Gaussian Process Approximations of Stochastic Differential Equations

Cedric Archambeau

c@ecs.soton.ac.uk
www.ecs.soton.ac.uk/people/ca
School of Electronics and Computer Science
University of Southampton

Joint work with D. Cornford, M. Opper and J. Shawe-Taylor.

Context

Stochastic differential equations:

- Describe the time dynamics of a state vector based on the (approximate) model of the real system.
- The driving noise process correspond to processes not known in the model, but present in the real system.
- Applications in environmental modelling, finance, physics, etc.
Target application: numerical weather prediction

- Numerical weather prediction models:
  - Based on the discretisation of coupled partial differential equations
  - Dynamical models are imperfect
  - State vectors have typically dimension $O(10^6)$.
  - Large number of data, but relatively few compared to dimension
- Previous approaches consider the models as deterministic or propagate only mean forward in time.
- Recent work attempts propagating uncertainty as well (e.g., approximate Monte Carlo methods).
- Most approaches do not deal with estimating unknown model parameters.
- We focus on a GP and a variational approximation and expect it can be applied to very large models, by exploiting localisation, hierarchical models and sparse representations.

Overview

- Basic setting
- Probability measures and state paths
- GP approximation of the posterior measure
- Variational approximation of the posterior measure
Basic setting

- **Stochastic differential equation:**
  \[ dx = f(x) dt + \sqrt{\Sigma} \, dW \]

- **Noise model (likelihood):**
  \[ p(y_n|x(t_n)) = \mathcal{N}(y_n|Hx(t_n), Q) \]

**Ito stochastic differential equation**

- **Discrete time form of Ito's SDE:**
  \[ x_{k+1} = x_k + f(x_k) \Delta t + \epsilon_k \sqrt{\Sigma} \Delta t \]
  \[ \text{with } \epsilon_k \sim \mathcal{N}(0, I) \]

- **The Wiener process is a Gaussian stochastic process with independent increments (if not overlapping):**
  \[ W(t_2) - W(t_1) \perp W(t'_2) - W(t'_1) \]
  \[ W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1) \]
Probability measures of state paths

- The nonlinear function $f$ induces a prior non-Gaussian probability measure over state paths in time:

\[
\begin{array}{c}
x(t) \\
t_0 \quad t_k \quad T
\end{array}
\]

- Inference problem:

\[
\frac{dp_{post}}{dp_{sde}} = \frac{1}{Z} \times \prod_{n=1}^{N} p(y_n|x(t_n))
\]

Gaussian approximation of the posterior measure

- Approximate the posterior measure by a Gaussian process:

\[
p_{post} \approx q_t(x) = \mathcal{N}(m(t), S(t))
\]

- Replace the non-Gaussian Markov process by a Gaussian one:

\[
dx = f_L(x)dt + \sqrt{\Sigma} \, dW
\]
with \( f_L(x, t) = A(t)x + b(t) \)

- Minimize Kullback-Leibler divergence along the state path:

\[
\begin{align*}
\text{KL} [ q \| p_{post} ] &= \int_0^T E(t)dt + \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln |Q| + \ln Z \\
\text{with } E_{sde}(t) &= \frac{1}{2} \langle \| f - f_L \|_S^2 \rangle_{q_t} \\
E_{obs}(t) &= \frac{1}{2} \sum_n \langle \| y_n - Hx(t) \|_Q^2 \rangle_{q_t} \delta(t-t_n)
\end{align*}
\]
Computing the KL divergence along a state path

- Discretized SDEs:
  \[ \Delta x_k = x_{k+1} - x_k = f(x_k) \Delta t + \sqrt{\Sigma \Delta t} \epsilon_k \]
  \[ \Delta x_k = x_{k+1} - x_k = f_L(x_k, t_k) \Delta t + \sqrt{\Sigma \Delta t} \epsilon_k \]

- Probability density of the discrete time path:
  \[ p(x_{1:K}) = \prod_k \mathcal{N}(x_{k+1} | x_k + f(x_k) \Delta t, \Sigma \Delta t) \]
  \[ q(x_{1:K}) = \prod_k \mathcal{N}(x_{k+1} | x_k + f_L(x_k, t_k) \Delta t, \Sigma \Delta t) \]

- KL along a discrete path:
  \[ \text{KL} [q(x_{1:K}) \| p_{\text{ode}}(x_{1:K})] \]
  \[ = \sum_k \int dx_k q(x_k) \int dx_{k+1} q(x_{k+1} | x_k) \ln \frac{q(x_{k+1} | x_k)}{p(x_{k+1} | x_k)} \]
  \[ = \frac{1}{2} \sum_k \int dx_k q(x_k) (f - f_L)^T \Sigma^{-1} (f - f_L) \Delta t \]

- Pass to a continuum by taking the limit \( \Delta t \to 0 \).

---

Gaussian process posterior moments

- GP approximation of the prior process:
  \[ \min \text{KL} [q \| p_{\text{ode}}] \to A(t) = - \langle \frac{df}{dx} \rangle q(t) \]
  \[ b(t) = - \langle f \rangle q(t) + A(t) m(t) \]

- Compute induced two-time kernel by solving its ordinary differential equations:
  \[ \frac{dK(t_1, t_2)}{dt_2} = -K(t_1, t_2) A^T(t_2) \quad \text{for} \ t_1 < t_2 \]
  \[ \frac{dK(t_1, t_2)}{dt_1} = -A(t_1) K(t_1, t_2) \quad \text{for} \ t_2 < t_1 \]

- Posterior moments (standard GP regression):
  \[ m_* = k_*^T (K + Q)^{-1} y \]
  \[ S_* = k(t_*, t_*) - k_*^T (K + Q)^{-1} k_* \]

---
Example 1: Ornstein-Uhlenbeck process

- Prior process:
  \[ f(x) = -\gamma x \]

- Solution to the kernel ODE:
  \[ K(t_1, t_2) = K(t_1, t_1) \exp\{-A(t_2 - t_1)\} \]

- Resulting induced kernel:
  \[ K(t_1, t_2) = \frac{\sigma^2}{2\gamma} \exp\{-\gamma|t_2 - t_1|\} \]
Example 2:
Double-well system

- **Prior process:**
  \[ f(x) = 4x(1 - x^2) \]

- **Stationary kernel:**
  \[ K(t_1, t_2) = \frac{g^2}{2\alpha} \exp\{-\alpha|t_2 - t_1|\} \]
  with \( \alpha = -4(1 - 3m_f^2 - 3s_f^2) \)

![Graphs of Stationary (OU) and Squared exponential kernels](image)
Variational approximation of the posterior moments

- Why?

- Constraint on the mean and covariance of the marginals:
  \[
  \frac{dm}{dt} = -A(t)m + b(t) \\
  \frac{dS}{dt} = -A(t)S - SA^T(t) + \Sigma
  \]

- Seeking for the stationary points of the Lagrangian leads to:
  \[
  \frac{\partial E}{\partial \lambda} - (\Psi + \Psi^T)S - \lambda m^T = 0 \\
  \frac{\partial E}{\partial b} + \lambda = 0, \\
  \frac{\partial E}{\partial S} - (\Psi^T + \Psi)A + \frac{d\Psi}{dt} = 0 \\
  \frac{\partial E}{\partial m} - A^T\lambda + \frac{d\lambda}{dt} = 0
  \]

A possible smoothing algorithm

Repeat until convergence:
1. Forward propagation of the mean and the covariance.
2. Backward propagation of the Lagrange multipliers:
   \[
   \frac{d\Psi}{dt} = (\Psi^T + \Psi)A - \frac{\partial E_{obs}}{\partial S} \\
   \frac{d\lambda}{dt} = A^T\lambda - \frac{\partial E_{obs}}{\partial m}
   \]
   Use jump conditions when there’s an observation:
   \[
   \frac{\partial E_{obs}}{\partial S} = \frac{1}{2}H^TQ^{-1}H \\
   \frac{\partial E_{obs}}{\partial m} = -H^TQ^{-1}(y_n - Hm(t_n))
   \]
3. Update the parameters of the approximate SDE:
   \[
   A(t) = -\langle \frac{df}{ds} \rangle_{qt} + \Sigma(\Psi(t) + \Psi^T(t)) \\
   b(t) = -\langle f \rangle_{qt} + A(t)m(t) - \Sigma A(t)
   \]
Example 1: Ornstein-Uhlenbeck process

\[ f(x) = -\gamma x \]

\[ f_L(x) = -Ax + b \]

Example 2: Double-well system

\[ f(x) = 4x(1 - x^2) \]
Conclusion

- Proper modelling requires to take into account that the prior process is a non-Gaussian process.

- A key quantity in the energy function is the KL divergence between processes over a time interval (i.e., between probability measures over paths!)

- Unlike in standard GP regression, the feature that the process is infinite dimensional plays a role in the inference.

- These results were preliminary ones, but the framework is a general one (not limited to smoothing in time).