Efficient and Flexible Inference for Stochastic Systems

Authored by:

Joachim M. Buhmann
Nico S. Gorbach
Stefan Bauer
Djordje Miladinovic

Abstract

Many real world dynamical systems are described by stochastic differential equations. Thus parameter inference is a challenging and important problem in many disciplines. We provide a grid free and flexible algorithm offering parameter and state inference for stochastic systems and compare our approch based on variational approximations to state of the art methods showing significant advantages both in runtime and accuracy.

1 Paper Body

A dynamical system is represented by a set of K stochastic differential equations (SDE?s) with model parameters ? that describe the evolution of K states X(t) = [x1 (t), x2 (t), . . . , xK (t)]T such that: dX(t) = f (X(t), ?)dt + ?dWt , (1)

where Wt is a Wiener process. A sequence of observations, y(t) is usually contaminated by some measurement error which we assume to be normally distributed with zero mean and variance for each of the K states, i.e. E ? N (0, D), with Dik = ?k2 ?ik . Thus for N distinct time points the overall system may be summarized as Y = AX + E, where X = [x(t1), . . . , x(tN)] = [x1 , . . . , xK]T Y = [y(t1), . . . , y(tN)] = [y1 , . . . , yK]T , where xk = [xk (t1), . . . , xk (tN)]T is the k?th state sequence and yk = [yk (t1), . . . , yk (tN)]T are the observations. Given the observations Y and the description of the dynamical system (1), the aim is to estimate both state variables X and parameters ?. Related Work. Classic approaches for solving the inverse problem i.e. estimating the parameters given some noisy observations of the process, include the Kalman Filter or its improvements [e.g. Evensen, 2003, Torn?e et al., 2005] and MCMC based approaches [e.g. Lyons et al., 2012]. However, ?

The first two authors contributed equally to this work.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

MCMC based methods do not scale well since the number of particles required for a given accuracy grows exponentially with the dimensionality of the inference problem [Snyder et al., 2008], which is why approximations to the inference problem became increasingly more popular in recent years. Archambeau et al. [2008] proposed a variational formulation for parameter and state inference of stochastic diffuion processes using a linear dynamic approximation: In an iterated two-step approach the mean and covariance of the approximate process (forward propagation) and in the second step the time evolution of the Lagrange multipliers, which ensure the consistency constraints for mean and variance (backward propagation), are calculated in order to obtain a smooth estimate of the states. Both forward and backward smoothing require the repeated solving of ODEs. In order to obtain a good accuracy a fine time grid is additionally needed, which makes the approach computational expensive and infeasible for larger systems [Vrettas et al., 2015]. For parameter estimation the smoothing algorithm is used in the inner loop of a conjugate gradient algorithm to obtain an estimate of the optimal approximation process (given a fixed set of parameters) while in the outer loop a gradient step is taken to improve the current estimate of the parameters. An extension of Archambeau et al. [2008] using local polynomial approximations and mean-field approximations was proposed in Vrettas et al. [2015]. Mean-field approximations remove the need of Lagrange multipliers and thus of the backward propagation while the polynomial approximations remove the need of solving ODEs iteratively in the forward propagation step which makes the smoothing algorithm and thus the inner loop for parameter estimation feasible, even for large systems while achieving a comparable accuracy [Vrettas et al., 2015]. Our contributions. While established methods often assume full observability of the stochastic system for parameter estimation, we solve the more difficult problem of inferring parameters in systems which include unobserved variables by combining state and parameter estimation in one step. Despite the fact that we compare our approach to other methods which solve a simpler problem, we offer improved accuracy in parameter estimation at a fraction of the computational cost.

2

Random Ordinary Differential Equations

Compared to stochastic differential equations, random ordinary differential equations (RODEs) have been less popular even though both frameworks are highly connected. RODEs are pathwise ordinary differential equations that contain a stochastic process in their vector field functions. In Kloeden and Jentzen [2007] RODEs have been studied to derive better numerical integration schemes for SDEs, which e.g. allows for stronger pathwise results compared to the L2 results given in Ito stochastic calculus. Moreover, RODEs sometimes have an advantage over SDEs by allowing more realistic noise for some applications e.g. correlated noise or noise with limited variance. Let (?, F, P) be a complete probability space, (?t)t?[0,T] be a Rm -valued stochastic process with continuous sample paths and f : Rm ? Rd ? Rd a continuous function. Then dx(t) =

f (x(t), ?t (?)) dt

(2)

is a scalar RODE, that is, an ODE dx(t) = F? (t, x) := f (x(t), ?(t)), (3) dt for all ? ? ?. Following Kloeden and Jentzen [2007], we likewise assume that f is arbitrary smooth i.e. f ? C ? and thus locally Lipschitz in x such that the initial value problem (3) has a unique solution, which we assume to exist on the finite time interval [0, T]. A simple example for a RODE is Example 1 (RODE). dx(t) = ?x + sin(Wt (?)), dt

(4)

where Wt is a Wiener process. Taylor-like schemes for directly solving RODEs (2) were derived e.g. in Gr?ne and Kloeden [2001], Jentzen and Kloeden [2009]. One approach for solving the RODE (2) is to use sampling to obtain many ODE?s (3) which can then be solved pathwise using deterministic calculus. However, this pathwise solution of RODEs implies that a massive amount of deterministic ODEs have to be solved efficiently. A study with a high performance focus was conducted in 2

Riesinger et al. [2016], where parallelized pathwise inference for RODEs was implemented using GPU?s. While in principle classic numerical schemes for deterministic systems e.g. Runge-Kutta can be used for each path, they will usually converge with a lower order since the vector field is not smooth enough in time [Asai et al., 2013]. Since the driving stochastic process ?t has at most H?lder continuous sample paths, the sample paths of the solution t ? x(t) are continuously differentiable but the derivatives of the solution sample paths are at most H?lder continuous in time. This is caused by the fact that F? (t, x) of the ODE (3) is usually only continuous, but not differentiable in t, no matter how smooth the function f is in its variables. RODEs offer the opportunity to use deterministic calculus (pathwise), yet being highly connected with an SDE since any RODE with a Wiener process can be written as SDE Jentzen and Kloeden [2011]. To illustrate the point, the example 1 above can be re-written as an SDE by: Example 2 (SDE transformed RODE).

Xt ?Xt + sin(Yt) 0 dWt . d = + Yt 0 1

(5)

It likewise holds that SDEs can be transformed into RODEs. This transformation was first described in Sussmann [1978] and Doss [1977] and generalized to all finite dimensional stochastic differential equations by Imkeller and Schmalfuss [2001]. RODEs can thus be used to find pathwise solutions for SDEs but SDEs can likewise be used to find better solution for RODEs Asai and Kloeden [2013]. Due to space limitations and to circumvent the introduction of a large mathematical framework, we only show the transformation for additive SDE?s following [Jentzen and Kloeden, 2011, chapter 2]. Proposition 1. Any finite dimensional SDE can be transformed into an RODE and the other way round: dxt = f (xt)dt + dWt ??

dz(t) = f (zt + Ot) + Ot , dt

(6)

where z(t) := xt ? Ot and Ot is the Ornstein-Uhlenbeck stochastic stationary process satisfying the linear SDE dOt = ?Ot dt + dWt

(7)

Typically a stationary Ornstein-Uhlenbeck process is used to replace the white noise of the SDE in its transformation to an RODE. By continuity and the Fundamental Theorem of Calculus it then follows that z(t) is pathwise differentiable. While we only showed the transformation for additive SDE?s, it generally holds true that any RODE with a Wiener process can be transformed into an SDE and any finite dimensional SDE with regular coefficients can be transformed into an RODE. This includes nonlinear drifts and diffusions and is true for univariate and multivariate processes [Han and Kloeden, 2017]. There are cases for which this does not hold e.g. a RODE which includes fractional Brownian motion as the driving noise. While the presented method is thus even more general since RODE?s can be solved, we limit ourselves to the problem of solving additive SDE?s by transforming them into a RODE. Since the solution of a RODE is continuously differentiable in time (but not further differentiable in time), classic numerical methods for ODEs rarely do achieve their traditional order and thus efficiency [Kloeden and Jentzen, 2007]. In the following we describe a scalable variational formulation to infer states and parameters of stochastic differential equations by providing an ensemble learning type algorithm for inferring the parameters of the corresponding random ordinary differential equation.

3

Variational Gradient Matching

Gradient matching with Gaussian processes was originally motivated in Calderhead et al. [2008] and offers a computationally efficient shortcut for parameter inference in deterministic systems. While the original formulation was based on sampling, Gorbach et al. [2017] proposed a variational formulation offering significant runtime and accuracy improvements. Gradient matching assumes that the covariance kernel C?k (with hyper-parameters ?k) of a Gaussian process prior on state variables is once differentiable to obtain a conditional distribution over state 3

Figure 1: Noise. The left plot shows three typical Wiener processes generated with mean zero and the corresponding Ornstein-Uhlenbeck (OU) process having the same Wiener process in its diffusion (right). The scale on the y-axis shows the mean-reverting behaviour of the OU process (compared to the Wiener process). derivatives using the closure property under differentiation of Gaussian processes: ? — X, ?) = p(X

Y

N (x? k — mk , Ak),

(8)

k

where the mean and covariance is given by: mk := 0 C?k C?1 ?k xk ,

0 Ak := C00?k ? 0 C?k C?1 ?k C?k ,

C00?k denotes the auto-covariance for each state-derivative with C0?k and 0 C?k denoting the crosscovariances between the state and its derivative. The posterior distribution over state-variables is p(X — Y, ?, ?) =

Y

N (?k (yk), ?k) ,

(9)

k

where ?k (yk) := C?k (C?k + ?k2 I)?1 yk and ?k := ?k2 C?k (C?k + ?k2 I)?1 . Inserting the GP based prior in the right hand side of a differential equation and assuming additive, normally distributed noise with state-specific error variance ?k one obtains a distribution of state derivatives ? — X, ?, ?) = p(X

Y

N (x? k — fk (X, ?), ?k I) .

(10)

k

which is combined with the smoothed distribution obtained from the data fit (9) in a product of experts approach: ? — X, ?, ?, ?) ? p(X ? — X, ?)p(X ? — X, ?, ?). p(X

After analytically integrating out the latent state-derivatives Y

p(? — X, ?, ?) ? p(?) N fk (X, ?) — mk , ??1 k) .

(11)

k

where ??1 k := Ak + ?k I one aims to determine the maximum a posteriori estimate (MAP) of the parameters Z ? ? : = arg max ln p(? — X, ?, ?)p(X — Y, ?)dX, (12) ?

Since the integral in (12) is in most cases analytically intractable (even for small systems due to the non-linearities and couplings induced by the drift function), a lower bound is established through the 4

introduction of an auxiliary distribution Q: Z ln p(? — X, ?, ?)p(X — Y, ?)dX R Z Q(X)dX (a) = ? Q(X)dX ln R p(? — X, ?, ?)p(X — Y, ?)dX Z (b) Q(X) ? ? Q(X) ln dX p(? — X, ?, ?)p(X — Y, ?) = H(Q) + EQ ln p(? — X, ?, ?) + EQ ln p(X — Y, ?) =: LQ (?) (13) R where H(Q) is the entropy. In (a) the auxiliary distribution Q(X), Q(X)dX = 1 is introduced and in (b) is using Jensens?s inequality. The lower bound holds with equality whenever p(? — X, ?, ?)p(X — Y, ?) (c) Q? (X) : = R = p(X — Y, ?, ?, ?), p(? — X, ?, ?)p(X — Y, ?)dX where in (c) Bayes rule is used. Unfortunately Q? is analytically intractable because its normalization given by the integral in the denominator is in most cases analytically intractable due to the strong couplings induced by the nonlinear drift function f in (1). Using mean-field approximations

Y Q := Q : Q(X, ?) = q(? — ?) q(xu — ? u) , (14) u

where ? and ? u are the variational parameters. Assuming that the drift in (1) is linear in the parameters ? and that states only appear as monomial factors in arbitrary large products of states the true conditionals p(? — X, Y, ?) and p(xu — ?, X?u , Y, ?) are Gaussian distributed, where X?u denotes all states excluding state xu (i.e. X?u := {x ? X — x 6= xu }) and thus q(? — ?) and q(xu — ? u) are designed to be Gaussian. b This Q Qposterior distribution over states is then approximated as p(X—Y, ?, ?, ?, ?) ? Q(X) = b?kt and the log transformed distribution over the ODE parameters given the observations as k tq ln p(?—Y, ?, ?, ?) ? LQ? (?).

Algorithm 1 Ensemble based parameter estimation for SDEs

1: Transform the SDE 1 into a RODE 2 2: Simulate a maximum number Nmax of OU-processes and insert them in 2 to obtain Nmax ODEs 3: For each ODE obtain approximate solutions using variational gradient matching [Gorbach et al.,

2017]

b to obtain an estimate of the parameters for the RODE 2 4: Combine the solutions ? 5: Transform the solutions of the RODE 2 back into solutions of the SDE 1.

Gorbach et al. [2017] then use an EM-type approach illustrated in figure 2 iteratively optimizing parameters and the variational lower bound LQ? (?). The variational parameters can be derived analytically and the algorithm scales linearly in the number states of the differential equation and is thus ideally suited to infer the solutions of the massive number of pathwise ODEs required for the pathwise solution of the RODE formulation of the SDE. Since solution paths of the RODE are only once differentiable, gradient matching (which only makes this assumption w.r.t. solution paths) is ideally suited for estimating the parameters. Our approach is summarized in algorithm 1. However, the application of variational gradient matching [Gorbach et al., 2017] for the pathwise solution of the RODE is not straightforward since e.g. in the case for scalar stochastic differential equations one has to solve dz(t) = f? (zt + Ot) + Ot , (15) dt for a sampled trajectory Ot of an Ornstein-Uhlenbeck process rather than the classic ODE formulation dz(t) dt = f (zt). We account for the increased uncertainty by assuming an additional state specific Gaussian noise factor ? i.e. assuming f (x + Ot) + Ot + ? for a sampled trajectory Ot in the gradient matching formulation (10). 5

74

5

Experiments

75 6 lower Discussion tight variational bounds that are analytically tractable provided that the ODE is such th state variables appear in quadratic form in equation 6. ODE?s such as the Lotka-Volterra s 76 ODE The contribution ofwhereas this paper is tosystems integratesuch out the latent state variables instead 80 full-fill such requirements other as the Fitz-High Nagumo sytemo 77 lower in previous work. over stateprovided variablesthat is not trac ??bounds ? (i?1) ?the integration 78 tight variational thatSince are analytically tractable theanalytically ODE is such ? ? 81 D Q (X) p(?, X — Y, ?, ?) KL 78 appear tight variational lower bounds that are analytically that the OD 79 state variables in quadratic form in equation 6. ODE?s tractable such as provided the Lotka-Volterr 79 state variables appear in quadratic form in equation ??. ODE?s such as the Lotk 80 full-fill 82 LQsuch (t) (?)ODE requirements whereas other systems such as the Fitz-High Nagumo syte 80 full-fill such ODE requirements whereas other systems such as the Fitz-High Nag ?? ? ?(i?1) ? tight variational lower bounds that are analytically tractable ?kernel ??) ?Y, ?p(?, ?X ? provided that the ODE is such that t GP 81 D83KLlogQ p(?,81 (X) — Y, ?, X —D ?, ?)dX (i?1) ? ? (X) p(?, X — Y, ?, ?) KL Q

78 79

Flexibility and Efficiency Algorithm 1 offers 78 parameters appear in quadratic form in equation 6. ODE?s such as the Lotka-Volterra syste a flexible framework for inference in stochas- 79 state variables state-derivative (i+1) 82 L84Qsuch state ?(?) ODE requirements noisesuch as the Fitz-High Nagumo sytem do n ? 82 LQobservations (?) whereas other systems b 80 full-fill tic dynamical systems e.g. if the parameters ? ??78 tight ? ? (i?1) bounds that are analytically tractable provided that ? variational ?? lower ? ? ? Q p(?, (X) X ?)dX — p(?, Y, ?, X??79 —83p(?, Y,log ?, X ?) — Y,appear ?, ?)dX are known they can be set to the true values in 81 D83KL log state variables in quadratic in equation 6.? ODE?s such 2as ?th ? k — Fkform max log N Y (Y, ?, ? I) N such Xk as— Y ? I 85 References k not analytically tractable 80 full-fill such ODE requirements whereas other systems thek ,Fitz-Hi (i) each iteration, and algorithm 1 then just corre- 82 L84Q ??(?) ? ? Y ??kbounds ?Girolami ? 78 84 tight variational lower that are analytically tractable provided that the ? References 86 B. Calderhead, M. and N. Lawrence, ?Accelerating bayesian inference over no (i?1) latent state ? ? ??) ? ? sponds to a smoothing algorithm. Compared to 83 log p(?, X —E-??step 81 DKL Q (X) p(?, X — Y, ??, Y, ?)dX 79 ?, state variableswith appear invariables quadratic form in equation 6. suchSystems, as the Lv 87 differential equations gaussian processes,? Neural Information Processing ? kODE?s = max H(Q) + E log N Y — F (Y, ?, ? I) 85 B. Calderhead, M. Girolami and N. Lawrence, ?Accelerating bayesian infere the smoothing algorithm in Archambeau et al. Q k k full-fill such(?) ODE?requirements whereas other systems such as the Fitz-High N 88 no.80429-443, 82 LQ2008. (i?1) 85 References 86 differential with gaussian ?? ? ? (i?1) equations k ? processes,? Neural Information Processin [2008] it does not require the computational ex- 84 ?? ?2008. ?S. 78Dondelinger, tight variational lower bounds are analytically tractable that the ad O ? provided 87 no. 429-443, M-??step 81 83 D Q p(?, X —that Y, ? 89 F. M. Filippone, Rogers and?, D.?) Husmeier, ?Ode parameter inference using KL logGirolami p(?, (X) X —and Y, ?, ?)dX 86 B. Calderhead, M. N. Lawrence, ?Accelerating bayesian inference over ?in ? ? pensive forward and backward propagation us? k — such 79gradient state variables appear in quadratic form equation 6.pp.ODE?s as the Lot 90 matching with with gaussian processes,? AISTATS, vol. 31,N 216?228, 2013. Y F (Y, ?, ? I k k ? ? 87 differential equations gaussian processes,? Neural Information Processing Systems 88 F. Dondelinger, M. Filippone, S. Rogers and D. Husmeier, ?Ode parameter infere (i) 82 LQ ?such (?)ODE requirements? 80 full-fill whereas such as the Fitz-High Na ?other ? systems ing an ODE solver. If the parameters are not 85 References ? + D Q(?) 84 ?? gradient KL 89 2008. matching with gaussian processes,? AISTATS, vol. 31, pp. 216?228, 2 88 no.E-??step 429-443, ? ??analytically tractable for? a ?? ? ? k — Fk (Y, ?, ?k I) 83 logQ(i?1) p(?,(X) X —?Y, ?, ?)dX ?p(?, known then algorithm 1 offers a grid free infer- 86 B. Calderhead, restricted of ?Accelerating ODE's k bayesian ?kN Y 81 D X —family Y, ?, ?) M. and N. Lawrence, inference over nonlin KLGirolami Y 89 F. Dondelinger, M. Filippone, S. Rogers and Neural D. Husmeier, ?Ode parameter equations with gaussian processes,? Information Processinginference

Systems, using vol. 2 ence procedure for estimating the parameters. 87 90differential 85? (i)References 82 L (?) gradient matching with gaussian processes,? AISTATS, vol. 31, pp. 216?228, 2013. 84 ? M-??step 2008. Q no. 429-443, Opposite to Vrettas et al. [2011] which consider 88 61 ? We can establish touching lower bounds since we can solve the in 86 B. X Calderhead, M. Girolami and N. Lawrence, ?Accelerating bayesian 83 log p(?, — S. Y,Rogers ?, ?)dX unobserved state variables in the case of smooth- 89 F. Dondelinger, M. 87Filippone, and D. with Husmeier, ?Ode parameterNeural inference using adapti differential equations gaussian processes,? Information Pro (ODE parameters) ? 8562 References gradient matching gaussian processes,? 429-443, 2008. AISTATS, vol. 31, pp. 216?228, 2013. ing but assume the system to be fully observed if 90 ? (i)88 withno. 84 ? 86 B. Calderhead, M. Girolami and N. Lawrence, ?Accelerating bayesian infe parameters are estimated, the outlined approach Figure 2: Illustration F. Dondelinger, M. climbing" Filippone, S. Rogers and D.Neural Husmeier, ?Ode parameter of theequations "hill algo87 89 differential with gaussian processes,? Information Process offers an efficient inference framework for the rithm in Gorbach matching with difference gaussian processes,? AISTATS, vol. 31, pp. 216? 88 90 no. 429-443, 2008.. The et gradient al. [2017] 85 References much more complicated problem of inferring between the lower bound L (?) andandthe log in(?) bM. 89 Calderhead, F. Dondelinger, Filippone, S.N. Rogers and D. ?Accelerating Husmeier, ?Odebayesian parameter infe Q 86 B. M. Girolami Lawrence, infer the parameters while not all states are observed 90by gradient matching with with gaussian gaussian processes,? vol. 31, pp. Processin 216?228 87 differential equations processes,?AISTATS, Neural Information tegral is given the Kullback-Leibler divergence and still scales linearly in the states if pathwise no. 429-443, 2008. (red line). 88 inference of the RODE is done in parallel. (t)

(t)

(t)

(i+1)

(i)

(i?1)

89

F. Dondelinger, M. Filippone, S. Rogers and D. Husmeier, ?Ode parameter infer

90 gradient matching with gaussian processes,? AISTATS, vol. 31, pp. 216?228, 2 The conceptual difference between the approach of Vrettas et al. [2015] and Gorbach et al. [2017] is illustrated in figure 3.

Figure 3: Conceptual Difference. The red line represents an artificial function which has to be approximated. Our approach (right) is grid free and based on the minimization of the differences of the slopes. That is why convergence is vertical with each iteration step corresponding to a dashed line (thickness of the line indicating the convergence direction). Vrettas et al. [2015] approximate 4 the true process by a linearized dynamic process which is discretized (left) and improved by iterated forward and backward smoothing.

4

Experiments

4

4

We compare our approach on two established benchmark models for stochastic systems especially 4 used for weather forecasts. Vrettas et al. [2011] provide an extensive comparison of the approach of Archambeau et al. [2008] and its improvements compared to classic Kalman filtering as well as more advanced and state of the art inference schemes like 4D-Var [Le Dimet and Talagrand, 1986]. We use the reported results there as a comparison measure.

4 4

The drift function for the Lorenz96 system consists of equations of the form: fk (x(t), ?) = (xk+1 ? xk?2)xk?1 ? xk + ? where ? is a scalar forcing parameter, x?1 = xK?1 , x0 = xK and xK+1 = x1 (with K being the number of states in the stochastic system (1)). The Lorenz96 system can be seen as a minimalistic weather model [Lorenz and Emanuel, 1998]. 6

4

4

The three dimensional Lorenz attractor is described by the parameter vector ? = (?, ?, ?) and the following time evolution: " # ?(x2 (t) ? x1 (t)) 1 dX(t) = ?x1 (t) ? x2 (t) ? x1 (t)x3 (t) dt + ? 2 dWt x1 (t)x2 (t) ? ?x3 (t) The runtime for state estimation using the approach of Vrettas et al. [2011] and our method is indicated in table 1. While parameter and state estimation are combined in one step in our approach, parameter estimation using the approach of Vrettas et al. [2011] would imply the iterative use of the smoothing algorithm and thus a multiple factor of the runtime indicated in table 1. While we solve a much more difficult problem by inferring parameters and states at the same time our runtime is only a fraction of the runtime awarded for a single run of the inner loop for parameter estimation in Vrettas et al. [2011]. Method VGPA_MF Our approach

L63/D=3 31s 2.4s

L96/D=40 6503s 14s

L96/D=1000 17345s 383s

Table 1: Runtime for one run of the smoothing algorithm of the approach of Vrettas et al. [2015] vs the runtime of our approach in parallel implementation (using 51 OU sample paths). While parameter estimation is done simultaneously in our approach, Vrettas et al. [2015] use the smoothing algorithm iteratively for state estimation in an inner loop such that the runtime for parameter estimations is multiple times higher than the indicated runtime for just one run of the smoothing algorithm. We use our method to infer the states and drift parameters for the Lorenz attractor where the dimension y is unobserved. The estimated state trajectories are shown in figure 4. Estimated Trajectory

z

z

Simulated Trajectory

x

y

x

y

Figure 4: Lorenz attractor. The Lorenz attractor trajectories are shown on the right -hand side for inferred solutions using an SDE solver, while the left-hand side plot shows the inferred trajectory using our method. Our method was able to accurately resolve the typical ?butterfly? pattern despite not observing the drift parameters as well as not observing the dimension y. Only the dimensions x and z were observed. The estimated trajectories for one sample path are also shown in the time domain in section 5.2 of the supplementary material. Our approach offers an appealing shortcut to the inference problem for stochastic dynamical systems and is robust to the noise in the diffusion term. Figure 5 shows the dependence of the inferred parameters on the variance in the diffusion term of the stochastic differential equation. Increasing the time interval of the observed process e.g. from 10 to 60 secs leads to a converging behaviour to the true parameters (figure 6). This is in contrast to the reported results of Archambeau et al. [2008], reported in Vrettas et al. [2011, Figure 29] and shows the asymptotic time consistency of our approach. Figure 5 shows, that in the near noiseless scenario we approximately identify sigma correctly. Estimating the ? term in Figure 6 is more difficult than the other two parameters in the drift 7

est

11

est

30

est

2.8 2.6

10

29 2.4

9 28

2.2

8 2

27 7

1.8 26

6

1.6

5

30

18

6

25

1

30

18

6

1.4

1

30
18
diffusion
diffusion
6
1
diffusion

Figure 5: Lorenz attractor. Boxplots indicate the median of the inferred parameters over 51 generated OU sample paths. Using a low variance for the diffusion term in simulating one random sample path from the SDE, our approach infers approximately the correct parameters and does not completely deteriorate if the variance is increased by a factor of 30. est

11
est
28.5
est
2.7 28
10
2.6 27.5 2.5
9 27
2.4
8 26.5 7
2.3
26
6
10
20
30
40
50
60
25.5
2.2
10
20
final time
30
40
50
60
2.1
10
20
final time
30
40

50

60

final time

Figure 6: Lorenz attractor. Increasing the time interval of the observed process leads to a convergence towards the true parameters opposed to the results in [Vrettas et al., 2011, Figure 29].

function of the Lorenz attractor system, since the variance of the diffusion and the observation noise unfortunately lead to an identifiability problem for the parameter sigma, which is why longer time periods in Figure 6 do not improve the estimation accuracy for ?. est

10

est

9 8

9

7 6

8 5 4

7

3 6

2 1 1

5

10

20

0

diffusion

fully obs.

2/3 observed

1/2 observed

1/3 observed

Figure 7: Lorenz96. Left hand side shows the accuracy of the parameter estimation with increasing diffusion variance (right to left) for a 40 dimensional system, while the plot on the right hand side shows the accuracy with decreasing number of observations. Red dots show the results of the approach of Archambeau et al. [2008] when available as reported in Vrettas et al. [2011]. The correct parameter has the value 8 and our approach performs significantly better, while having a lower runtime and is furthermore able to include unobserved variables (right) For the Lorenz96 system our parameter estimation approach is likewise robust to the variance in the diffusion term (figure 7). It furthermore outperforms the approach of Archambeau et al. [2008] in the cases where results were reported in Vrettas et al. [2011]. The performance level is equal when, for our approach, we assume that only one third of the variables are unobserved. 8

The estimated trajectories for one sample path of the Lorenz96 system are shown in section 5.3 of the supplementary material.

5

Discussion

Parameter inference in stochastic systems is a challenging but important problem in many disciplines. Current approaches are based on exploration in the parameter space which is computationally expensive and infeasible for larger systems. Using a gradient matching formulation and adapting it to the inference of random ordinary differential equations, our proposal is a flexible framework which allows to use deterministic calculus for inference in stochastic systems. While our approach tackles a much more difficult problem by combining state and parameter estimation in one step, it offers improved accuracy and is orders of magnitude faster compared to current state of the art methods based on variational inference.

Acknowledgements This research was partially supported by the Max Planck ETH Center for Learning Systems and the SystemsX.ch project SignalX.

9

2 References

C?dric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John S Shawe-taylor. Variational inference for diffusion processes. Neural Information Processing Systems (NIPS), 2008. Yusuke Asai and Peter E Kloeden. Numerical schemes for random odes via stochastic differential equations. Commun. Appl. Anal, 17(3):521?528, 2013. Yusuke Asai, Eva Herrmann, and Peter E Kloeden. Stable integration of stiff random ordinary differential equations. Stochastic Analysis and Applications, 31(2):293?313, 2013. Ben Calderhead, Mark Girolami and Neil D. Lawrence. Accelerating bayesian inference over nonliner differential equations with gaussian processes. Neural Information Processing Systems (NIPS), 2008. Halim Doss. Liens entre ?quations diff?rentielles stochastiques et ordinaires. In Annales de l?IHP Probabilit?s et statistiques, volume 13, pages 99?125, 1977. Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 53(4):343?367, 2003. Nico S Gorbach, Stefan Bauer, and Joachim M Buhmann. Scalable variational inference for dynamical systems. arXiv preprint arXiv:1895944, 2017. Lars Gr?ne and PE Kloeden. Pathwise approximation of random ordinary differential equations. BIT Numerical Mathematics, 41(4):711?721, 2001. Xiaoying Han and Peter E Kloeden. Random ordinary differential equations and their numerical solution, 2017. Peter Imkeller and Bj?rn Schmalfuss. The conjugacy of stochastic and random differential equations and the existence of global attractors. Journal of Dynamics and Differential Equations, 13(2): 215?249, 2001. Arnulf Jentzen and Peter E Kloeden. Pathwise taylor schemes for random ordinary differential equations. BIT Numerical Mathematics, 49(1):113?140, 2009. Arnulf Jentzen and Peter E Kloeden. Taylor approximations for stochastic partial differential equations. SIAM, 2011. Peter E Kloeden and Arnulf Jentzen. Pathwise convergent higher order numerical schemes for random ordinary differential equations. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 463, pages 2929?2944. The Royal Society, 2007. Fran?ois-Xavier Le Dimet and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A: Dynamic Meteorology and Oceanography, 38(2):97?110, 1986. Edward N Lorenz and Kerry A Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. Journal of the Atmospheric Sciences, 55(3):399?414, 1998. Simon Lyons, Amos J Storkey, and Simo S?rkk?. The coloured noise expansion and parameter estimation of diffusion processes. Neural Information Processing Systems (NIPS), 2012. Christoph Riesinger, Tobias Neckel, and Florian Rupp. Solving random ordinary differential equations on gpu clusters using multiple levels of parallelism. SIAM Journal on Scientific Computing, 38(4): C372?C402, 2016. Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. Monthly Weather Review, 136(12):4629?4640, 2008. H?ctor J Sussmann. On the gap between deterministic and stochastic ordinary differential equations. The Annals of Probability, pages 19?41, 1978. 10

Christoffer W Torn?e, Rune V Overgaard, Henrik Agers?, Henrik A Nielsen, Henrik Madsen, and R implementation, application, E Niclas Jonsson. Stochastic differential equations in nonmem : and comparison with ordinary differential equations. Pharmaceutical research, 22(8):1247?1258, 2005. Michail D Vrettas, Dan Cornford, and Manfred Opper. Estimating parameters in stochastic systems: A variational bayesian approach. Physica D: Nonlinear Phenomena, 240(23):1877?1900, 2011. Michail D Vrettas, Manfred Opper, and Dan Cornford. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. Physical Review E, 91(1):012148, 2015.

11