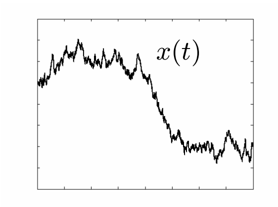
Gaussian Process Approximations of Stochastic Differential Equations

Cedric Archambeau

ca@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



$$dx = f(x)dt + \sqrt{\Sigma}\, dW$$

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.

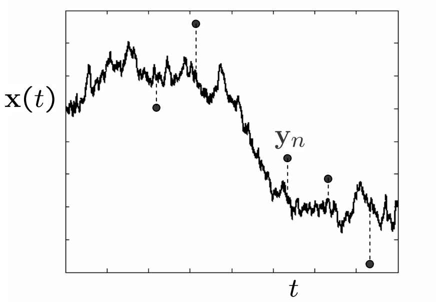
Target application: numerical weather prediction

- Numerical weather prediction models:
  - Based on the discretisation of coupled partial differential equations
  - Dynamical models are imperfect
  - State vectors have typically dimension $\mathcal{O}(10^6)$.
  - 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:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \sqrt{\boldsymbol{\Sigma}}\, d\mathbf{W}$$

- Noise model (likelihood):

$$p(\mathbf{y}_n|\mathbf{x}(t_n)) = \mathcal{N}(\mathbf{y}_n|\mathbf{H}\mathbf{x}(t_n), \mathbf{Q})$$

(Ito) stochastic differential equation

- Discrete time form of Ito's SDE:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)\Delta t + \boldsymbol{\epsilon}_k \sqrt{\boldsymbol{\Sigma}\ \Delta t}$$

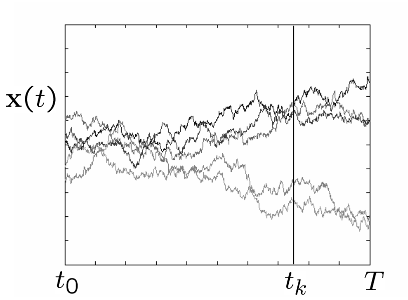
with $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

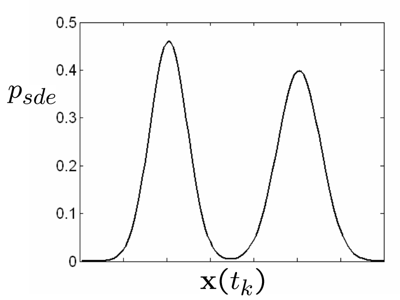
- The Wiener process is a Gaussian stochastic process with independent increments (if not overlapping):

$$W(t_2) - W(t_1) \perp W(t_2') - W(t_1')$$
$$W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1)$$

Probability measures of state paths

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





- Inference problem:

$$\frac{dp_{post}}{dp_{sde}} = \frac{1}{Z} \times \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{x}(t_n))$$

Gaussian approximation of the posterior measure

- Approximate the posterior measure by a Gaussian process:

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

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

$$d\mathbf{x} = \mathbf{f}_L(\mathbf{x})dt + \sqrt{\Sigma}\, d\mathbf{W}$$

with $\mathbf{f}_L(\mathbf{x}, t) = \mathbf{A}(t)\mathbf{x} + \mathbf{b}(t)$

- Minimize Kullback-Leibler divergence along the state path:

$$\mathrm{KL}\left[q \| p_{post}\right] = \int_0^T E(t)dt + \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{Q}| + \ln Z$$

with $E_{sde}(t) = \frac{1}{2}\left\langle \|\mathbf{f} - \mathbf{f}_L\|^2_{\Sigma} \right\rangle_{q_t}$

$$E_{obs}(t) = \frac{1}{2}\sum_n \left\langle \|\mathbf{y}_n - \mathbf{H}\mathbf{x}(t)\|^2_{\mathbf{Q}} \right\rangle_{q_t} \delta(t - t_n)$$

Computing the KL divergence along a state path

- Discretized SDEs:

$$\Delta \mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}(\mathbf{x}_k)\Delta t + \sqrt{\boldsymbol{\Sigma}\Delta t}\,\boldsymbol{\epsilon}_k$$
$$\Delta \mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}_L(\mathbf{x}_k, t_k)\Delta t + \sqrt{\boldsymbol{\Sigma}\Delta t}\boldsymbol{\epsilon}_k$$

- Probability density of the discrete time path:

$$p(\mathbf{x}_{1:K}) = \textstyle\prod_k \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)\Delta t, \boldsymbol{\Sigma}\Delta t)$$
$$q(\mathbf{x}_{1:K}) = \textstyle\prod_k \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + \mathbf{f}_L(\mathbf{x}_k, t_k)\Delta t, \boldsymbol{\Sigma}\Delta t)$$

- KL along a discrete path:

$$\begin{aligned}&\mathrm{KL}\left[q(\mathbf{x}_{1:K})\|p_{sde}(\mathbf{x}_{1:K})\right]\\ &\quad= \textstyle\sum_k \int d\mathbf{x}_k\ q(\mathbf{x}_k) \int d\mathbf{x}_{k+1}\ q(\mathbf{x}_{k+1}|\mathbf{x}_k) \ln \frac{q(\mathbf{x}_{k+1}|\mathbf{x}_k)}{p(\mathbf{x}_{k+1}|\mathbf{x}_k)}\\ &\quad= \tfrac{1}{2}\textstyle\sum_k \int d\mathbf{x}_k\ q(\mathbf{x}_k)\ (\mathbf{f} - \mathbf{f}_L)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \mathbf{f}_L)\Delta t\end{aligned}$$

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

Gaussian process posterior moments

- GP approximation of the prior process:

$$\min \mathrm{KL}\,[q\|p_{sde}] \to \mathbf{A}(t) = -\left\langle \frac{d\mathbf{f}}{d\mathbf{x}} \right\rangle_{q_t}$$
$$\mathbf{b}(t) = -\langle \mathbf{f} \rangle_{q_t} + \mathbf{A}(t)\mathbf{m}(t)$$

- Compute induced two-time kernel by solving its ordinary differential equations:

$$\frac{d\mathbf{K}(t_1,t_2)}{dt_2} = -\mathbf{K}(t_1,t_2)\mathbf{A}^{\top}(t_2) \quad \text{for } t_1 \le t_2$$
$$\frac{d\mathbf{K}(t_1,t_2)}{dt_1} = -\mathbf{A}(t_1)\mathbf{K}(t_1,t_2) \quad \text{for } t_2 \le t_1$$

- Posterior moments (standard GP regression):

$$\mathbf{m}_* = \mathbf{k}_*^{\top}\,(\mathbf{K} + \mathbf{Q})^{-1}\,\mathbf{y}$$
$$\mathbf{S}_* = k(t_*, t_*) - \mathbf{k}_*^{\top}\,(\mathbf{K} + \mathbf{Q})^{-1}\,\mathbf{k}_*$$

Example 1:
Ornstein-Uhlenbeck process

- Prior process:

$$f(x) = -\gamma x$$

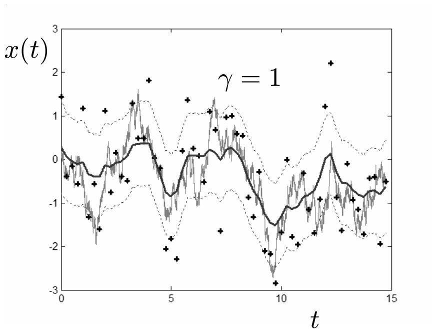
- Solution to the kernel ODE:

$$K(t_1, t_2) = K(t_1, t_1)\exp\{-A(t_2 - t_1)\}$$

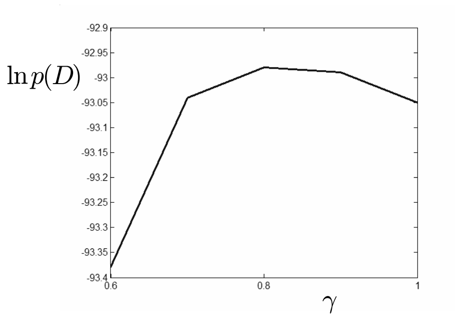
- Resulting induced kernel:

$$K(t_1, t_2) = \frac{\sigma^2}{2\gamma}\exp\{-\gamma|t_2 - t_1|\}$$

Ornstein-Uhlenbeck kernel



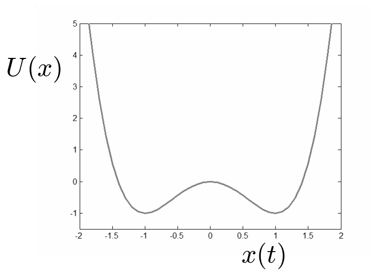
Evidence



Example 2:
Double-well system

- Prior process:

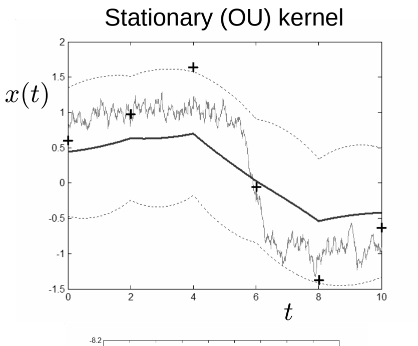
$$f(x) = 4x\,(1 - x^2)$$

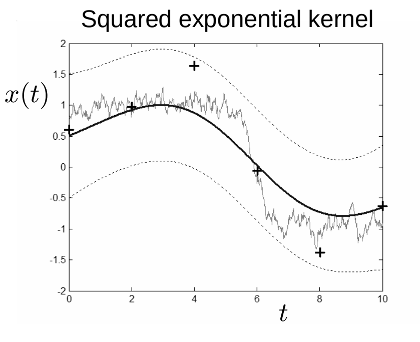


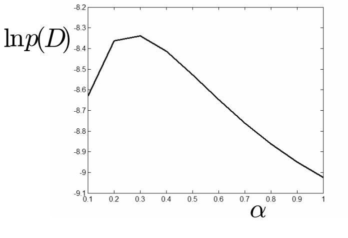
- Stationary kernel:

$$K(t_1, t_2) = \frac{\sigma^2}{2\alpha}\,\mathrm{exp}\{-\alpha|t_2 - t_1|\}$$

with $\alpha = -4(1 - 3m_f^2 - 3s_f^2)$







Variational approximation of the posterior moments

- Why?

- Constraint on the mean and covariance of the marginals:

$$\frac{d\mathbf{m}}{dt} = -\mathbf{A}(t)\mathbf{m} + \mathbf{b}(t)$$
$$\frac{d\mathbf{S}}{dt} = -\mathbf{A}(t)\mathbf{S} - \mathbf{S}\mathbf{A}^\top(t) + \mathbf{\Sigma}$$

- Seeking for the stationary points of the Lagrangian leads to:

$$\frac{\partial E}{\partial \mathbf{A}} - (\mathbf{\Psi} + \mathbf{\Psi}^\top)\mathbf{S} - \boldsymbol{\lambda}\mathbf{m}^\top = 0$$
$$\frac{\partial E}{\partial \mathbf{b}} + \boldsymbol{\lambda} = 0,$$
$$\frac{\partial E}{\partial \mathbf{S}} - (\mathbf{\Psi}^\top + \mathbf{\Psi})\mathbf{A} + \frac{d\mathbf{\Psi}}{dt} = 0$$
$$\frac{\partial E}{\partial \mathbf{m}} - \mathbf{A}^\top\boldsymbol{\lambda} + \frac{d\boldsymbol{\lambda}}{dt} = 0$$

A possible smoothing algorithm

Repeat until convergence:

1. Forward propagation of the mean and the covariance.
2. Backward propagation of the Lagrange multipliers:

$$\frac{d\mathbf{\Psi}}{dt} = (\mathbf{\Psi}^\top + \mathbf{\Psi})\mathbf{A} - \frac{\partial E_{sde}}{\partial \mathbf{S}}$$
$$\frac{d\boldsymbol{\lambda}}{dt} = \mathbf{A}^\top\boldsymbol{\lambda} - \frac{\partial E_{sde}}{\partial \mathbf{m}}$$

   Use jump conditions when there's an observation:

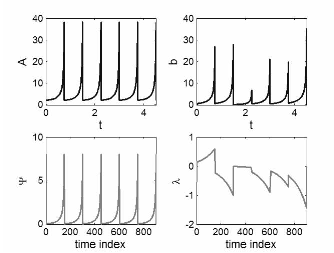
$$\frac{\partial E_{obs}}{\partial \mathbf{S}} = \frac{1}{2}\mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{H}$$
$$\frac{\partial E_{obs}}{\partial \mathbf{m}} = -\mathbf{H}^\top\mathbf{Q}^{-1}(\mathbf{y}_n - \mathbf{H}\mathbf{m}(t_n))$$

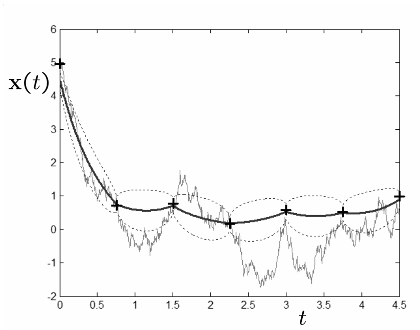
3. Update the parameters of the approximate SDE:

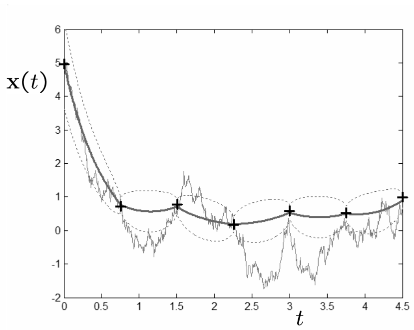
$$\mathbf{A}(t) = -\left\langle \frac{d\mathbf{f}}{d\mathbf{x}} \right\rangle_{q_t} + \mathbf{\Sigma}(\mathbf{\Psi}(t) + \mathbf{\Psi}^\top(t))$$
$$\mathbf{b}(t) = -\langle \mathbf{f} \rangle_{q_t} + \mathbf{A}(t)\mathbf{m}(t) - \mathbf{\Sigma}\boldsymbol{\lambda}(t)$$

Example 1:
Ornstein-Uhlenbeck process

$$f(x) = -\gamma x$$
$$f_L(x) = -Ax + b$$



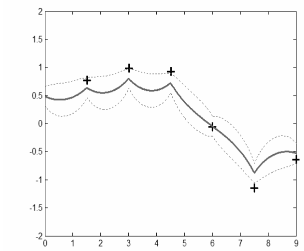




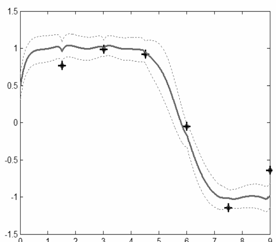
Example 2:
Double-well system

$$f(x) = 4x(1 - x^2)$$

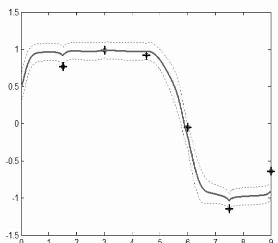
GP initialization



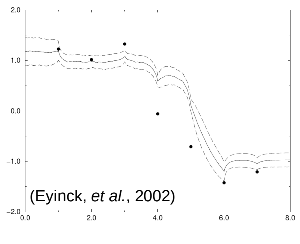
1 FW-BW sweep

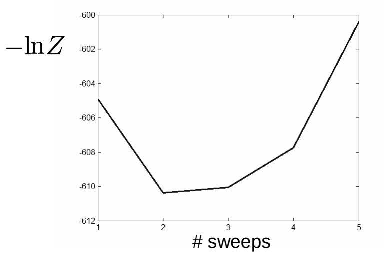


2 FW-BW sweep



Ensemble Kalman smoother





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).