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

Lea Duncker 1  Gergő 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

\[
\log p(y|\theta) = \log \int dx \, p(y|x)p(x|f)p(f|u, \theta)p(u|\theta) \\
\geq \int dx \, \int df \, \int du \, q(x, f, u) \log \frac{p(y|x)p(x|f)p(f|u, \theta)p(u|\theta)}{q(x, f, u)} \\
defF^* = \frac{1}{T} \sum_{t=1}^{T} \log \frac{p(y|x)p(x|f)}{q(x, f, u)}
\]

where \( p(f|u, \theta)p(u|\theta) = \prod_k p(f_k|u_k, \theta)p(u_k|\theta) \). Choosing a factorised variational distribution of the form

\[
q(x, f, u) = q_x(x) \prod_k p(f_k|u_k, \theta)q_u(u_k)
\]

we can rewrite the bound as

\[
\mathcal{F}^* = \int dx \, \int df \, \int du \, q(x, f, u) \log \frac{p(y|x)p(x|f)}{q_x(x) \prod_k p(u_k|\theta)q_u(u_k)} \\
= \langle \log p(y|x) \rangle_{q_x} - \langle \mathbb{KL}[q_x(x)||p(x|f)] \rangle_{q_f} \\
- \sum_k \mathbb{KL}[q_u(u_k)||p(u_k)]
\]

where

\[
q_f(f) = \prod_k \int du_k p(f_k|u_k, \theta)q_u(u_k)
\]

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

\[
p(x_{t+1}|x_t, f) = \mathcal{N}(x_{t+1}|x_t + f(x_t)\Delta t, \Sigma \Delta t) \\
qu_x(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}|x_t + f_q(x_t)\Delta t, \Sigma \Delta t)
\]

We can hence write

\[
\mathbb{KL}[q_x(x)||p(x)] = \sum_{t=1}^{T-1} \int dx_t q(x_t) \int dx_{t+1} q(x_{t+1}|x_t) \log \frac{q(x_{t+1}|x_t)}{p(x_{t+1}|x_t)} \\
= \frac{1}{2} \sum_{t=1}^{T-1} \Delta t \langle (f - f_q)^T \Sigma^{-1} (f - f_q) \rangle_{q_x}
\]

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

\[
\mathbb{KL}[q_x(x)||p(x)] = \frac{1}{2} \int_T dt \langle (f - f_q)^T \Sigma^{-1} (f - f_q) \rangle_{q_x}
\]

B. Inference Details

B.1. Lagrangian

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

\[
\mathcal{L} = \mathcal{F}^* - C_1 - C_2
\]

\[
C_1 = \int_T dt \left[ \mathbb{Tr} \left[ (\Theta (M_x - \delta) - \frac{d\Psi}{dt} S_x \right) \right] \\
+ \mathbb{Tr} [\Psi(T) S_x(T)] - \mathbb{Tr} [\Psi(0) S_x(0)]
\]

\[
C_2 = \int_T dt \left[ (\mathbf{M}^T \mathbf{M}) - \frac{d\mathbf{M}^T}{dt} \mathbf{M} \right] \\
+ \mathbf{M}(T)^T M_x(0)
\]

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

\[
\mathcal{F} = \sum_i \langle \log p(y_i|x_i) \rangle_{q_x} - \mathbb{KL}[q_x(x)||p(x)]
\]

and

\[
\mathcal{F}^* = \langle \mathcal{F} \rangle_{q_f} - \sum_{k=1}^{K} \mathbb{KL}[q_u(u_k)||p(u_k|\theta)]
\]

The Kullback-Leibler divergences can be evaluated as

\[
\mathbb{KL}[q_u(u_k)||p(u_k|\theta)] = \frac{1}{2} \left[ \mathbb{Tr} \left[ (\mathbf{M}_u^{-1} - M + \log |\mathbf{M}_u| \right] \\
+ (\mathbf{M}_u - \mathbf{M}_u^{-1} M \mathbf{M}_u) \right]
\]

with

\[
\mathbf{M}_u = \mathbf{K}_{zz} - \mathbf{K}_{zs} \mathbf{K}_{ss}^{-1} \mathbf{K}_{sz} \\
\mathbf{M}_u = \mathbf{K}_{zz} \mathbf{K}_{ss}^{-1} \mathbf{v}_k
\]

and

\[
\langle \mathbb{KL}[q_x(x)||p(x)] \rangle_{q_f} = \frac{1}{2} \int_T dt \langle (f - f_q)^T \Sigma^{-1} (f - f_q) \rangle_{q_x q_f}
\]
For later convenience, we denote this term as
\[
\langle KL[q_x(x)||p(x)] \rangle_q = \mathcal{E}(m_x, S_x)
\]
Using the identity
\[
\langle (f)_{q_f}(x - m_x)^T \rangle_q = \left\langle \frac{\partial (f)_{q_f}}{\partial x} \right\rangle_q S_x
\]
the integrand can be evaluated as
\[
\langle (f - f_q)^T (f - f_q) \rangle_{q_f, q_f} = \langle f^T f \rangle_{q_f, q_f} + 2 \operatorname{Tr} \left[ A^T \left\langle \frac{\partial f}{\partial x} \right\rangle_q S(t) \right]
\]
\[
+ \operatorname{Tr} \left[ A^T A (S_x + m_x m_x^T) \right] + 2 m_x^T A^T (f)_{q_f, q_f} + b^T b - 2 b^T (f - 2 b^T A m_x)
\]

For the expected log-likelihood terms, in general, there will be terms that are continuous in \( x \), and terms that depend only on evaluations of \( x \) at specific locations \( t_i \), which we will denote by \( \ell_{\text{cont}} \) and \( \ell_{\text{jump}} \), respectively. We can write
\[
\langle \log p(y|x) \rangle_{q_x} = \ell_{\text{cont}}(m_x, S_x) + \ell_{\text{jump}}(m_x, S_x)
\]
Thus, the variational free energy can be expressed as
\[
\mathcal{F}^* = \ell_{\text{cont}}(m_x, S_x) + \ell_{\text{jump}}(m_x, S_x) - \mathcal{E}(m_x, S_x)
\]
\[- \sum_{k=1}^{K} \langle KL[q_u(u_k)||p(u_k|\theta)] \rangle_{u_k}
\]

**B.1.1. Example: Gaussian likelihood**

In the case of a Gaussian likelihood, there is no continuous term in the likelihood:
\[
\ell_{\text{cont}} = 0
\]
\[
\ell_{\text{jump}} = \sum_i \int_{T_0}^{T_{\text{end}}} dt \delta(t - t_i)(m_x(t)^T C^T \Gamma^{-1} (y_t - d)\right) - \frac{1}{2} \operatorname{Tr} \left[ C^T \Gamma^{-1} C \sum_i (S_x(t) + m_x(t) m_x(t)^T) \right]
\]

**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 \( t_1^{(n)}, \ldots, t_{\phi(n)}^{(n)} \) for the \( n \)th output dimension:
\[
\ell_{\text{cont}} = - \sum_n \int_{T_0}^{T_{\text{end}}} \exp \left( c_n^T m_x + \frac{1}{2} c_n^T S_x c_n \right) dt
\]
\[
\ell_{\text{jump}} = \sum_{n=1}^{N} \sum_{i=1}^{\phi(n)} \int_{T_0}^{T_{\text{end}}} (c_n^T m_x(t) + d_n) \delta(t - t_i^{(n)}) dt
\]

**B.2. Symmetric variations in \( S_x \)**

To arrive at the fixed point equations given in the main paper, we need to take variational derivatives of the Lagrangian with respect to \( m_x \) and \( S_x \). In contrast to Archambeau et al. (2007), we take the symmetric variations in \( S_x \) into account. Also note that the Lagrange multiplier \( \Psi \) is symmetric. We can write
\[
\frac{\partial \mathcal{C}_1}{\partial S_x} = (\Psi A + A^T \Psi - \frac{\partial \Psi}{\partial t}) \odot \ddot{p}
\]
where \( \odot \) denotes the elementwise Hadamard product and \( \ddot{p}_{ij} = 2 \) for \( i \neq j \) and 1 otherwise. Differentiating the entire Lagrangian with respect to the symmetric matrix \( S_x \) and setting to zero we get
\[
0 = \frac{\partial \mathcal{F}^*}{\partial S_x} \odot \ddot{p} - \Psi A - A^T \Psi + \frac{\partial \Psi}{\partial t}
\]
matching the equation given in the main text with \( \ddot{p}_{ij} = \frac{1}{2} \) if \( i \neq j \) and 1 otherwise. Note that the derivatives of the free energy with respect to \( S_x \) 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 \( U = [u_1 \ldots u_K] \) and \( \langle U \rangle_{q_u} = M_u \), such that we can define \((M + L + LK) \times K \) matrices stacking all inducing variables, zero function values, and Jacobians as
\[
U_{u,f,i} = \begin{bmatrix} U_u \\ 0 \\ \vdots \\ J^{(i)} \end{bmatrix}, \quad \langle U_{u,f,i} \rangle_{q_u} = M_{u,f,i} = \begin{bmatrix} M_u \\ 0 \\ \vdots \\ J^{(i)} \end{bmatrix}
\]
The required expectations can then be evaluated as
\[
\langle f(x) \rangle_{q_x,q_f} = \langle a^\theta(x) \rangle_{q_x} M_{u,f,i}
\]
\[
\langle \frac{\partial f(x)}{\partial x} \rangle_{q_x,q_f} = \langle \nabla_x a^\theta(x) \rangle_{q_x} M_{u,f,i}
\]
\[
\langle f(x)^T f(x) \rangle_{q_x,q_f} = \sum_k \langle f_k^\theta(x) \rangle_{q_x,q_f} = \kappa(x,x')
\]
\[
+ \operatorname{Tr} \left[ \left( \langle U_{u,f,i} U_{u,f,i}^T \rangle_{q_u} - K_{xx}^\theta \right) \langle a^\theta(x) a^\theta(x) \rangle_{q_x} \right]
\]
The above expressions still involve computing expectations of covariance functions and their derivatives, which can be computed analytically for choices such as the exponentially 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 \( r \)th point of the time grid between \( T_0 \) and \( T_{end} \) taking steps of size \( \Delta t \). Note that the derivatives of the terms in \( \mathcal{E}^{jump} \) will need to be discretized appropriately. 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 \( \Delta t \) 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 \( f \) are the KL-divergence between the continuous-time processes and the KL-divergence relating to the inducing points for \( f \).

C.1.1. Inducing point covariances

Collecting the terms that contain \( S_k^x \) we have

\[
\frac{\partial}{\partial S_k^x} \mathcal{E} = \frac{1}{2} \int_T dt \frac{\partial}{\partial S_k^x} \text{Tr} \left[ \begin{bmatrix} S_k^0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a^y_k(x)^T a^y_k(x) \\ a^y_k(x)^T a^y_k(x) \end{bmatrix} q_x \right] = \frac{1}{2} \int_T dt \langle a^y_k(x)^T a^y_k(x) \rangle_{q_x:M:M}.
\]

where the last line selects the first \( M \times M \) block from \( \langle a^y_k(x)^T a^y_k(x) \rangle_{q_x:M:M} \). We hence obtain the closed form update

\[
S_k^x = \left( \Omega_k^{-1} + \int_T dt \langle a^y_k(x)^T a^y_k(x) \rangle_{q_x:M:M} \right)^{-1}
\]

C.1.2. Inducing points and Jacobians

To find the update efficiently, let \( J_k = [J_{k;1}, \ldots, J_{k:L}]^T \) so that we can write

\[
\mu_k = K_{zz} K_{ss}^{-1} v_k = K_{zz} K_{ss}^{-1} \begin{bmatrix} 0 \\ J_k \end{bmatrix} = GJ_k
\]

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

\[
\sum_k (\mu_k - m_k^k)^T \Omega_k^{-1} (\mu_k - m_k^k) = \sum_k \begin{bmatrix} m_k^k \\ J_k \end{bmatrix}^T \begin{bmatrix} \Omega_k^{-1} & - \Omega_k^{-1} G \\ - G^T \Omega_k^{-1} & G^T \Omega_k^{-1} G \end{bmatrix} \begin{bmatrix} m_k^k \\ J_k \end{bmatrix} = \text{Tr} \left[ (M_{u,J}^T \Omega_{u,J} M_{u,J}) \right]
\]

with \( M_{u,J} = \begin{bmatrix} m_1^k & \cdots & m_K^k \\ J_1 & \cdots & J_K \end{bmatrix} \) and derivative

\[
\frac{\partial}{\partial M_{u,J}} \sum_k \mathcal{KL}[q(u_k)||p(u_k|\theta)] = \hat{\Omega} M_{u,J}
\]

\[
\frac{\partial \mathcal{E}}{\partial M_{u,J}} = \int_T dt \left[ \begin{bmatrix} a^y_k(x)^T a^y_k(x) \\ a^y_k(x)^T a^y_k(x) \end{bmatrix}_{[i,i]} M_{u,J} + \int_T dt \left[ \begin{bmatrix} \nabla_x a^y_k(x) \\ \nabla_x a^y_k(x) \end{bmatrix}_{[i,i]} S_x A^T \right. \\
- \left. \int_T dt \langle a^y_k(x) \rangle_{q_x[M:M]} (-A m_x + b)^T \right]
\]

Putting all terms together, we obtain the update

\[
M_{u,J} = B_1^{-1} (B_2 - B_3)
\]

with

\[
B_1 = \left( \hat{\Omega} + \int_T dt \left[ \begin{bmatrix} a^y_k(x)^T a^y_k(x) \end{bmatrix}_{[u,u]} \right] \right)
\]

\[
B_2 = \int_T dt \left[ \langle a^y_k(x) \rangle_{q_x[M:M]} (f \cdot q_x)^T \right]
\]

\[
B_3 = \int_T dt \left[ \langle \nabla_x a^y_k(x) \rangle_{q_x[M:M]} S_x A^T \right]
\]

and we have defined an indexing operation where \( [X]_{[u,u]} \) selects the first \( M \times M \) and last \( LK \times LK \) block of \( X \) and \( [X]_{[i,j]} \) 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 \( f \) is modelled by a classic sparse Gaussian Process, i.e. using inducing points without the additional conditioning on fixed points and Jacobians. We have

\[
S_k^x = K_{zz} \left( K_{zz} + \int_T dt \langle k(Z,x)k(x,Z) \rangle_{q_x[M:M]} \right)^{-1}
\]

\[
M_u = S_k^{u} K_{zz}^{-1} \left( \int_T dt \Phi_1 f_q^T - \int_T dt \Phi_1 S_x A^T \right)
\]

where \( \Phi_1 = \langle k(x,Z) \rangle_{q_z} \) and \( \Phi_{d1} = \langle \frac{\partial}{\partial x} k(x,Z) \rangle_{q_z} \).

C.3. Linear dynamics

Our modelling framework also easily extends to other parameterisation of \( f \). For example, in a continuous-time linear dynamical system with \( f(x) = -Ax + b \) direct minimisation of the KL-divergence between the continuous-time...
Algorithm 1 Inference algorithm

**Input:** data \( \{y_i, t_i\}_{i=1}^T, m_{x0}, S_{x0}, q_f(f), \Delta t, T_0, T_{\text{end}} \)

Initialize \( A(t), b(t) \)
\( R = T_0 - T_{\text{end}} \)

repeat
  for \( r = 0 \) to \( R - 1 \) do
    \( m_{x,r+1} \leftarrow m_{x,r} - \Delta t (A_r m_{x,r} - b_r) \)
    \( S_{x,r+1} \leftarrow S_{x,r} - \Delta t (A_r S_{x,r} + S_{x,r} A_r^\top - I) \)
  end for
  for \( r = R \) to \( 1 \) do
    \( \lambda_{r-1} \leftarrow \lambda_r - \Delta t \left( A_r \lambda_r + \left( \frac{\partial \ell}{\partial \lambda_r} \right) \right) \)
    \( \Psi_{r-1} \leftarrow \Psi_r - \Delta t \left( A_r \Psi_r + \Psi_r A_r + \frac{1}{\Delta t} \left( \frac{\partial \ell}{\partial \Psi_r} \right) \right) \)
  end for
\( A = \left( \frac{\partial f}{\partial x} \right)_{q_f} + 2 \Psi \)
\( b = \left( \frac{\partial f}{\partial x} \right)_{q_f} + A m_x - \lambda \)
until convergence in \( F^* \)
return: \( \{A_r, b_r, \lambda_r, \Psi_r, m_{x,r}, S_{x,r}\}_{r=1}^R \)

C.4. Output mapping

We consider an observation model of the form

\[
y(t_i) = C x(t_i) + d + \epsilon_i
\]

where \( \epsilon_i \sim N(0, \Gamma) \). Dropping all terms that are constant in \( C, d \) from the expression for the variational free energy, we have

\[
F^* = -\frac{1}{2} \sum_i \langle (y_i - C x_i - d)^\top \Gamma^{-1} (y_i - C x_i - d) \rangle_{q_x}
\]

Differentiating and setting to zero gives

\[
C^{\text{new}} = \left( \sum_i (y_i - d)m_i^\top \right) \left( \sum_i (S_{x,t} + m_{x,t} m_{x,t}^\top) \right)^{-1}
\]
\[
d^{\text{new}} = \frac{1}{T} \sum_i (y_i - C^{\text{new}} m_{x,t})
\]

D. Chemical reaction dynamics

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

\[
\frac{d[\Gamma^-]_A}{dt} = (k_a [\Gamma^-]_A + k_b [\Gamma^-]_A^2) (S_0 - [\Gamma^-]_A)
+ F_1 [\Gamma^-]_0 \frac{V_A}{V_A} - (F_3 + F_4) [\Gamma^-]_A + \frac{F_4 [\Gamma^-]_D}{V_A}
\]
\[
\frac{d[\Gamma^-]_D}{dt} = (k_a [\Gamma^-]_D + k_b [\Gamma^-]_D^2) (S_0 - [\Gamma^-]_D)
+ \frac{F_4 [\Gamma^-]_A}{V_D} - \frac{F_3 [\Gamma^-]_D}{V_D}
\]

To generate the simulations, we use the parameter settings

\[
[\Gamma^-]_0 = 4.4 \times 10^{-5}, \quad k_0 = 2.7 \times 10^{-3}
V_A = 4 \times 10^1, \quad F_1 = \frac{1}{2} F_3
V_D = 1, \quad F_3 = k_0 V_A
k_a = 2.1425 \times 10^{-1}, \quad F_2 = \frac{1}{2} F_3
k_b = 2.1425 \times 10^4
S_0 = \frac{1}{2} ([\Gamma^-]_0 + 1.42 \times 10^{-3})
\]

References