Supplementary Material: Learning interpretable continuous-time models of latent stochastic dynamical systems

Lea Duncker [1] **Gergö Bohner** [1] **Julien Boussard** [2] **Maneesh Sahani** [1]

A. Variational Lower Bound

We derive a variational lower bound to the marginal log-likelihood of our model using Jensen's inequality

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \log \int d\boldsymbol{x} d\boldsymbol{f} d\boldsymbol{u}\; p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{u},\boldsymbol{\theta}) p(\boldsymbol{u}|\boldsymbol{\theta})$$
$$\geq \int d\boldsymbol{x} d\boldsymbol{f} d\boldsymbol{u} q(\boldsymbol{x},\boldsymbol{f},\boldsymbol{u}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{u},\boldsymbol{\theta}) p(\boldsymbol{u}|\boldsymbol{\theta})}{q(\boldsymbol{x},\boldsymbol{f},\boldsymbol{u})}$$
$$\stackrel{def}{=} \mathcal{F}^*$$

where $p(\boldsymbol{f}|\boldsymbol{u},\boldsymbol{\theta}) p(\boldsymbol{u}|\boldsymbol{\theta}) = \prod_k p(f_k|\boldsymbol{u}_k,\boldsymbol{\theta}) p(\boldsymbol{u}_k|\boldsymbol{\theta})$. Choosing a factorised variational distribution of the form

$$q(\boldsymbol{x},\boldsymbol{f},\boldsymbol{u}) = q_x(\boldsymbol{x}) \prod_k p(f_k|\boldsymbol{u}_k,\boldsymbol{\theta}) q_u(\boldsymbol{u}_k)$$

we can rewrite the bound as

$$\mathcal{F}^* = \int d\boldsymbol{x} d\boldsymbol{f} d\boldsymbol{u} q(\boldsymbol{x},\boldsymbol{f},\boldsymbol{u}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{f}) \prod_k p(\boldsymbol{u}_k|\boldsymbol{\theta})}{q_x(\boldsymbol{x}) \prod_k q_u(\boldsymbol{u}_k)}$$
$$= \langle \log p(\boldsymbol{y}|\boldsymbol{x}) \rangle_{q_x} - \langle \mathsf{KL}[q_x(\boldsymbol{x}) \| p(\boldsymbol{x}|\boldsymbol{f})] \rangle_{q_f} - \sum_k \mathsf{KL}[q_u(\boldsymbol{u}_k) \| p(\boldsymbol{u}_k)]$$

where

$$q_f(\boldsymbol{f}) = \prod_k \int d\boldsymbol{u}_k p(f_k|\boldsymbol{u}_k,\boldsymbol{\theta}) q_u(\boldsymbol{u}_k)$$

and $q_x(\boldsymbol{x})$ is described by (3) and (4). We can derive the Kullback-Leibler divergence between the distributions over SDE paths $q_x(\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{f})$ by discretising time in steps of Δt. The discretised paths have Markovian structure with

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t,\boldsymbol{f}) = \mathcal{N}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t + \boldsymbol{f}(\boldsymbol{x}_t)\Delta t, \boldsymbol{\Sigma}\Delta t)$$
$$q_x(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t + \boldsymbol{f}_q(\boldsymbol{x}_t)\Delta t, \boldsymbol{\Sigma}\Delta t)$$

We can hence write

$$\mathsf{KL}[q_x(\boldsymbol{x}) \| p(\boldsymbol{x})]$$
$$= \sum_{t=1}^{T-1} \int d\boldsymbol{x}_t q(\boldsymbol{x}_t) \int d\boldsymbol{x}_{t+1} q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \log \frac{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}$$
$$= \frac{1}{2} \sum_{t=1}^{T-1} \Delta t \left\langle (\boldsymbol{f} - \boldsymbol{f}_q)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{f} - \boldsymbol{f}_q) \right\rangle_{qX}$$

Taking the limit as $\Delta t \to 0$, we obtain

$$\mathsf{KL}[q_x(\boldsymbol{x}) \| p(\boldsymbol{x})] = \frac{1}{2} \int_{\mathcal{T}} dt \left\langle (\boldsymbol{f} - \boldsymbol{f}_q)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{f} - \boldsymbol{f}_q) \right\rangle_{q_x}$$

B. Inference Details

B.1. Lagrangian

The full Lagrangian, after applying integration by parts to the constraints in (8), has the form

$$\mathcal{L} = \mathcal{F}^* - \mathcal{C}_1 - \mathcal{C}_2$$
$$\mathcal{C}_1 = \int_{\mathcal{T}} dt \left(\mathsf{Tr}\left[\boldsymbol{\Psi}(\boldsymbol{A}\boldsymbol{S}_x + \boldsymbol{S}_x \boldsymbol{A}^\mathsf{T} - I) - \frac{d\boldsymbol{\Psi}}{dt} \boldsymbol{S}_x \right] \right) + \mathsf{Tr}\left[\boldsymbol{\Psi}(T)\boldsymbol{S}_x(T)\right] - \mathsf{Tr}\left[\boldsymbol{\Psi}(0)\boldsymbol{S}_x(0)\right]$$
$$\mathcal{C}_2 = \int_{\mathcal{T}} dt \left(\boldsymbol{\lambda}^\mathsf{T}(A\boldsymbol{m}_x - \boldsymbol{b}) - \frac{d\boldsymbol{\lambda}}{dt}^\mathsf{T} \boldsymbol{m}_x \right) + \boldsymbol{\lambda}(T)^\mathsf{T} \boldsymbol{m}_x(T) - \boldsymbol{\lambda}(0)^\mathsf{T} \boldsymbol{m}_x(0)$$

For the variational free energy term $\mathcal{F}^*$, we have from before

$$\mathcal{F} = \sum_i \langle \log p(\boldsymbol{y}_i|\boldsymbol{x}_i) \rangle_{q_x} - \mathsf{KL}[q_x(\boldsymbol{x}) \| p(\boldsymbol{x})]$$

and

$$\mathcal{F}^* = \langle \mathcal{F} \rangle_{q_f} - \sum_{k=1}^{K} \mathsf{KL}[q_u(\boldsymbol{u}_k) \| p(\boldsymbol{u}_k|\boldsymbol{\theta})]$$

The Kullback-Leibler divergences can be evaluated as

$$\mathsf{KL}[q(\boldsymbol{u}_k) \| p(\boldsymbol{u}_k|\boldsymbol{\theta})] = \frac{1}{2} \Big(\mathsf{Tr}\left[{\boldsymbol{\Omega}_u}^{-1} \boldsymbol{S}_u^k \right] - M + \log \frac{|\boldsymbol{\Omega}_u|}{|\boldsymbol{S}_u^k|} + (\boldsymbol{\mu}_u^k - \boldsymbol{m}_u^k)^\mathsf{T} {\boldsymbol{\Omega}_u}^{-1} (\boldsymbol{\mu}_u^k - \boldsymbol{m}_u^k) \Big)$$

with

$$\boldsymbol{\Omega}_u = \boldsymbol{K}_{zz} - \tilde{\boldsymbol{K}}_{zs} \tilde{\boldsymbol{K}}_{ss}^{-1} \tilde{\boldsymbol{K}}_{sz}$$
$$\boldsymbol{\mu}_u^k = \tilde{\boldsymbol{K}}_{zs} \tilde{\boldsymbol{K}}_{ss}^{-1} \boldsymbol{v}_k^\theta$$

and

$$\langle \mathsf{KL}[q_x(\boldsymbol{x}) \| p(\boldsymbol{x})] \rangle_{q_f} = \frac{1}{2} \int_0^T dt \, \langle (\boldsymbol{f} - \boldsymbol{f}_q)^\mathsf{T} (\boldsymbol{f} - \boldsymbol{f}_q) \rangle_{q_x q_f}$$

For later convenience, we denote this term as

$$\left\langle \mathsf{KL}[q_x(\boldsymbol{x})\|p(\boldsymbol{x})]\right\rangle_{q_f} = \mathcal{E}(\boldsymbol{m}_x, \boldsymbol{S}_x)$$

Using the identity

$$\langle\langle \boldsymbol{f}\rangle_{q_f}(\boldsymbol{x}-\boldsymbol{m}_x)^{\mathsf{T}}\rangle_{q_x} = \left\langle \frac{\partial \langle \boldsymbol{f}\rangle_{q_f}}{\partial \boldsymbol{x}} \right\rangle_{q_x} \boldsymbol{S}_x$$

the integrand can be evaluated as

$$\begin{aligned}
&\langle (\boldsymbol{f}-\boldsymbol{f}_q)^{\mathsf{T}}(\boldsymbol{f}-\boldsymbol{f}_q)\rangle_{q_x q_f} \\
&= \langle \boldsymbol{f}^{\mathsf{T}}\boldsymbol{f}\rangle_{q_x q_f} + 2\mathsf{Tr}\left[\boldsymbol{A}^{\mathsf{T}}\left\langle \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right\rangle_{q_x q_f} \boldsymbol{S}(t)\right] \\
&\quad + \mathsf{Tr}\left[\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\left(\boldsymbol{S}_x + \boldsymbol{m}_x\boldsymbol{m}_x^{\mathsf{T}}\right)\right] + 2\,\boldsymbol{m}_x^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\langle \boldsymbol{f}\rangle_{q_x q_f} \\
&\quad + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{b} - 2\boldsymbol{b}^{\mathsf{T}}\langle \boldsymbol{f}\rangle - 2\boldsymbol{b}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{m}_x
\end{aligned}$$

For the expected log-likelihood terms, in general, there will be terms that are continuous in $\boldsymbol{x}$, and terms that depend only on evaluations of $\boldsymbol{x}$ at specific locations t_i, which we will denote by ℓ^{cont} and ℓ^{jump}, respectively. We can write

$$\langle \log p(\boldsymbol{y}|\boldsymbol{x})\rangle_{q_x} = \ell^{cont}(\boldsymbol{m}_x, \boldsymbol{S}_x) + \ell^{jump}(\boldsymbol{m}_x, \boldsymbol{S}_x)$$

Thus, the variational free energy can be expressed as

$$\begin{aligned}
\mathcal{F}^* = \ell^{cont}(\boldsymbol{m}_x, \boldsymbol{S}_x) + \ell^{jump}(\boldsymbol{m}_x, \boldsymbol{S}_x) - \mathcal{E}(\boldsymbol{m}_x, \boldsymbol{S}_x) \\
- \sum_{k=1}^{K} \mathsf{KL}[q_u(\boldsymbol{u}_k)\|p(\boldsymbol{u}_k|\boldsymbol{\theta})]
\end{aligned}$$

B.1.1. Example: Gaussian likelihood

In the case of a Gaussian likelihood, there is no continuous term in the likelihood:

$$\begin{aligned}
\ell^{cont} &= 0 \\
\ell^{jump} &= \sum_i \int_{\mathcal{T}_0}^{\mathcal{T}_{end}} dt\delta(t-t_i)\big(\boldsymbol{m}_x(t)^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\Gamma^{-1}(\boldsymbol{y}_t - \boldsymbol{d}) \\
&\quad - \frac{1}{2}\mathsf{Tr}\left[\boldsymbol{C}^{\mathsf{T}}\Gamma^{-1}\boldsymbol{C}\sum_i \left(\boldsymbol{S}_x(t) + \boldsymbol{m}_x(t)\boldsymbol{m}_x(t)^{\mathsf{T}}\right)\right]\big)
\end{aligned}$$

B.1.2. Example: Multivariate Poisson Process likelihood

In the case of a multivariate Poisson Process, with $g(\cdot) = \exp(\cdot)$ and observed event times $t_1^{(n)}, \dots t_{\phi(n)}^{(n)}$for the nth output dimension:

$$\begin{aligned}
\ell^{cont} &= -\sum_n \int_{\mathcal{T}_0}^{\mathcal{T}_{end}} \exp\left(\boldsymbol{c}_n^{\mathsf{T}}\boldsymbol{m}_x + \frac{1}{2}\boldsymbol{c}_n^{\mathsf{T}}\boldsymbol{S}_x\boldsymbol{c}_n\right) dt \\
\ell^{jump} &= \sum_{n=1}^{N}\sum_{i=1}^{\phi(n)} \int_{\mathcal{T}_0}^{\mathcal{T}_{end}} \left(\boldsymbol{c}_n^{\mathsf{T}}\boldsymbol{m}_x(t) + d_n\right)\delta(t - t_i^{(n)})dt
\end{aligned}$$

B.2. Symmetric variations in $\boldsymbol{S}_x$

To arrive at the fixed point equations given in the main paper, we need to take variational derivatives of the Lagrangian with respect to $\boldsymbol{m}_x$ and $\boldsymbol{S}_x$. In contrast to Archambeau et al. (2007), we take the symmetric variations in $\boldsymbol{S}_x$ into account. Also note that the Lagrange multiplier $\boldsymbol{\Psi}$ is symmetric. We can write

$$\frac{\partial \mathcal{C}_1}{\partial \boldsymbol{S}_x} = \left(\boldsymbol{\Psi}\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Psi} - \frac{d\boldsymbol{\Psi}}{dt}\right) \odot \tilde{\mathbb{P}}$$

where $\odot$ denotes the elementwise Hadamard product and $\tilde{\mathbb{P}}_{ij} = 2$ for $i \neq j$ and 1 otherwise. Differentiating the entire Lagrangian with respect to the symmetric matrix $\boldsymbol{S}_x$ and setting to zero we get

$$0 = \frac{\partial \mathcal{F}^*}{\partial \boldsymbol{S}_x} \odot \mathbb{P} - \boldsymbol{\Psi}\boldsymbol{A} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Psi} + \frac{d\boldsymbol{\Psi}}{dt}$$

matching the equation given in the main text with $\mathbb{P}_{ij} = \frac{1}{2}$ if $i \neq j$ and 1 otherwise. Note that the derivatives of the free energy with respect to $\boldsymbol{S}_x$ will also need to take into account the symmetry of the covariance matrix. The derivations for (20)-(22) follow those of Archambeau et al. (2007).

B.3. Expected values of dynamics

The inference algorithm requires evaluating several expectations with respect to q_x and q_f. Let $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{u}_1 & \dots & \boldsymbol{u}_K\end{bmatrix}$ and $\langle \boldsymbol{U}\rangle_{q_u} = \boldsymbol{M}_u$, such that we can define $(M+L+LK) \times K$ matrices stacking all inducing variables, zero function values, and Jacobians as

$$\boldsymbol{U}_{u,f_s,J} = \begin{bmatrix}\boldsymbol{U}_u \\ \boldsymbol{0} \\ \boldsymbol{J}_s^{(1)} \\ \vdots \\ \boldsymbol{J}_s^{(L)}\end{bmatrix}, \quad \langle \boldsymbol{U}_{u,f_s,J}\rangle_{q_u} = \boldsymbol{M}_{u,f_s,J} = \begin{bmatrix}\boldsymbol{M}_u \\ \boldsymbol{0} \\ \boldsymbol{J}_s^{(1)} \\ \vdots \\ \boldsymbol{J}_s^{(L)}\end{bmatrix}$$

The required expectations can then be evaluated as

$$\langle \boldsymbol{f}(\boldsymbol{x})\rangle_{q_x q_f}^{\mathsf{T}} = \left\langle \boldsymbol{a}_z^{\theta}(\boldsymbol{x})\right\rangle_{q_x} \boldsymbol{M}_{u,f_s,J}$$

$$\left\langle \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right\rangle_{q_x q_f}^{\mathsf{T}} = \left\langle \nabla_x \boldsymbol{a}_z^{\theta}(\boldsymbol{x})\right\rangle_{q_x} \boldsymbol{M}_{u,f_s,J}$$

$$\begin{aligned}
&\left\langle \boldsymbol{f}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{f}(\boldsymbol{x})\right\rangle_{q_x q_f} = \sum_k \left\langle f_k^2(\boldsymbol{x})\right\rangle_{q_x q_f} = \kappa(\boldsymbol{x}, \boldsymbol{x}') \\
&+ \mathsf{Tr}\left[\left(\left\langle \boldsymbol{U}_{u,f_s,J}\boldsymbol{U}_{u,f_s,J}^{\mathsf{T}}\right\rangle_{q_u} - \mathsf{K}_{zz}^{\theta}\right)\langle \boldsymbol{a}_z^{\theta}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{a}_z^{\theta}(\boldsymbol{x})\rangle_{q_x}\right]
\end{aligned}$$

The above expressions still involve computing expectations of covariance functions and their derivatives, which can be computed analytically for choices such as the exponentiated quadratic covariance function.

B.4. Inference algorithm

The full inference algorithm involves solving a set of ODEs forward and backward in time, which we do using the forward Euler method. We provide the full approach in Algorithm 1, where the subscript r denotes the evaluation of the functions at the rth point of the time grid between $\mathcal{T}_0$ and $\mathcal{T}_{end}$ taking steps of size Δt. Note that the derivatives of the terms in ℓ^{jump} will need to be discretized appropriately as well. Using the same time-grid as was used for solving the ODEs, the delta-functions will contribute a factor of $\frac{1}{\Delta t}$, such that the Δt terms cancel in the update written in Algorithm 1.

C. Learning Details

C.1. Conditioned Sparse Gaussian Process dynamics

The only term in the variational free energy that depends on the parameters in $\boldsymbol{f}$ are the KL-divergence between the continuous-time processes and the KL-divergence relating to the inducing points for $\boldsymbol{f}$.

C.1.1. Inducing point covariances

Collecting the terms that contain $\boldsymbol{S}_u^k$ we have

$$\frac{\partial}{\partial \boldsymbol{S}_u^k}\mathsf{KL}[q(\boldsymbol{u}_k)\|p(\boldsymbol{u}_k|\boldsymbol{\theta})] = \frac{1}{2}{\boldsymbol{\Omega}_u}^{-1} - \frac{1}{2}{\boldsymbol{S}_u^k}^{-1}$$

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial \boldsymbol{S}_u^k} &= \frac{1}{2}\int_{\mathcal{T}} dt \frac{\partial}{\partial \boldsymbol{S}_u^k}\mathsf{Tr}\left[\begin{bmatrix}\boldsymbol{S}_u^k & 0\\ 0 & 0\end{bmatrix}\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}\right]\\ &= \frac{1}{2}\int_{\mathcal{T}} dt[\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}]_{:M,:M}\end{aligned}$$

where the last line selects the first $M \times M$ block from $\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}$. We hence obtain the closed form update

$$\boldsymbol{S}_u^k = \left({\boldsymbol{\Omega}_u}^{-1} + \int_{\mathcal{T}} dt[\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}]_{:M,:M}\right)^{-1}$$

C.1.2. Inducing points and Jacobians

To find the update efficiently, let $\boldsymbol{J}_k = [\boldsymbol{J}_{k,:}^{(1)}, \ldots, \boldsymbol{J}_{k,:}^{(L)}]^\top$ so that we can write

$$\boldsymbol{\mu}_u^k = \tilde{\boldsymbol{K}}_{zs}\tilde{\boldsymbol{K}}_{ss}^{-1}\boldsymbol{v}_k^\theta = \tilde{\boldsymbol{K}}_{zs}\tilde{\boldsymbol{K}}_{ss}^{-1}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{J}_k\end{bmatrix} = \boldsymbol{G}\boldsymbol{J}_k$$

We can rewrite the quadratic terms in the Kullback-Leibler divergences of the inducing points as

$$\begin{aligned}&\sum_k (\boldsymbol{\mu}_u^k - \boldsymbol{m}_u^k)^\top {\boldsymbol{\Omega}_u}^{-1}(\boldsymbol{\mu}_u^k - \boldsymbol{m}_u^k)\\ &\quad= \sum_k \begin{bmatrix}\boldsymbol{m}_u^k\\ \boldsymbol{J}_k\end{bmatrix}^\top \begin{bmatrix}{\boldsymbol{\Omega}_u}^{-1} & -{\boldsymbol{\Omega}_u}^{-1}\boldsymbol{G}\\ -\boldsymbol{G}^\top{\boldsymbol{\Omega}_u}^{-1} & \boldsymbol{G}^\top{\boldsymbol{\Omega}_u}^{-1}\boldsymbol{G}\end{bmatrix}\begin{bmatrix}\boldsymbol{m}_u^k\\ \boldsymbol{J}_k\end{bmatrix}\\ &\quad= \mathsf{Tr}\left[\boldsymbol{M}_{u,J}^\top\tilde{\boldsymbol{\Omega}}\boldsymbol{M}_{u,J}\right]\end{aligned}$$

with $\boldsymbol{M}_{u,J} = \begin{bmatrix}\boldsymbol{m}_u^1 & \ldots & \boldsymbol{m}_u^K\\ \boldsymbol{J}_1 & \ldots & \boldsymbol{J}_K\end{bmatrix}$ and derivative

$$\frac{\partial}{\partial \boldsymbol{M}_{u,J}}\sum_k \mathsf{KL}[q(\boldsymbol{u}_k)\|p(\boldsymbol{u}_k|\boldsymbol{\theta})] = \tilde{\boldsymbol{\Omega}}\boldsymbol{M}_{u,J}$$

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial \boldsymbol{M}_{u,J}} &= \int_{\mathcal{T}} dt\left[\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}\right]_{[i,i]}\boldsymbol{M}_{u,J}\\ &\quad+ \int_{\mathcal{T}} dt\left[\left\langle \nabla_x \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,i]}^\top \boldsymbol{S}_x\boldsymbol{A}^\top\\ &\quad- \int_{\mathcal{T}} dt\left[\left\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,i]}^\top(-\boldsymbol{A}\boldsymbol{m}_x + \boldsymbol{b})^\top\end{aligned}$$

Putting all terms together, we obtain the update

$$\boldsymbol{M}_{u,J} = \boldsymbol{B}_1^{-1}(\boldsymbol{B}_2 - \boldsymbol{B}_3)$$

with

$$\begin{aligned}\boldsymbol{B}_1 &= \left(\tilde{\boldsymbol{\Omega}} + \int_{\mathcal{T}} dt\left[\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}\right]_{[uj,uj]}\right)\\ \boldsymbol{B}_2 &= \int_{\mathcal{T}} dt\left[\left\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,uj]}^\top \langle \boldsymbol{f}_q\rangle_{q_x}^\top\\ \boldsymbol{B}_3 &= \int_{\mathcal{T}} dt\left[\left\langle \nabla_x \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,uj]}^\top \boldsymbol{S}_x\boldsymbol{A}^\top\end{aligned}$$

and we have defined an indexing operation where $[X]_{[uj,uj]}$ selects the first $M \times M$ and last $LK \times LK$ block of X and $[X]_{[:,uj]}$ selects the first M and last LK columns of X. Hence, this selects the appropriate block matrices for the updates. The one-dimensional integrals can be computed efficiently using Gauss-Legendre quadrature.

C.2. Sparse Gaussian Process dynamics

Similarly, closed form updates are available in the simpler case, when $\boldsymbol{f}$ is modelled by a classic sparse Gaussian Process, i.e. using inducing points without the additional conditioning on fixed points and Jacobians.

$$\begin{aligned}\boldsymbol{S}_u^k &= \boldsymbol{K}_{zz}\left(\boldsymbol{K}_{zz} + \int_{\mathcal{T}} dt\,\langle \boldsymbol{\kappa}(\boldsymbol{Z},\boldsymbol{x})\boldsymbol{\kappa}(\boldsymbol{x},\boldsymbol{Z})\rangle_{q_x}\right)^{-1}\boldsymbol{K}_{zz}\\ \boldsymbol{M}_u &= \boldsymbol{S}_u^k\boldsymbol{K}_{zz}^{-1}\left(\int_{\mathcal{T}} dt\boldsymbol{\Phi}_1\boldsymbol{f}_q^\top - \int_{\mathcal{T}} dt\boldsymbol{\Phi}_{d1}\boldsymbol{S}_x\boldsymbol{A}^\top\right)\end{aligned}$$

Where $\boldsymbol{\Phi}_1 = \langle k(\boldsymbol{x},\boldsymbol{Z})\rangle_{q_x}$ and $\boldsymbol{\Phi}_{d1} = \left\langle \frac{\partial}{\partial \boldsymbol{x}}k(\boldsymbol{x},\boldsymbol{Z})\right\rangle_{q_x}$.

C.3. Linear dynamics

Our modelling framework also easily extends to other parameterisation of $\boldsymbol{f}$. For example, in a continuous-time linear dynamical system with $\boldsymbol{f}(\boldsymbol{x}) = -\tilde{\boldsymbol{A}}\boldsymbol{x} + \tilde{\boldsymbol{b}}$ direct minimisation of the KL-divergence between the continuous-time

Algorithm 1 Inference algorithm

Input: data $\{y_i, t_i\}_{i=1}^T$, $\boldsymbol{m}_{x,0}$, $\boldsymbol{S}_{x,0}$, $q_f(\boldsymbol{f})$, Δt, $\mathcal{T}_0$, $\mathcal{T}_{end}$
Initialize $\boldsymbol{A}(t), \boldsymbol{b}(t)$
$R = \frac{\mathcal{T}_0 - \mathcal{T}_{end}}{\Delta t}$
repeat
  for $r = 0$ **to** $R-1$ **do**
    $\boldsymbol{m}_{x,r+1} \leftarrow \boldsymbol{m}_{x,r} - \Delta t\left(\boldsymbol{A}_r \boldsymbol{m}_{x,r} - \boldsymbol{b}_r\right)$
    $\boldsymbol{S}_{x,r+1} \leftarrow \boldsymbol{S}_{x,r} - \Delta t\left(\boldsymbol{A}_r \boldsymbol{S}_{x,r} + \boldsymbol{S}_{x,r} \boldsymbol{A}_r{}^\top - I\right)$
  end for
  for $r = R$ **to** 1 **do**
    $\boldsymbol{\lambda}_{r-1} \leftarrow \boldsymbol{\lambda}_r - \Delta t\left(\boldsymbol{A}_r{}^\top \boldsymbol{\lambda}_r + \left(\frac{\partial \ell^{cont}}{\partial \boldsymbol{m}_x} - \frac{\partial \mathcal{E}}{\partial \boldsymbol{m}_x}\right)|_{t=r\Delta t}\right) - \Delta t \frac{\partial \ell^{jump}}{\partial \boldsymbol{m}_x}\Big|_{t=(r-1)\Delta t}$
    $\boldsymbol{\Psi}_{r-1} \leftarrow \boldsymbol{\Psi}_r - \Delta t\left(\boldsymbol{A}_r{}^\top \boldsymbol{\Psi}_r + \boldsymbol{\Psi}_r \boldsymbol{A}_r + \mathbb{P} \odot \left(\frac{\partial \ell^{cont}}{\partial \boldsymbol{S}_x} - \frac{\partial \mathcal{E}}{\partial \boldsymbol{S}_x}\right)|_{t=r\Delta t}\right) - \Delta t \mathbb{P} \odot \frac{\partial \ell^{jump}}{\partial \boldsymbol{S}_x}\Big|_{t=(r-1)\Delta t}$
  end for
  $\boldsymbol{A} = \left\langle \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right\rangle_{q_x q_f} + 2\boldsymbol{\Psi}$
  $\boldsymbol{b} = \langle \boldsymbol{f}(\boldsymbol{x}) \rangle_{q_x q_f} + \boldsymbol{A}\boldsymbol{m}_x - \boldsymbol{\lambda}$
until convergence in $\mathcal{F}^*$
return: $\{\boldsymbol{A}_r, \boldsymbol{b}_r, \boldsymbol{\lambda}_r, \boldsymbol{\Psi}_r, \boldsymbol{m}_{x,r}, \boldsymbol{S}_{x,r}\}_{r=1}^R$

processes leads to the closed form updates

$$\tilde{\boldsymbol{A}} = \left(\int_{\mathcal{T}} dt \left(\boldsymbol{b}\langle \boldsymbol{x} \rangle^\top - \langle \boldsymbol{f}_q(\boldsymbol{x})\boldsymbol{x}^\top \rangle\right)\right)\left(\int_{\mathcal{T}} dt \langle \boldsymbol{x}\boldsymbol{x}^\top \rangle\right)^{-1}$$
$$\tilde{\boldsymbol{b}} = \frac{1}{T}\int_{\mathcal{T}} dt \left(\langle \boldsymbol{f}_q(\boldsymbol{x}) \rangle + \boldsymbol{A}\langle \boldsymbol{x} \rangle\right)$$

reminiscent of the update equations for the generative parameters of a discrete-time Linear Dynamical System.

C.4. Output mapping

We consider an observation model of the form

$$\boldsymbol{y}(t_i) = \boldsymbol{C}\boldsymbol{x}(t_i) + \boldsymbol{d} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\epsilon|0, \Gamma)$. Dropping all terms that are constant in $\boldsymbol{C}, \boldsymbol{d}$ from the expression for the variational free energy, we have

$$\mathcal{F}^* = -\frac{1}{2}\sum_t \left\langle (\boldsymbol{y}_t - \boldsymbol{C}\boldsymbol{x}_t - \boldsymbol{d})^\top \Gamma^{-1} (\boldsymbol{y}_t - \boldsymbol{C}\boldsymbol{x}_t - \boldsymbol{d}) \right\rangle_{q_x}$$

Differentiating and setting to zero gives

$$\boldsymbol{C}^{new} = \left(\sum_t (\boldsymbol{y}_t - \boldsymbol{d})\boldsymbol{m}_t^\top\right)\left(\sum_t (\boldsymbol{S}_{x,t} + \boldsymbol{m}_{x,t}\boldsymbol{m}_{x,t}^\top)\right)^{-1}$$
$$\boldsymbol{d}^{new} = \frac{1}{T}\sum_t \left(\boldsymbol{y}_t - \boldsymbol{C}^{new}\boldsymbol{m}_{x,t}\right)$$

D. Chemical reaction dynamics

The dynamical system used to generate the data in section 5.4 is of the form

$$\frac{d\lfloor \mathrm{I}^- \rfloor_A}{dt} = \left(k_a \lfloor \mathrm{I}^- \rfloor_A + k_b \lfloor \mathrm{I}^- \rfloor_A^2\right)\left(S_0 - \lfloor \mathrm{I}^- \rfloor_A\right) + \frac{F_1 \lfloor \mathrm{I}^- \rfloor_0}{V_A} - \frac{(F_3 + F_4)\lfloor \mathrm{I}^- \rfloor_A}{V_A} + \frac{F_4 \lfloor \mathrm{I}^- \rfloor_D}{V_A}$$
$$\frac{d\lfloor \mathrm{I}^- \rfloor_D}{dt} = \left(k_a \lfloor \mathrm{I}^- \rfloor_D + k_b \lfloor \mathrm{I}^- \rfloor_D^2\right)\left(S_0 - \lfloor \mathrm{I}^- \rfloor_D\right) + \frac{F_4 \lfloor \mathrm{I}^- \rfloor_A}{V_D} - \frac{F_4 \lfloor \mathrm{I}^- \rfloor_D}{V_D}$$

To generate the simulations, we use the parameter settings

$$\lfloor \mathrm{I}^- \rfloor_0 = 4.4 \times 10^{-5} \qquad k_0 = 2.7 \times 10^{-3}$$
$$V_A = 4 \times 10^1 \qquad F_4 = 3.25 \times 10^{-3}$$
$$V_D = 1 \qquad F_3 = k_0 V_a$$
$$k_a = 2.1425 \times 10^{-1} \qquad F_1 = \frac{1}{2}F_3$$
$$k_b = 2.1425 \times 10^4 \qquad F_2 = \frac{1}{2}F_3$$
$$S_0 = \frac{1}{2}\left(\lfloor \mathrm{I}^- \rfloor_0 + 1.42 \times 10^{-3}\right)$$

References

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. Gaussian process approximations of stochastic differential equations. *Journal of machine learning research*, 1:1–16, 2007.