Gaussian Process Approximations
of Stochastic Differential Equations

Cédric Archambeau

Centre for Computational Statistics and Machine Learning
University College London
c.archambeau@cs.ucl.ac.uk

CSML 2007 Reading Group on SDEs
Joint work with Manfred Opper (TU Berlin), John Shawe-Taylor (UCL) and Dan Cornford (Aston).
Context: numerical weather prediction (data assimilation)

- Model equations for the atmosphere are ordinary differential equations (deterministic)
  
  \[
  \begin{align*}
  \frac{Dv}{Dt} &= -\frac{1}{\rho} \nabla \rho - \nabla \phi - 2 \Omega \times v + F, \text{ momentu}m \\
  \frac{\partial p}{\partial t} &= \nabla \cdot (\rho \nu), \text{ mass} \\
  \frac{D T}{D t} &= \frac{1}{\rho c_p} \frac{D p}{D t} + \frac{Q}{c_p}, \text{ energy (2nd LoT)} \\
  \frac{\partial \rho q}{\partial t} &= -\nabla \cdot (\rho q v) + \rho (E - C), \text{ water vapour} \\
  \rho &= \rho R T, \text{ ideal gas law}
  \end{align*}
  \]

- Stochasticity arises from:
  - Dynamics simplifications
  - Discretisation of continuous process
  - Physics (unknown parameters)
  - Observations are satellite radiances (indirect, possibly non-linear)
  - Measurement error

- Decision making must take the uncertainty into account!
Issues with the current methods

Double-well potential

Ensemble Kalman smoother

\[ U(x) \]

\[ x \]

\[ t \]

(Eyink, et al., 2002)
Motivation

Data Assimilation:
- Improve current prediction tools
- Current methods cannot exploit large data sets
- Current methods fail on (highly) non-linear data
- Parameters are unknown (model, noise, observation operator)

Machine Learning:
- Deal with uncertainty in a principled way
- Discover optimal data representation:
  - find sparse solution
  - learn the kernel
  - ...
- Deal with non-linear data
- Deal with huge amount of data
- Deal with high dimensional data
- Deal with large noise
Outline

- Diffusion processes
- Gaussian approximation of the prior process
- Approximate posterior process after observing the data
- Smoothing algorithm (E-step)
- Parameter estimation (M-step)
- Conclusion
Continuous-time continuous-state stochastic processes

Wiener Process:

\[ W_0 = 0 \quad \text{w.p. 1} \]
\[ \langle W_t \rangle = 0 \]
\[ dW_t \sim \mathcal{N}(0, dt) \]
- Almost surely continuous
- Almost surely non-differentiable

- Stochastic differential equation for a diffusion process:

\[ dX_t = f(t, X_t) \, dt + \sqrt{\Sigma} \, dW_t \]

- To be interpreted as an (Ito) stochastic integral:

\[ X_t - X_{t_0} = \int_{t_0}^{t} f(t, X_t) \, dt + \int_{t_0}^{t} \sqrt{\Sigma} \, dW_t \quad \text{w.p. 1} \]
**Fokker-Planck equation:**
- Time evolution of the transition density
- Drift: instantaneous rate of change of the mean
- Diffusion: instantaneous rate of change of squared fluctuations

**Stochastic differential equation:**
- Depends on same drift and diffusion terms
- Alternative representation of the transition density
- (Non-)linear SDE leads to (non-)Gaussian transition density

**Kernel representation:**
- Captures correlations between data
- Particular SDE induces a specific (two-time) kernel
- Time varying approximation of an SDE is equivalent to learning the kernel (hot topic!)
Gaussian approximation of the non-Gaussian process

\[ dX_t \approx f_L(t, X_t) \, dt + \sqrt{\Sigma} \, dW_t \]

- Time varying linear approximation of the drift:
  \[ f(t, x) \approx f_L(t, x) = -A(t)x + b(t) \]

- Gaussian marginal density:
  \[ q_0(x) = \mathcal{N}(x|m_0, S_0) \Rightarrow q_t(x) = \mathcal{N}(x|m(t), S(t)) \]

- Time evolution of the means and the covariances (consistency):
  \[
  \begin{align*}
  \frac{d m}{dt} &= -A m + b \\
  \frac{d S}{dt} &= -AS - SA^T + \Sigma
  \end{align*}
  \]

(see for example second CSML reading group on SDEs)
Optimal (Gaussian) approximation of the prior process

- Euler-Maruyama discrete approximation:

\[
\Delta x_k \equiv x_{k+1} - x_k = f(t_k, x_k) \Delta t + \sqrt{\Sigma} \Delta t \epsilon_k \quad \text{with } \epsilon_k \sim \mathcal{N}(0, I)
\]

- Probability density of a discrete-time sample path:

\[
p(x_{1:K}) = \prod_k \mathcal{N}(x_{k+1} \mid x_k + f(t_k, x_k) \Delta t, \Sigma \Delta t)
\]

\[
q(x_{1:K}) = \prod_k \mathcal{N}(x_{k+1} \mid x_k + f_L(t_k, x_k) \Delta t, \Sigma \Delta t)
\]

- Optimal prior process:

\[
KL[q(x_{1:K}) \mid \mid p(x_{1:K})] = \frac{1}{2} \sum_k \langle (f - f_L)^T \Sigma^{-1} (f - f_L) \rangle_{q(x_k)} \Delta t
\]

... path integral Kullback-Leibler divergence!
Comment on the Kullback-Leibler divergence

Is this measure a good criterion?

$$\text{KL}[q\|p] = \int q(x) \ln \frac{q(x)}{p(x)} \, dx$$

![Graph showing the Kullback-Leibler divergence](image)
Including the observations

- Posterior process:
  \[
  \frac{dp_{post}}{dp_{sde}} = \frac{1}{Z} \times \prod_n p(y_n|x_n)
  \]

- Gaussian likelihood with linear observation operator (for simplicity):
  \[
  p(y_n|x_n) = \mathcal{N}(y_n|Hx_n, R)
  \]

- EM-type training algorithm:
  \[
  \ln p(y|\theta) = -\mathcal{F}(q, \theta) + \text{KL}[q||p_{post}]
  \]
  where the lower bound is defined as
  \[
  \mathcal{F}(q, \theta) = -\langle \ln p(y, X_t|\theta) \rangle_q + \mathbb{H}[q]
  \]
E-step: estimating the optimal (latent) state path

\[
\text{KL}[q\|p_{\text{post}}] = \frac{1}{2} \int_{t_0}^{t} \langle (f - f_L)^\top \Sigma^{-1} (f - f_L) \rangle_{q_t} dt + \frac{1}{2} \sum_n \langle (y_n - x_n)^\top R^{-1} (y_n - x_n) \rangle_{q_t}
\]

- Find optimal variational functions

- Constrained optimization problem:
  - ODE for the marginal means
  - ODE for the marginal covariances

- Integrating the Lagrangian by parts and differentiating leads to:
  - ODEs for the Lagrange multipliers
  - Gradient for the variational functions
Smoothing algorithm:

1. Fix initial conditions.
2. Repeat until convergence:
   - **Forward sweep:**
     Propagate means and covariances forward in time:
     \[ A, b \rightarrow S, m \]
   - **Backward sweep:**
     Propagate Lagrange multipliers backward in time (adjoint operation):
     \[ A, b, S, m \rightarrow \Psi, \lambda \]
   - Use jump conditions when observations:
     \[ \Psi(t^+) = \Psi(t^-) + \Delta \Psi \]
     \[ \lambda(t^+) = \lambda(t^-) + \Delta \lambda \]
   - **Update variational functions:**
     \[ A, b \] (scaled gradient step)
Double-well example

Gaussian process regression

Variational Gaussian approximation
Comparison to MCMC simulation

Hybrid Monte Carlo approach:

- Reference solution
- Generate sample paths from posterior
- Modify scheme in order to increase acceptance rate (molecular dynamics)
- Still requires to generate 100,000 for good results
- Hard to check convergence

(Y. Shen, Aston)
M-step: learning the parameters by maximum likelihood

\[
\langle \ln p(y, X_t | \theta) \rangle_q = \sum_n \langle \ln p(y_n | x_n, \theta) \rangle_{q_n} + \langle \ln p_{sde} \rangle_q
\]

- Linear transformation \( H \)
- Observations noise \( R \)
- Stochastic noise?

\( \hat{r} = 0.11 \)
When things go wrong...

\[ \sigma = 0.6 \]
Conclusion

- Machine Learning for Data Assimilation
- Modeling the uncertainty is essential
- Learning the kernel is a challenging new concern
- Bunch of suboptimal/intermediate good solutions?
- Promising results: quality of the solution & potential extensions

Future work includes:
- High(er) dimensional data
- Check gradient approach (cf. stochastic noise)
- Simplifications when the force (drift) derives from a potential
- Investigate a full variational Bayesian treatment
- Combine the variational approach with MCMC?

Paper available from: www.cs.ucl.ac.uk/staff/C.Archambeau

Research project: VISDEM (multi-site, EPSRC funded)