Supplement: Differentiable Likelihoods for Fast Inversion of ‘Likelihood-Free’ Dynamical Systems

A. Short Introduction to Gaussian ODE Filtering

A.1. Gaussian Filtering for Generic Time Series

In signal processing, a Bayesian Filter (Särkkä, 2013, Chapter 4) does Bayesian inference of the discrete state $\{x_i;\ i = 1, \dots, N\} \subset \mathbb{R}^n$ from measurements $\{y_i;\ i = 1, \dots, N\} \subset \mathbb{R}^n$ in a *probabilistic state space model* consisting of

$$\text{a dynamic model} \quad x_i \sim p(x_i \mid x_{i-1}), \quad \text{and} \tag{A.1}$$

$$\text{a measurement model} \quad y_i \sim p(y_i \mid x_i). \tag{A.2}$$

Usually, the state x_i is assumed to be the discretization of a continuous signal $x : [0, T] \to \mathbb{R}^n$ which is *a priori* modeled by a stochastic process. Absent very specific expert knowledge, this prior is usually chosen to be a linear time-invariant (LTI) stochastic differential equation (SDE):

$$p(x) \ \sim \ X(t) = FX(t)\,\mathrm{d}t + L\,\mathrm{d}B(t), \tag{A.3}$$

where F and L are the drift and diffusion matrix, respectively. The corresponding dynamic model (eq. (A.1)) can be easily constructed by discretization of the LTI SDE (eq. (A.3)), as described in Särkkä & Solin (2019, Chapter 6.2). If an LTI SDE prior with Gaussian initial condition is used, $p(x)$ is a GP which implies a Gaussian dynamic model

$$p(x_i \mid x_{i-1}) = \mathcal{N}(Ax_{i-1}, Q) \tag{A.4}$$

for matrices A, Q that are implied by F, L from eq. (A.3). If additionally the measurement model (eq. (A.2)) is Gaussian, i.e.

$$p(y_i \mid x_i) = \mathcal{N}(Hx_i, R) \tag{A.5}$$

for matrices H, R, the filtering distributions $p(x_i \mid y_{1:i})$, $i = 1, \dots, N$, can be computed by Gaussian filtering in linear time. Note that the filtering distribution $p(x_i \mid y_{1:i})$ is not the full posterior distribution $p(x_i \mid y_{1:N})$ which can, however, also be computed in linear time by running a smoother after the filter. See e.g. Särkkä (2013) for more information.

A.2. Gaussian ODE Filtering

A Gaussian ODE filter is simply a Gaussian filter, as defined in Section A.1, with a specific kind of probabilistic state space model eqs. (A.1) and (A.2), to infer the solution $x : [0, T] \to \mathbb{R}^d$ of the ODE eq. (1), at the discrete time grid $\{0 \cdot h, \dots, N \cdot h\}$ with step size $h > 0$. The dynamic model is—as usual, recall eqs. (A.3) and (A.4)—constructed from a GP defined by a LTI SDE that incorporates the available prior information on x. The measurement model, however, is specific to ODEs as we will see next: Recall that, after $i - 1$ steps, the Gaussian filter has computed the $(i - 1)$-th filtering distribution

$$p(x_{i-1} \mid y_{1:i-1}) = \mathcal{N}(m_{i-1}, P_{i-1}), \tag{A.6}$$

which is Gaussian with mean m_{i-1} and covariance matrix P_{i-1}, and computes the predictive distribution

$$p(x_i \mid y_{1:i-1}) = \mathcal{N}(m_i^-, P_i^-) \tag{A.7}$$

by inserting eq. (A.4) into eq. (A.6). Analogous to the logic

$$f(\hat{x}(t)) \approx f(x(t)) = \dot{x}(t) \tag{A.8}$$

of classical solvers, the Gaussian ODE Filter treats evaluations at the predictive mean m_i^-—which is a numerical approximation like $\hat{x}$—as data on $\dot{x}(ih)$. This yields the measurement model

$$p(y_i \mid x_i) = \mathcal{N}(Hx_i, R), \tag{A.9}$$

with data

$$y_i := f(m_i^-) \approx \dot{x}(ih). \tag{A.10}$$

The probabilistic state space model is thereby completely defined. Gaussian ODE filtering is equivalent to running a Gaussian filter on this probabilistic state space model.
For more details on Gaussian ODE filters, see Kersting et al. (2019) or **?**. An extension to more Bayesian filters—such as particle filters—is provided by Tronarp et al. (2019).

B. Equivalent Form of Filtering Distribution by GP Regression

Recall from Section A that any Gaussian filter computes a sequence of filtering distributions

$$p(x_i \mid y_{1:i}) = \mathcal{N}(m_i, P_i) \tag{B.1}$$

from a GP prior on x eq. (A.3) and a linear Gaussian measurement model (eq. (A.5)) with derivative data (eq. (A.10)). Hence, the classical framework for GP regression with derivative observations, as introduced in Solak et al. (2003), is applicable. It *a priori* models the state x and its derivative $\dot{x}$ as a multi-task GP:

$$p\left(\begin{bmatrix} x \\ \dot{x} \end{bmatrix}\right) = \mathcal{GP}\left(\begin{bmatrix} x \\ \dot{x} \end{bmatrix}; \begin{bmatrix} \mu \\ \dot{\mu} \end{bmatrix}, \begin{bmatrix} k & k^{\partial} \\ {}^{\partial}k & {}^{\partial}k^{\partial} \end{bmatrix}\right), \tag{B.2}$$

with

$${}^{\partial}k = \frac{\partial k(t,t')}{\partial t},\ k^{\partial} = \frac{\partial k(t,t')}{\partial t'},\ {}^{\partial}k^{\partial} = \frac{\partial^2 k(t,t')}{\partial t \partial t'}. \tag{B.3}$$

B.1. Kernels for Derivative Observations

In this paper, we model the solution x with a integrated Brownian motion kernel k or, in other words, we model $\dot{x}$ by the Brownian Motion (a.k.a. Wiener process) kernel, i.e.

$${}^{\partial}k^{\partial}(t,t') = \sigma^2_{\text{dif}} \min(t,t'), \qquad \forall t,t' \in [0,T]. \tag{B.4}$$

Here, $\sigma_{\text{dif}} > 0$ denotes the output variance which scales the diffusion matrix L in the equivalent SDE (eq. (A.3)). Integration with respect to both arguments yields the integrated Brownian motion (IBM) kernel

$$k(t,t') = \sigma^2_{\text{dif}}\left(\frac{\min^3(t,t')}{3} + |t-t'|\frac{\min^2(t,t')}{2}\right) \tag{B.5}$$

to model x. The once-differentiated kernels in eq. (B.2) are given by

$$k^{\partial}(t,t') = {}^{\partial}k(t',t) = \sigma^2_{\text{dif}} \begin{cases} t \leq t' : \frac{t^2}{2}, \\ t > t' : tt' - \frac{t'^2}{2} \end{cases}. \tag{B.6}$$

A detailed derivation of eqs. (B.4) to (B.6) can be found in Schober et al. (2014, Supplement B).

B.2. GP Form of Filtering Distribution

Now, GP regression with prior (eq. (B.2)), likelihood (eq. (B.1)) and data $y_{1:i}$ yields an equivalent form of the filtering distribution eq. (B.1):

$$\begin{aligned} m_i =& \mu + k^{\partial}(h:ih, ih)^{\intercal}\left[{}^{\partial}K^{\partial}(h:ih) + R \cdot I_i\right]^{-1} \\ & \times [y_1 - \dot{\mu}(h), \dots, y_i - \dot{\mu}(ih)]^{\intercal}, \end{aligned} \tag{B.7}$$

$$\begin{aligned} P_i =& \begin{bmatrix} k(h,h) & \dots & k(ih,ih) \\ \vdots & \ddots & \vdots \\ k(ih,h) & \dots & k(ih,ih) \end{bmatrix} - k^{\partial}(h:ih,ih)^{\intercal} \\ & \times \left[{}^{\partial}K^{\partial}(h:ih) + R \cdot I_l\right]^{-1} k^{\partial}(h:ih,ih), \end{aligned} \tag{B.8}$$

with $y_{1:i} = [y_1, \dots, y_i]^{\intercal}$, where we used the notations from eqs. (15) and (16). The derivation of eq. (18) is hence concluded by eq. (B.8).

B.3. Derivation of Equation (10)

In this subsection, we will use the ODE-specific notation from above instead of the generic filtering notation—e.g. $m_\theta(ih)$ instead of m_i, $f(m^{-}(ih))$ instead of y_i etc. To derive the missing eq. (10), we first observe that, by eq. (B.7), $m(ih)$ is linear in the data residuals:

$$\begin{aligned} m_\theta(ih) &= \mu + \beta_{ih} \times \\ &\qquad \left[f(m^{-}(h)) - \dot{\mu}(h), \dots, f(m^{-}(ih)) - \dot{\mu}(ih)\right]^{\intercal} \\ \beta_{ih} &:= k^{\partial}(h:ih,ih)^{\intercal}\left[{}^{\partial}K^{\partial}(h:ih) + R \cdot I_i\right]^{-1}. \end{aligned} \tag{B.9}$$

Now recall that, in ODE filtering, the prior mean in eq. (B.2) is set to be $[\mu, \dot{\mu}] \equiv [x_0; f(x_0)]$ (or $[\mu, \dot{\mu}] \equiv [m_0; f(m_0)]$ for some estimate m_0 of x_0, in the case of unknown x_0). Consequently, application of Assumption 1 to eq. (B.9) yields

$$m_\theta(ih) = x_0 + J_{ih}\theta, \quad \text{with} \tag{B.10}$$

$$\begin{aligned} J_{ih} &:= \beta_{ih} \begin{bmatrix} f_1(m_\theta^{-}(h)) - f_1(x_0) & \dots & f_n(m_\theta^{-}(h)) - f_n(x_0) \\ \vdots & \ddots & \vdots \\ f_1(m_\theta^{-}(ih)) - f_1(x_0) & \dots & f_n(m_\theta^{-}(ih)) - f_n(x_0) \end{bmatrix} \\ &= \beta_{ih} Y_{1:i} \quad , \end{aligned} \tag{B.11}$$

where $Y_{1:i}$ denotes the first i rows of Y; see eq. (17). We omit the dependence of J_{ih} on θ to obtain a linear form. Recall from Section 3 that we may w.l.o.g. assume that the time points $\{t_1, \dots, t_M\}$ lie on the filter time grid, i.e. $t_i = l_i h$ from some $l_i \in \mathbb{N}$. Therefore, eq. (B.10) implies

$$m_\theta(t_i) \overset{eq.\ (14)}{=} x_0 + \tilde{\kappa}_i Y_{1:i} \overset{eq.\ (13)}{=} x_0 + \kappa_i Y \tag{B.12}$$

for all data time points t_i, $i = 1, \dots, M$. Here, we used that $\tilde{\kappa}_i$ is equal to $\beta_{l_i h}$ by eq. (14). We conclude the derivation of eq. (10) by observing that the i-th entry of eq. (10) reads eq. (B.12) for all $i = 1, \dots, M$.

C. Proof of Theorem 1

Proof. We start by computing the rows of

$$D\boldsymbol{m}_\theta = [\nabla_\theta m(t_1), \dots, \nabla_\theta m(t_M)]^{\intercal}. \tag{C.1}$$

By eqs. (10) and (11) and the fact that the kernel prefactor K does not depend on θ, we obtain, for all $i = 1, \dots, M$, that

$$\begin{aligned} \nabla_\theta m(t_i) &= \nabla(\tilde{\kappa}(i)^{\intercal} v(\theta)) \\ &= [Dv(\theta)]^{\intercal}\, \tilde{\kappa}(i) + \underbrace{[D\tilde{\kappa}(i)]^{\intercal}}_{=0} v(\theta) \qquad \text{(C.2)} \\ &= [Dv(\theta)]^{\intercal}\, \tilde{\kappa}(i), \qquad \text{(C.3)} \end{aligned}$$

with $v(\theta) = \tilde{Y}\theta$. Here,

$$\tilde{Y} = Y[1:l_i, :] = [Y_1(\theta), \dots, Y_{l_i}(\theta)]^{\intercal} \tag{C.4}$$

is defined by

$$Y_j(\theta) = [y_{j1}, \dots, y_{jn}]^\intercal \in \mathbb{R}^n, \tag{C.5}$$

the j-th row of $Y = Y(\theta)$ (recall eq. (17)), for $j = 1, \dots, l_i$. Next, we again compute the rows of the missing Jacobian of eq. (C.3)

$$Dv(\theta) = [\nabla_\theta[v(\theta)]_1, \dots, \nabla_\theta[v(\theta)]_{l_i}]^\intercal \tag{C.6}$$

by the chain rule, for all $j \in \{1, \dots, l_i\}$:

$$\nabla_\theta[v(\theta)]_j = \nabla_\theta[Y_j(\theta)^\intercal \theta] = [DY_j(\theta)]^\intercal \, \theta + Y_j(\theta). \tag{C.7}$$

Again, we compute the rows of the final missing Jacobian

$$DY_j(\theta) = [\nabla_\theta y_{j1}(\theta), \dots, \nabla y_{jn}(\theta)]^\intercal. \tag{C.8}$$

The definition of y_{ij} from eq. (17) implies, in the notation of eq. (21), that

$$[\nabla_\theta y_{jk}(\theta)]_l = \lambda_{lk}(jh), \tag{C.9}$$

for all $l = 1, \dots, n$. Now, we can insert backwards. First, we insert eq. (C.9) into eq. (C.8) which yields

$$DY_j(\theta) = \Lambda_j, \tag{C.10}$$

where $\Lambda_j = \left[\lambda_{kl}(jh)\right]_{k,l=1,\dots,n}$. Second, insertion of eq. (C.10) into eq. (C.7) provides that

$$\nabla_\theta[v(\theta)]_j = \Lambda_j^\intercal \theta + Y_j(\theta). \tag{C.11}$$

Third, insertion of eq. (C.11) into eq. (C.6) implies that

$$Dv(\theta) = \left[\Lambda_1^\intercal \theta, \dots, \Lambda_{l_i}^\intercal \theta\right]^\intercal + Y[: l_i, :], \tag{C.12}$$

where

$$Y[: l_i, :] \overset{eq.\ (C.11)}{=} [Y_1(\theta), \dots, Y_{l_i}(\theta)]^\intercal \overset{eq.\ (C.5)}{=} \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{l_i 1} & \cdots & y_{l_i n} \end{bmatrix}.$$

Fourth, we insert eq. (C.12) into eq. (C.3) and obtain

$$\begin{aligned} \nabla_\theta m(t_i) &= \left([Y[: l_i, :]]^\intercal + \left[\Lambda_1^\intercal \theta, \dots, \Lambda_{l_i}^\intercal \theta\right]\right) \tilde{\kappa}_i \\ &= [Y[: l_i, :]]^\intercal \, \tilde{\kappa}_i + \left[\Lambda_1^\intercal \theta, \dots, \Lambda_{l_i}^\intercal \theta\right] \tilde{\kappa}_i. \end{aligned} \tag{C.13}$$

By eq. (13), it follows that

$$[Y[: l_i, :]]^\intercal \, \tilde{\kappa}_i \overset{eq.\ (17)}{=} Y^\intercal \kappa_i, \qquad \text{and} \tag{C.14}$$

$$\left[\Lambda_1^\intercal \theta, \dots, \Lambda_{l_i}^\intercal \theta\right] \tilde{\kappa}_i \overset{eq.\ (20)}{=} S^\intercal \kappa_i. \tag{C.15}$$

This implies via eq. (C.13) that

$$\nabla_\theta m(t_i) = (Y^\intercal + S^\intercal) \, \kappa_i, \tag{C.16}$$

Fifth and finally, we, by insertion of eq. (C.16) into eq. (C.1) and application of eq. (12), obtain

$$Dm_\theta = K(Y + S) \overset{eq.\ (11)}{=} J + KS. \tag{C.17}$$

□

D. Proof of Theorem 2

We first show some preliminary technical lemmas in Section D.1 which are needed to prove bounds on $\|K\|$ and $\|S\|$ in Section D.2 and Section D.3, respectively. Having proved these bounds, the core proof of Theorem 2 simply consists of combining them by Theorem 1, as executed in Section D.4.

D.1. Preliminary lemmas

The following lemma will be needed in Section D.2 to bound $\|K\|$.

Lemma 1. *Let $Q > 0$ be a symmetric positive definite and $Q' \geq 0$ a symmetric positive semi-definite matrix in $\mathbb{R}^{m \times n}$. Then, it holds true that*

$$\left\| [Q + Q']^{-1} \right\|_* \leq \|Q^{-1}\|_*, \tag{D.1}$$

for the nuclear norm

$$\|A\|_* = \operatorname{trace} \sqrt{A^* A} = \sum_{i=1}^{m \wedge n} \sigma_i(A), \tag{D.2}$$

where $\sigma_i(A)$, $i \in \{1, \dots, m \wedge n\}$, are the singular values of A.

Proof. Recall that, for all symmetric positive semi-definite matrices, the singular values are the eigenvalues. Therefore

$$\begin{aligned} \left\| [Q + Q']^{-1} \right\|_* &= \sum_{i=1}^{m \wedge n} \frac{1}{\lambda_i(Q + Q')} \\ &\leq \sum_{i=1}^{m \wedge n} \frac{1}{\lambda_i(Q)} = \|Q^{-1}\|_*. \end{aligned} \tag{D.3}$$

In eq. (D.3), we exploited the fact that $Q \leq Q + Q'$ (i.e. that $(Q + Q') - Q = Q'$ is positive semi-definite) and therefore $\lambda_i(Q) \leq \lambda_i(Q + Q')$ for ordered eigenvalues $\lambda_1(Q) \leq \cdots \leq \lambda_{m \wedge n}(Q)$ counted by algebraic multiplicity. This fact is an immediate consequence of Theorem 8.1.5. in Golub & Van Loan (1996). □

The next lemma will be necessary to prove a bound on $\|S\|$ in Section D.3.

Lemma 2. *Let $g(x, \lambda) \in C\left([0, T] \times \Lambda; \mathbb{R}\right)$ on non-empty compact $\Lambda \subset \mathbb{R}^n$ with continuous first-oder partial derivatives w.r.t. the components of λ. If*

$$\sup_{\lambda \in \Lambda} g(x, \lambda) \in \mathcal{O}(h(x)) \tag{D.4}$$

for some constant $C > 0$ and some strictly positive $h : [0, T] \to \mathbb{R}$, then also

$$\sup_{\lambda \in \Lambda^\circ} \left| \frac{\partial}{\partial \lambda_k} g(x, \lambda) \right| \in \mathcal{O}(h(x)), \tag{D.5}$$

where Λ° denotes the interior of Λ.

Proof. Assume not. Then, there is a $k \in \{1, \dots, n\}$ and a $\tilde{\lambda} \in \Lambda^{\mathrm{o}}$ such that

$$\left| \frac{\partial}{\partial \lambda_k} g(x, \tilde{\lambda}) \right| \notin \mathcal{O}(h(x)). \tag{D.6}$$

Since, for all $x \in [0, T]$, $\frac{\partial}{\partial \lambda_k}(x, \cdot)$ is uniformly continuous over the bounded domain Λ^{o}, there is a $\delta > 0$ such that

$$\left| \frac{\partial}{\partial \lambda_k} g(x, \tilde{\lambda}) \right| \notin \mathcal{O}(h(x)), \quad \text{for all } \lambda \in B_{2\delta}(\tilde{\lambda}). \tag{D.7}$$

Let us w.l.o.g. (otherwise consider $-g$) assume that

$$\frac{\partial}{\partial \lambda_k} g(x, \tilde{\lambda}) \geq 0, \quad \text{for all } \lambda \in B_{2\delta}(\tilde{\lambda}). \tag{D.8}$$

Now, on the one hand, we know by the fundamental theorem of calculus that

$$\begin{aligned} &\int_{-\delta}^{0} \frac{\partial}{\partial \lambda_k} g(x_n, \tilde{\lambda} + \tilde{\delta} e_k) \, \mathrm{d}\tilde{\delta} \\ &\quad = \underbrace{g(x, \tilde{\lambda})}_{\in \mathcal{O}(h(x))} - \underbrace{g(x, \tilde{\lambda} - \delta e_k)}_{\in \mathcal{O}(h(x))} \in \mathcal{O}(h(x)). \end{aligned} \tag{D.9}$$

However, on the other hand, we know from our assumption that

$$0 \overset{eq.\ (\mathrm{D.8})}{\leq} \int_{-\delta}^{0} \frac{\partial}{\partial \lambda_k} g(x_n, \tilde{\lambda} + \tilde{\delta} e_k) \, \mathrm{d}\tilde{\delta} \tag{D.10}$$

$$\leq \int_{-\delta}^{0} \underbrace{\left| \frac{\partial}{\partial \lambda_k} g(x_n, \tilde{\lambda} + \tilde{\delta} e_k) \right|}_{\notin \mathcal{O}(h(x)), \text{ by } eq.\ (\mathrm{D.7})} \mathrm{d}\tilde{\delta} \notin \mathcal{O}(h(x)), \tag{D.11}$$

which implies

$$\int_{-\delta}^{0} \frac{\partial}{\partial \lambda_k} g(x_n, \tilde{\lambda} + \tilde{\delta} e_k) \, \mathrm{d}\tilde{\delta} \notin \mathcal{O}(h(x)). \tag{D.12}$$

The desired contradiction is now found between eqs. (D.9) and (D.12). □

D.2. Bound on $\|K\|$

Lemma 3. *Under Assumption 3 and for all $R > 0$, it holds true that*

$$\|K\| \leq C(T), \tag{D.13}$$

where $C(T) > 0$ is a constant that depends on T.

Proof. First, recall eqs. (12) to (16) and observe that

$$\left\| k^{\partial}(h : t_i, t_i) \right\| \leq C \frac{\sigma^2}{2} \left\| [h^2, \dots, T^2] \right\|_{\infty} = C \left(2^{-\frac{1}{2}} \sigma T \right)^2,$$

for all $i = 1, \dots, M$. Second, Lemma 1 implies that

$$\begin{aligned} \left\| \left[{}^{\partial}K^{\partial}(h : t_i) + R \cdot I_{l_i} \right]^{-1} \right\| &\overset{eq.\ (\mathrm{D.1})}{\leq} C \left\| R^{-1} \cdot I_{l_i - 1} \right\|_* \\ &\leq C \left\| R^{-1} \cdot I_{\bar{N} - 1} \right\|_* \leq C R \bar{N}. \end{aligned}$$

Now, by eq. (13), we observe

$$\begin{aligned} \|\kappa_i\|_1 &= \|\tilde{\kappa}_i\|_1 \\ &\leq \left\| \left[{}^{\partial}K^{\partial}(h : t_i) + R \cdot I_{l_i} \right]^{-1} \right\| \cdot \left\| k^{\partial}(h : t_i, t_i) \right\| \\ &\leq C(T), \end{aligned} \tag{D.14}$$

where we inserted the above inequalities in the last step. Finally, we obtain eq. (D.13) by plugging eq. (D.14) into

$$\|K\| \leq C \|K\|_{\infty} \overset{eq.\ (12)}{=} \max_{1 \leq i \leq M} \|\kappa_i\|_1. \tag{D.15}$$

□

D.3. Bound on $\|S\|$

Before estimating $\|S\|$, we need to bound how far the entries of S (recall eq. (20)) deviate from the true sensitivities $\frac{\partial}{\partial \theta_k} x_\theta(T)$.

Lemma 4. *If $\Theta \subset \mathbb{R}^n$ is compact, then it holds true, under Assumptions 1 and 2, that*

$$\sup_{\theta \in \Theta^{\mathrm{o}}} \left\| \frac{\partial}{\partial \theta_k} m_\theta^-(T) - \frac{\partial}{\partial \theta_k} x_\theta(T) \right\| \in \mathcal{O}(h). \tag{D.16}$$

Proof. First, recall that the convergence rates of $\mathcal{O}(h)$ provided by Theorem 6.7 in Kersting et al. (2019) only depend on f through the dependence of the constant $K(T) > 0$ on the Lipschitz constant L of f. But this L is independent of θ by Assumption 1. Hence, Theorem 6.7 from Kersting et al. (2019) yields under Assumption 2 that

$$\sup_{\theta \in \Theta^{\mathrm{o}}} m_\theta^-(T) - x_\theta(T) \in \mathcal{O}(h). \tag{D.17}$$

Moreover, Theorem 8.49 in Kelley & Peterson (2010) is applicable under Assumption 1 and implies that $x_\theta(t)$ is continuous and has continuous first-order partial derivatives with respect to of θ_k. By construction—recall eq. (10)—the filtering mean $m_\theta(t)$ has the same regularity too. Hence, application of Lemma 2 with $x = h$, $\Lambda = \Theta$, $\lambda = \theta$, $g(x, \lambda) = m_\theta^-(T) - x_\theta(T)$ is possible, which yields eq. (D.16) from eq. (D.17). □

Lemma 5. *If $\Theta \subset \mathbb{R}^n$ is compact, then it holds true, under Assumptions 1 to 3, that*

$$\|S\| \leq C \left(\|\nabla_\theta x_\theta\| + h \right), \tag{D.18}$$

for sufficiently small $h > 0$.

Proof. By Assumption 3 and the equivalence of all matrix norms, we observe

$$\begin{aligned} \|S\| &\leq C\|S\|_2 = C\|S^\intercal\|_2 \leq C\|S^\intercal\|_{2,1} && \text{(D.19)}\\ &\overset{\text{eq. (20)}}{=} C\sum_{j=1}^{\bar{N}} \left\|\Lambda_j^\intercal\theta\right\|_2 && \text{(D.20)}\\ &\leq C\sum_{j=1}^{\bar{N}} \left\|\Lambda_j^\intercal\right\|_2 \underbrace{\|\theta\|_2}_{\leq C,\text{ since }\Theta\text{ bounded}}, && \text{(D.21)} \end{aligned}$$

where $\|\cdot\|_{2,1}$ denotes the $L_{2,1}$ norm. We conclude, using Assumption 2 and Lemma 4, that

$$\begin{aligned} \left\|\Lambda_j^\intercal\right\|_2 &\overset{\text{eq. (21)}}{\leq} L\max_{jk}\left[\frac{\partial}{\partial\theta_k} m_\theta^-(jh)\right] && \text{(D.22)}\\ &\overset{\text{eq. (D.16)}}{\leq} C\left(\|\nabla_\theta x_\theta\| + h\right). && \text{(D.23)} \end{aligned}$$

□

D.4. Proof of Theorem 2

Proof. By Theorem 1 and the sub-multiplicativity of the induced p-norm $\|\cdot\|_p$, we observe that

$$\begin{aligned} \|J - D\boldsymbol{m}_\theta\| = \|KS\| \leq C\|KS\|_p &\leq \|K\|_p\|S\|_q\\ &\leq C\|K\|\|S\|, && \text{(D.24)} \end{aligned}$$

for some $p, q \geq 1$. Application of Lemmas 3 and 5 concludes the proof. □

E. Gradient and Hessian Estimators for the Bayesian Case

In the main paper, we only consider the maximum likelihood objective; see eq. (23). Nonetheless, the extension to the Bayesian objective, with a prior $\pi(\theta)$, is straightforward:

$$-\log\left(p(\boldsymbol{z}\mid\theta)\pi(\theta)\right) = -\log\left(p(\boldsymbol{z}\mid\theta)\right) - \log\left(\pi(\theta)\right)$$

Accordingly, the gradients and Hessian of this objective are

$$\begin{aligned} \nabla_\theta\left[-\log\left(p(\boldsymbol{z}\mid\theta)\pi(\theta)\right)\right] &\overset{\text{eq. (26)}}{=} \hat{\nabla}_\theta E(\boldsymbol{z}) - \nabla_\theta\log\left(\pi(\theta)\right),\\ \nabla_\theta^2\left[-\log\left(p(\boldsymbol{z}\mid\theta)\pi(\theta)\right)\right] &\overset{\text{eq. (27)}}{=} \hat{\nabla}_\theta^2 E(\boldsymbol{z}) - \nabla_\theta^2\log\left(\pi(\theta)\right). \end{aligned}$$

Hence, for a Gaussian prior $\pi(\theta) = \mathcal{N}(\theta;\mu_\theta, V_\theta)$, the Bayesian version of the gradients and Hessian estimators in eqs. (26) and (27) are hence given by

$$\begin{aligned} \hat{\nabla}_\theta E(\boldsymbol{z})_{\text{Bayes}} &:= -J^\intercal\left[\boldsymbol{P} + \sigma^2 I_M\right]^{-1}\left[\boldsymbol{z} - \boldsymbol{m}_\theta\right]\\ &\quad - V_\theta^{-1}\left[\theta - \mu_\theta\right], \quad\text{and} && \text{(E.1)}\\ \hat{\nabla}_\theta^2 E(\boldsymbol{z})_{\text{Bayes}} &:= J^\intercal\left[\boldsymbol{P} + \sigma^2 I_M\right]^{-1} J + V_\theta^{-1}. && \text{(E.2)} \end{aligned}$$

F. Glucose Uptake in Yeast

The Glucose uptake in yeast (GUiY) is described by mass-action kinetics. In the notation of Schillings et al. (2015), the underlying ODE is given by:

$$\begin{aligned} \dot{x}^e_{\text{Glc}} &= -k_1 x^e_E x^e_{\text{Glc}} + k_{-1} x^e_{\text{E–Glc}}\\ \dot{x}^i_{\text{Glc}} &= -k_2 x^i_E x^i_{\text{Glc}} + k_{-2} x^i_{\text{E–Glc}}\\ \dot{x}^i_{\text{E–G6P}} &= k_4 x^i_E x^i_{\text{G6P}} + k_{-4} x^i_{\text{E–G6P}}\\ \dot{x}^i_{\text{E–Glc–G6P}} &= k_3 x^i_{\text{E–Glc}} x^i_{\text{G6P}} - k_{-3} x^i_{\text{E–Glc–G6P}}\\ \dot{x}^i_{\text{G6P}} &= -k_3 x^i_{\text{E–Glc}} x^i_{\text{G6P}} + k_{-3} x^i_{\text{E–Glc–G6P}}\\ &\quad - k_4 x^i_E x^i_{\text{G6P}} + k_{-4} x^i_{\text{E–Glc}}\\ \dot{x}^e_{\text{E–Glc}} &= \alpha\left(x^i_{\text{E–Glc}} - \dot{x}^e_{\text{E–Glc}}\right) + k_1 x^e_E x^e_{\text{Glc}}\\ &\quad - k_{-1} x^e_{\text{E–Glc}}\\ \dot{x}^i_{\text{E–Glc}} &= \alpha\left(x^e_{\text{E–Glc}} - \dot{x}^i_{\text{E–Glc}}\right) - k_3 x^i_{\text{E–Glc}} x^i_{\text{G6P}}\\ &\quad + k_{-3} x^i_{\text{E–Glc–G6P}} + k_2 x^i_{\text{E}} x^i_{\text{Glc}} - k_{-2} x^i_{\text{E–Glc}}\\ \dot{x}^e_{\text{E}} &= \beta\left(x^i_{\text{E}} - x^e_{\text{E}}\right) - k_1 x^e_{\text{E}} x^e_{\text{Glc}} + k_{-1} x^e_{\text{E–Glc}}\\ \dot{x}^i_{\text{E}} &= \beta\left(x^e_{\text{E}} - x^i_{\text{E}}\right) - k_4 x^i_{\text{E}} x^i_{\text{G6P}} + k_{-4} x^i_{\text{E–G6P}}\\ &\quad - k_2 x^i_{\text{E}} x^i_{\text{Glc}} + k_{-2} x^i_{\text{E–Glc}}, \end{aligned}$$

where k_1, k_{-1}, k_2, k_{-2}, k_3, k_{-3}, k_4, k_{-4}, α, and β are the 10 parameters. Note that this system satisfies Assumption 1. Following Schillings et al. (2015) and Gorbach et al. (2017), we used this ODE with initial value $x_0 = \mathbb{1}_M$, time interval $[0., 100.]$ and true parameter $\theta^* = [0.1, 0.0, 0.4, 0.0, 0.3, 0.0, 0.7, 0.0, 0.1, 0.2]$. To generate data by eq. (3), we added Gaussian noise with variance $\sigma^2 = 10^{-5}$ to the corresponding solution at time points $[1., 2., 4., 5., 7., 10., 15., 20., 30., 40., 50., 60., 80., 100.]$. The optimizers and samplers were initialized at $\theta^0 = 1.2\cdot\theta^* = [0.12, 0, 0.48, 0, 0.36, 0, 0.84, 0, 0.12, 0.24]$, and the forward solutions for all likelihood evaluations were computed with step size $h = 0.05$. To create a good initialization, we accepted the first 30 proposals for PHMC and PLMC.

References

Golub, G. and Van Loan, C. *Matrix computations*. Johns Hopkins University Press, 4th edition, 1996.

Gorbach, N. S., Bauer, S., and Buhmann, J. M. Scalable variational inference for dynamical systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kelley, W. and Peterson, A. *The Theory of Differential Equations: Classical and Qualitative*. Springer, 2010.

Kersting, H., Sullivan, T. J., and Hennig, P. Convergence rates of Gaussian ODE filters. *arXiv:1807.09737v2 [math.NA]*, 2019.

Särkkä, S. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

Särkkä, S. and Solin, A. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

Schillings, C., Sunnaker, M., Stelling, J., and Schwab, C. Efficient characterization of parametric uncertainty of complex (bio)chemical networks. *PLOS Computational Biology*, 11, 2015.

Schober, M., Duvenaud, D., and Hennig, P. Probabilistic ODE solvers with Runge–Kutta means. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Schober, M., Särkkä, S., and Hennig, P. A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing*, 29(1):99–122, 2019.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2003.

Tronarp, F., Kersting, H., Särkkä, S., and Hennig, P. Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. *Statistics and Computing*, 29(6):1297–1315, 2019.