme with surrogate burnin phase and DAMH sampling phase (see Section 3.5) with gradient matching filter step in the first phase of eq. 7.

Considering the setup for sampling using the proposed scheme, we refer to Algorithm 1 with the following values: $\text{PSRF}_1 = 1.1, N_{surr} = 10000, N_{pre} = 200, \text{PSRF}_2 = 1.05, N_{corr} = 10000, N_{samp} =$

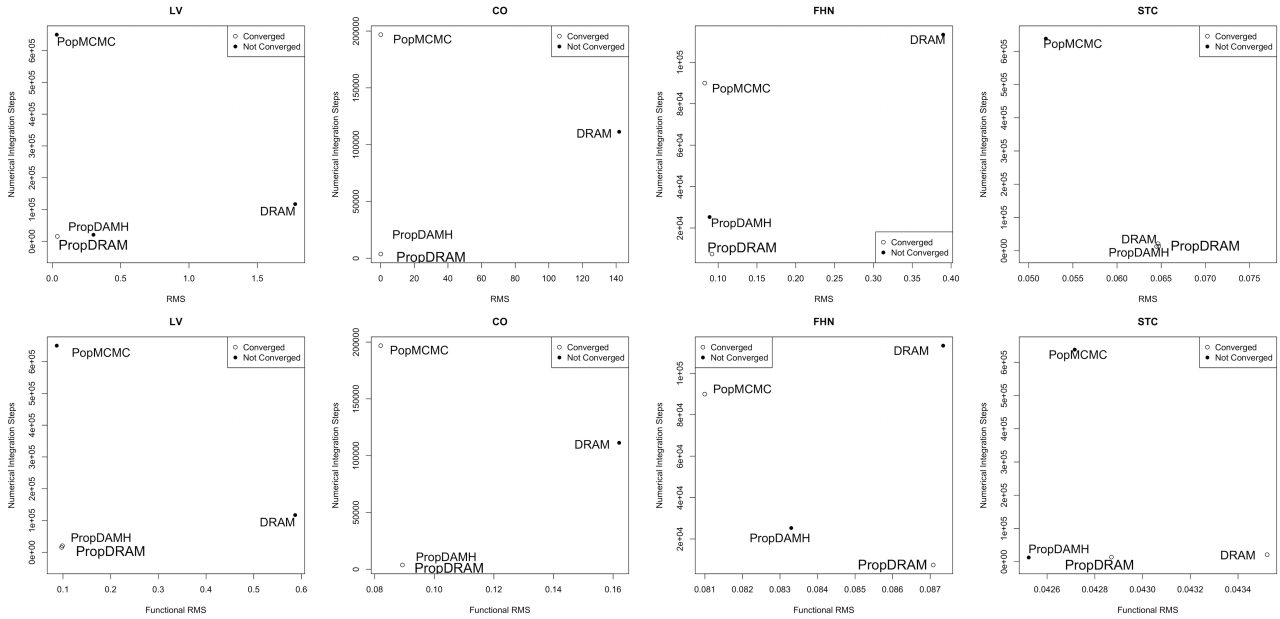

Figure 4: Top: The figure shows the RMS scores obtained by combining the posterior distributions from 10 independent data instantiations versus number of numerical integration steps. This allows consideration of accuracy (horizontal axis) relative to computational complexity (vertical axis) for each of the four methods. Good performance is signified by a method appearing in the bottom left corner (with the exception of the STC data since RMS differences are so small). The multi phase proposed scheme (propDRAM) is the only method consistently appearing in the bottom left corner. Bottom: Plots of number of numerical integration steps (vertical axis) versus functional RMS (horizontal axis).

$5000, \text{PSRF}_3 = 1.01$. The benchmark methods had stopping rule PSRF=1.01 or 50000 MCMC steps. In all cases, initial conditions were assumed known, providing optimal conditions for DRAM and popMCMC since these no longer have to infer the initial conditions and parameters simultaneously. For each ODE system, we simulated ten datasets and perform parameter inference for each.

Figure 3 allows consideration of the accuracy of the proposed method compared with traditional DRAM MCMC with the exact likelihood. In the first 3 ODE models, periodicity in the signal means the proposed propDRAM outperforms DRAM. For the STC data, a lack of identifiability leads us to infer the ratio k_3/k_2 instead of each parameter individually. The lack of periodicity in the signal results in a smoother likelihood surface and so accuracy of the two methods is very similar. However, Figure 4[1] shows the reduction in computational expense when using the proposed scheme.

Multimodality of likelihood surfaces (see Figure 1 and Section 1.2 of the SM) makes it important for us to consider algorithms designed for evading these local optima (detailed bias boxplots for each of the methods across the four models are given in Section 1.3 of the SM)[2]. We aim for efficiency and accuracy in the inference scheme, leading to consideration of accuracy relative to computational complexity. Figure 4 presents root mean square deviation (RMS) in parameter space versus number of numerical integrations of the ODEs. Appearance in the bottom left corner of the plot indicates good performance of the method. With the exception of the signal transduction cascade, PropDRAM is the only method consistently displaying this trait. However, for the signal transduction cascade, RMS difference between the right and left corners of the plot is relatively small (0.012) in comparison to the average parameter value of 0.2074 (6%) and popMCMC incurs far greater computational cost. A legend is provided for indication of convergence in each of the different models. PropDRAM was the only method to converge in all cases. Considering function space performance, we evaluated the functional RMS value based on posterior samples $RMS_{func} = \sqrt{\frac{1}{n}\sum_i |\mathbf{x} - \hat{\mathbf{x}}_i|^2}$, where $\mathbf{x}$ is the true signal and $\hat{\mathbf{x}}_i$ is the signal at the ith posterior sample

[1] We used $\beta = 0.2$, as suggested by one of the reviewers. Our orginal results with $\beta = 0$ can be found in the SM.

[2] In all cases, we use uninformative gamma priors for the ODE parameters to satisfy their positivity constraints and conjugate inverse-gamma priors for the noise variance parameters.

obtained by solving the ODE at our samples. The bottom row of Figure 4 shows number of numerical integrations versus Functional RMS. As before, appearance in the bottom left means good performance. PropDRAM is the only method to consistently achieve this.

6 Discussion

This paper has presented, in detail, two alternative implementations of gradient matching. One of them is based on delayed acceptance [17] (propDAMH). Considering eq. 7, it becomes apparent that certain properties of the surrogate distribution lead to inefficient sampling from the posterior. In cases where the surrogate has heavier tails than the exact likelihood, $\alpha_1(x, y)$ is high and so points pass the filter step and require a decision be made based on the exact likelihood with a high rejection rate. More explicitly, use of the DAMH algorithm moves towards an inefficient, low acceptance rate, Metropolis-Hastings algorithm. The situation with weak tails manifests itself differently to the scenario above. $\alpha_1(x, y)$ will be low, resulting in a low proportion of points making it through the initial filter step, giving a poor representation of the exact posterior distribution as we fail to sample from the tails of the distribution.

The key result was the proposal of a multiphase approach to parameter inference that extends the paradigm of gradient matching to sampling from the correct posterior distribution. Using a gradient matching likelihood function, we efficiently sample from an approximation of the exact posterior distribution, driving us towards the posterior mode (as displayed in Figure 1). This minimizes the number of evaluations of the exact likelihood function required to converge to the stationary distribution. We compared the accuracy of the proposed method relative to the computational expense in Figure 4. Providing a more thorough exploration of the performance compared with traditional DRAM allowed us to observe the proposed scheme's ability to avoid local optima.

7 Conclusion

Parameter inference in systems of coupled ODEs is intrinsically computationally expensive due to the need for a numerical integration at every parameter adaptation. The methodological framework of gradient matching systematically bypasses this computational bottleneck by inferring gradients from a smooth interpolant. To reduce the bias inherent in this framework, past approaches have tried to use the ODEs as a regularizer, but this is computationally expensive and never eliminates the bias completely. The approach we have proposed here is to use a surrogate function based on gradient matching only in the burn-in phase of a multi-phase MCMC scheme, and use the correct likelihood obtained from a (computationally expensive) numerical integration of the ODEs as a final adjustment. Our simulation studies, based on four standard benchmark ODE systems, suggest that the novel approach achieves similar accuracy as the standard scheme that is entirely based on the true likelihood, but at substantially reduced computational costs. We have obtained insight into the source of the improvement from log likelihood contour plots, like Figure 1. The surrogate log likelihood tends to have its probability mass close to the mode of the true log likelihood, with a much smoother surface akin to what could have been obtained (at higher computational costs) with an annealing scheme. Consequently, the burn-in phase achieves two objectives simultaneously: an increase in the computational efficiency (due to bypassing the numerical integration), and an avoidance of entrapment in local optima (due to the smoother surface). These enable improved accuracy in the final posterior samples, as shown in Figures 3 and 4.

It appears natural to combine the proposed method with the delayed acceptance MCMC scheme recently proposed by Sherlock et al. [17], as the surrogate likelihood from gradient matching appears to provide the natural function for the first-stage rejection step in this algorithm. This has turned out to be counterproductive, though, as the mismatch between the two likelihood functions, which is illustrated in Fig. 1, leads to a systematic undersampling from the tails of the true posterior distribution, resulting in a net loss of computational efficiency. From this, we would argue that the critical dependence on a close match between the true and the surrogate likelihood is an intrinsic limitation of the delayed acceptance MCMC scheme per se, which does not become immediately clear from its mathematical formulation, and our study is therefore also of interest to current methodological research in MCMC beyond the particular settings of ODE parameter estimation.

8 Acknowledgements

This project was funded by EPSRC, project numbers EP/L020319/1 and EP/N014642/1.

References

[1] A. J. Lotka. The growth of mixed populations: Two species competing for a common food supply. *Journal of the Washington Academy of Sciences*, 22(16):110–120, 1932.

[2] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1, 1961.

[3] J S. Nagumo, S Arimoto, and S Yoshizawa. An active pulse transmission line simulating a nerve axon. In *Proceedings of the IRE*, 1962.

[4] B. C. Goodwin. Oscillatory behavior in enzymatic control processes. *Advances in Enzyme Regulation*, 3:425–438, November 1965.

[5] V. Vyshemirsky and M. A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008.

[6] B. Calderhead, M. A. Girolami, and N. D. Lawrence. ODE parameter inference using adaptive gradient matching with Gaussian processes. 2008.

[7] F. Dondelinger, D. Husmeier, S. Rogers, and M. Filippone. ODE parameter inference using adaptive gradient matching with Gaussian processes. In *Proceedings of Machine Learning Research*, volume 31, pages 216–228. 2013.

[8] Mu Niu, Simon Rogers, Maurizio Filippone, and Dirk Husmeier. Fast parameter inference in nonlinear dynamical systems using iterative gradient matching. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1699–1707, New York, USA, 20–22 Jun 2016. PMLR.

[9] X. Xun, J. Cao, B. Mallick, A. Maity, and R. J. Carroll. Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, 108(503), 2013.

[10] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[11] C. E. Rasmussen and K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA., 2006.

[12] J. M. Varah. A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.

[13] B. Macdonald and D. Husmeier. Gradient matching methods for computational inference in mechanistic models for systems biology: A review and comparative analysis. *Frontiers in Bioengineering and Biotechnology*, 3, November 2015.

[14] Y. Wang and D. Barber. Gaussian processes for bayesian estimation in ordinary differential equations. *Proceeding of Machine Learning Research ICML 2014*, 32(2):1485–1493, 2014.

[15] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[16] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[17] C. Sherlock, A. Golightly, and D. A. Henderson. Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *Journal of Computational and Graphical Statistics*, 26(2):434–444, 2017.

[18] M. Banterle, C. Grazian, A. Lee, and C. P. Robert. Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. *ArXiv e-prints*, March 2015.

[19] D. Campbell and R. J. Steele. Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, 22(2):429–443, November 2012.

[20] A. Woller, D. Gonze, and T. Erneux. Strong feedback limit of the Goodwin circadian oscillator. *Physical Review*, 87, March 2013.

[21] M. Girolami, B. Calderhead, and V. Vyshemirsky. System Identification and Model Ranking: The Bayesian Perspective. In N. D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti, editors, *Learning and Inference in Computational Systems Biology*, chapter 8, pages 201–230. MIT Press, Cambridge, MA, 2010.

[22] H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

[23] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.