Gaussian processes for inference in stochastic differential equations

Manfred Opper, AI group, TU Berlin

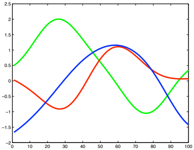
November 6, 2017

Gaussian Process models in machine learning

- Gaussian processes provide prior distributions over latent functions $f(\cdot)$ to be learnt from data.

Gaussian Process models in machine learning

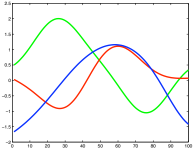
- Gaussian processes provide prior distributions over latent functions $f(\cdot)$ to be learnt from data.



- $f(\cdot) \sim \mathcal{GP}(m, K)$ with $m(x) = E[f(x)]$ and $K(x, x') = Cov\,[f(x), f(x')]$

Gaussian Process models in machine learning

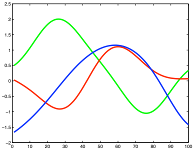
- Gaussian processes provide prior distributions over latent functions $f(\cdot)$ to be learnt from data.



- $f(\cdot) \sim \mathcal{GP}(m, K)$ with $m(x) = E[f(x)]$ and $K(x, x') = Cov\,[f(x), f(x')]$
- Well known examples:
  1. Regression: $y_i = f(x_i) + \nu_i$

Gaussian Process models in machine learning

- Gaussian processes pro-
  vide prior distributions over latent functions $f(\cdot)$ to be learnt from data.



- $f(\cdot) \sim \mathcal{GP}(m, K)$ with $m(x) = E[f(x)]$ and $K(x, x') = Cov\,[f(x), f(x')]$
- Well known examples:
  1. Regression: $y_i = f(x_i) + \nu_i$
  2. Classification $(y_i \in \{0, 1\})$: $P[y_i = 1|f(x_i)] = \sigma\,(f(x_i))$
  3. ...

Inference

- We would like to predict $f(x)$ given observations $\mathbf{y} \doteq (y_1, \ldots, y_n)$ at $x_1, \ldots, x_n$.
- GP prior is infinite dimensional !

Inference

- We would like to predict $f(x)$ given observations $\mathbf{y} \doteq (y_1, \ldots, y_n)$ at $x_1, \ldots, x_n$.
- GP prior is infinite dimensional !
- For these simple models, inference reduces to n dimensional problem Let $f_i \doteq f(x_i)$

$$p(f(x)|\mathbf{y}) = \int p(f(x)|f_1, \ldots, f_n) p(f_1, \ldots, f_n|\mathbf{y}) df_1 \cdots df_n$$

A more complicated likelihood

- Poisson process: Likelihood for set of n points $\mathcal{D} = (x_1, x_2, \ldots, x_n) \in \mathcal{T}$ is given by

$$L(\mathcal{D}|\lambda) = \exp\left\{-\int_{\mathcal{T}} \lambda(x)dx\right\} \prod_{i=1}^{n} \lambda(x_i)$$

A more complicated likelihood

- Poisson process: Likelihood for set of n points $\mathcal{D} = (x_1, x_2, \ldots, x_n) \in \mathcal{T}$ is given by

$$L(\mathcal{D}|\lambda) = \exp\left\{-\int_{\mathcal{T}} \lambda(x)dx\right\} \prod_{i=1}^{n} \lambda(x_i)$$

- $\lambda(x)$ is the unknown **intensity** or **rate** of the process.
- Gaussian Process Modulated Poisson Processes Model (Lloyd et al, 2015) $\lambda(x) = f^2(x)$, where f is a Gaussian process.

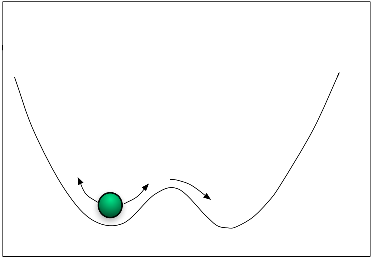
Outline

- Stochastic differential equations
- Drift estimation for dense observations
- Drift estimation for sparse observations
- Sparse GP approximation
- Drift estimation from empirical distribution
- Outlook

Stochastic differential equations

$$\frac{dX}{dt} = f(X) + \text{'white noise'}$$

E.g. $f(x) = -\frac{dV(x)}{dx}$



Prior process: Stochastic differential equations (SDE)

- Mathematicians prefer **Ito** version

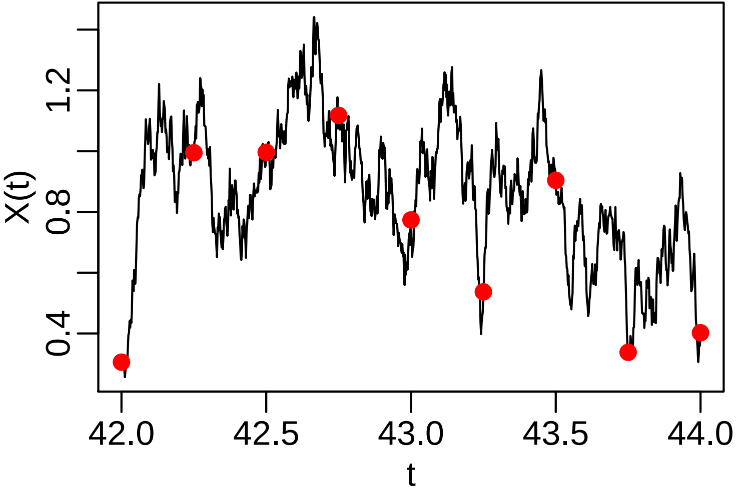
$$dX_t = \underbrace{f(X_t)}_{\text{Drift}} dt + \underbrace{D^{1/2}(X_t)}_{\text{Diffusion}} \times \underbrace{dW_t}_{\text{Wiener process}}$$

for $X_t \in R^d$

- Limit of discrete time process X_k

$$X_{t+\Delta t} - X_t = f(X_t)\Delta t + D^{1/2}(X_t)\sqrt{\Delta t}\,\epsilon_t \ .$$

ϵ_t i.i.d. Gaussian.



Path with observations.

Nonparametric drift estimation

- Infer the drift function $f(\cdot)$ under smoothness assumptions from observations of the process X.

Nonparametric drift estimation

- Infer the drift function $f(\cdot)$ under smoothness assumptions from observations of the process X.
- **Idea** (see e.g. Papaspilioupoulis, Pokern, Roberts & Stuart (2012) Assume a Gaussian Process prior $f(\cdot) \sim \mathcal{GP}(0, K)$ with covariance kernel $K(x, x')$.

Simple for observations dense in time

- Euler discretization of SDE
  $X_{t+\Delta t} - X_t = f(X_t)\Delta t + \sqrt{\Delta t}\,\epsilon_t$, for $\Delta t \to 0$.

Simple for observations dense in time

- Euler discretization of SDE
  $X_{t+\Delta t} - X_t = f(X_t)\Delta t + \sqrt{\Delta t}\,\epsilon_t$, for $\Delta t \to 0$.
- Likelihood (assume **densely observed** path $X_{0:T}$) is Gaussian

$$p(X_{0:T}|f) \propto \exp\left[-\frac{1}{2\Delta t}\sum_t ||X_{t+\Delta t} - X_t||^2\right] \times$$
$$\exp\left[-\frac{1}{2}\sum_t ||f(X_t)||^2\,\Delta t + \sum_t f(X_t)\cdot(X_{t+\Delta t} - X_t)\right].$$

- Posterior process is also a GP !

Simple for observations dense in time

- Euler discretization of SDE
  $X_{t+\Delta t} - X_t = f(X_t)\Delta t + \sqrt{\Delta t}\,\epsilon_t$, for $\Delta t \to 0$.
- Likelihood (assume **densely observed** path $X_{0:T}$) is Gaussian

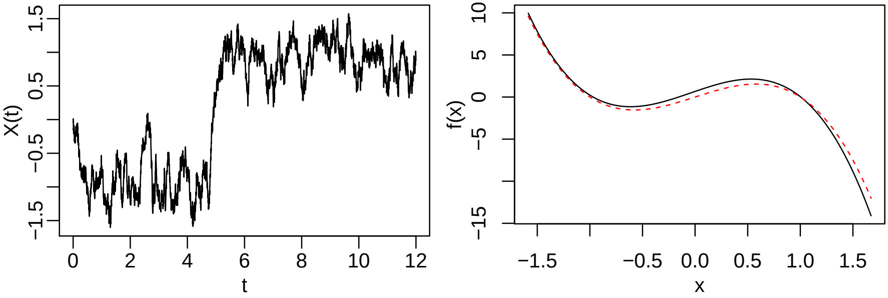
$$p(X_{0:T}|f) \propto \exp\left[-\frac{1}{2\Delta t}\sum_t ||X_{t+\Delta t} - X_t||^2\right] \times$$

$$\exp\left[-\frac{1}{2}\sum_t ||f(X_t)||^2 \Delta t + \sum_t f(X_t)\cdot(X_{t+\Delta t} - X_t)\right].$$

- Posterior process is also a GP !
- Solves the regression problem
  $f(x) \approx E\left[\frac{X_{t+\Delta t}-X_t}{\Delta t}|X_t = x\right]$. Works well for $\Delta t \to 0$.

Example: Double well model

$n = 6000$ data points with $\Delta_t = 0.002$, GP with polynomial kernel of order 4.



Estimation of diffusion

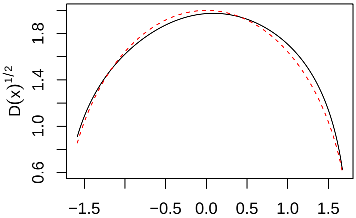
- Euler discretized SDE

$$X_{t+\Delta} - X_t = f(X_t)\Delta t + \sqrt{\Delta t}\,\epsilon_t$$

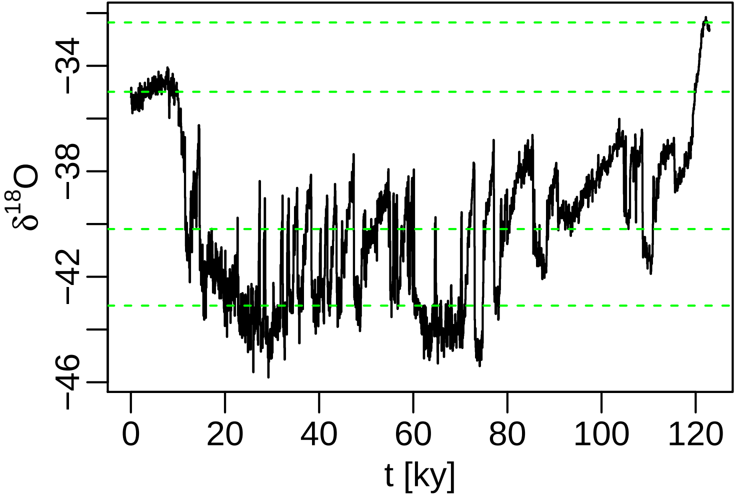
- Diffusion

$$D(x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{Var}\left(X_{t+\Delta t} - X_t | X_t = x\right) =$$
$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{E}\left[(X_{t+\Delta t} - X_t)^2 | X_t = x\right].$$

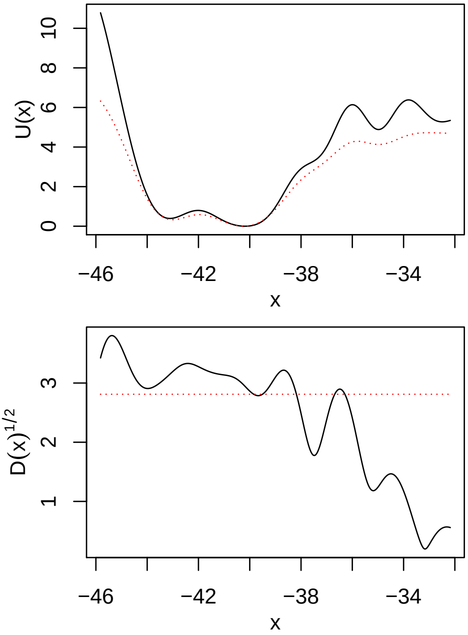
- Independent of drift !
- Estimate $D(x)$ with GPs by regression with data $y_t = (X_{t+\Delta t} - X_t)^2$



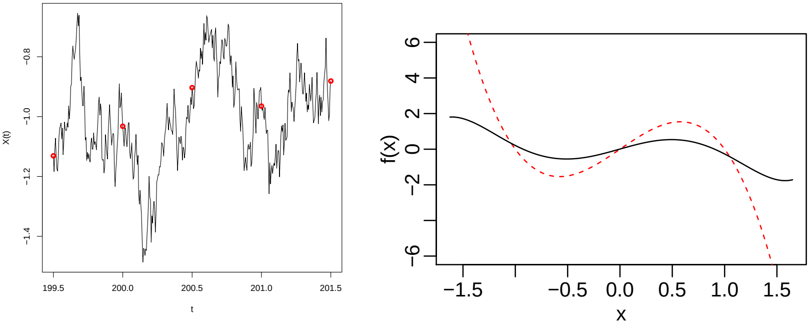
Ice–core data



GP inference using RBF kernels



For larger Δt ...



Problem: We can compute $p(X_{0:T}|f)$ but NOT $P(\mathbf{y}|f)$!

EM–Algorithm

Treat unobserved path X_t for times t between observations as **latent** random variables (Batz, Ruttor & Opper NIPS 2013).

EM algorithm:

1. **E-step**: Compute expected complete data likelihood

$$\mathcal{L}(f, f_{old}) = -E_{p_{old}}\left[\ln L(X_{0:T}|f)\right] \tag{1}$$

   where p_{old} = posterior $p(X_{0:T}|\mathbf{y})$ computed with the previous estimate f_{old} of the drift.

2. **M-Step**: Recompute the drift function as

$$f_{new} = \arg\min_f \left(\mathcal{L}(f, f_{old}) - \ln P_0(f)\right) \tag{2}$$

Likelihood for a complete path

$$p(X_{0:T}|f) \propto \exp\left[-\frac{1}{2\Delta t}\sum_t ||X_{t+\Delta t} - X_t||^2\right] \times$$

$$\underbrace{\exp\left[-\frac{1}{2}\sum_t ||f(X_t)||^2 \Delta + \sum_t f(X_t)\cdot(X_{t+\Delta t} - X_t)\right]}_{L(X_{0:T}|f)}$$

The complete likelihood

$$
\begin{aligned}
&\qquad\qquad -\mathrm{E}_p\left[\ln L(X_{0:T}|f)\right] = \\
&\lim_{\Delta t\to 0}\frac{1}{2}\sum_t \mathrm{E}_p\left[||f(X_t)||^2\right]\Delta t - 2\mathrm{E}_p\left[(f(X_t), X_{t+\Delta t} - X_t)\right] \\
&= \frac{1}{2}\int_0^T \mathrm{E}_p\left[||f(X_t)||^2\right] - 2\mathrm{E}_p\left[(f(X_t), g_t(X_t))\right] dt \\
&= \frac{1}{2}\int ||f(x)||^2 A(x)dx - \int (f(x), z(x))\, dx. \qquad (3)
\end{aligned}
$$

where

$$
g_t(x) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathrm{E}_p[X_{t+\Delta t} - X_t | X_t = x],
$$

as well as the functions

$$
A(x) = \int_0^T q_t(x)dt \qquad b(x) = \int_0^T g_t(x)q_t(x)dt.
$$

Problems

- E-step requires posterior marginal densities q_t for diffusion processes with *arbitrary* prior drift functions $f(x)$.

$$p(X_{0:T}|\mathbf{y}, f) \propto p(X_{0:T}|f) \prod_{k=1}^{n} \delta(y_k - X_{k\tau}), \tag{4}$$

- GP has to deal with an infinite amount of densely imputed data.

Approximate solutions

- Linearize drift between consecutive observations (Ornstein Uhlenbeck bridge). Hence, we consider the approximate process for t between two observations

$$dX_t = [f(y_k) - \Gamma_k(X_t - y_k)]dt + D_k^{1/2} dW \tag{5}$$

  with $\Gamma_k = -\nabla f(y_k)$ and $D_k = D(y_k)$. For this process, the transition density is a multivariate Gaussian !

- Work with sparse GP approximation.

Variational sparse GP approximation

L. Csato (2002), M. Titsias (2009)

- Assume measure over functions f of the form $dP(f) = \frac{1}{Z} dP_0(f)\, e^{-U(f)}$.

Variational sparse GP approximation

L. Csato (2002), M. Titsias (2009)

- Assume measure over functions f of the form $dP(f) = \frac{1}{Z} dP_0(f)\, e^{-U(f)}$.
- Approximate by $dQ(f) = \frac{1}{Z_s} dP_0(f)\, e^{-U_s(\mathbf{f}_s)}$. U_s depends only on **sparse** set $\mathbf{f}_s = \{f(x)\}_{x \in S}$ of dim m.

Variational sparse GP approximation

L. Csato (2002), M. Titsias (2009)

- Assume measure over functions f of the form $dP(f) = \frac{1}{Z} dP_0(f)\, e^{-U(f)}$.
- Approximate by $dQ(f) = \frac{1}{Z_s} dP_0(f)\, e^{-U_s(\mathbf{f}_s)}$. U_s depends only on **sparse** set $\mathbf{f}_s = \{f(x)\}_{x \in S}$ of dim m.
- Minimize KL–divergence $D(Q\|P) = \int dQ(f) \ln \frac{dQ(f)}{dP(f)}$

Variational sparse GP approximation

L. Csato (2002), M. Titsias (2009)

- Assume measure over functions f of the form $dP(f) = \frac{1}{Z} dP_0(f)\ e^{-U(f)}$.
- Approximate by $dQ(f) = \frac{1}{Z_s} dP_0(f)\ e^{-U_s(\mathbf{f}_s)}$. U_s depends only on **sparse** set $\mathbf{f}_s = \{f(x)\}_{x \in S}$ of dim m.
- Minimize KL–divergence $D(Q\|P) = \int dQ(f) \ln \frac{dQ(f)}{dP(f)}$
- Integrating over $dQ(f|\mathbf{f}_s) = dP_0(f|\mathbf{f}_s)$ yields optimal

$$U_s(\mathbf{f}_s) = \mathrm{E}_0[U(f)|\mathbf{f}_s]$$

Variational sparse GP approximation

L. Csato (2002), M. Titsias (2009)

- Assume measure over functions f of the form $dP(f) = \frac{1}{Z} dP_0(f)\ e^{-U(f)}$.
- Approximate by $dQ(f) = \frac{1}{Z_s} dP_0(f)\ e^{-U_s(\mathbf{f}_s)}$. U_s depends only on **sparse** set $\mathbf{f}_s = \{f(x)\}_{x \in S}$ of dim m.
- Minimize KL–divergence $D(Q\|P) = \int dQ(f) \ln \frac{dQ(f)}{dP(f)}$
- Integrating over $dQ(f|\mathbf{f}_s) = dP_0(f|\mathbf{f}_s)$ yields optimal

$$U_s(\mathbf{f}_s) = \mathrm{E}_0[U(f)|\mathbf{f}_s]$$

- Can be computed analytically for Gaussian prior P_0 and quadratic log–likelihood U

Some details

- We can show that

$$\mathrm{E}_0[f(x)|\mathbf{f}_s] = \mathbf{k}_s^\top(x)(\mathbf{K}_s)^{-1}\mathbf{f}_s \tag{6}$$

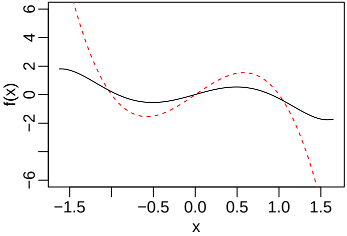
- Hence, if

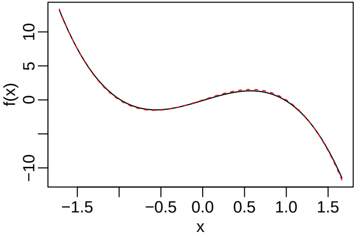
$$U(f) = \frac{1}{2}\int f^2(x)A(x)dx - \int f(x)b(x)dx,$$

we get

$$\begin{aligned} U_s(\mathbf{f}_s) = \mathrm{E}_0[U(\mathbf{f})|\mathbf{f}_s] &= \frac{1}{2}\mathbf{f}_s^\top\mathbf{K}_s^{-1}\left\{\int \mathbf{k}_s(x)\ A(x)\ \mathbf{k}_s^\top(x)dx\right\}\mathbf{K}_s^{-1}\mathbf{f}_s \\ &- \mathbf{f}_s^\top\mathbf{K}_s^{-1}\int \mathbf{k}_s(x)\ b(x)\ dx. \end{aligned}$$

GP estimation after one iteration of EM.





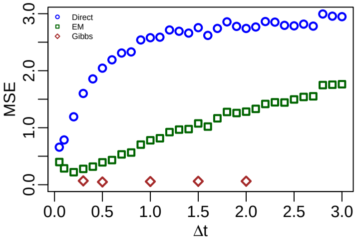


Figure: (color online) Comparison of the MSE for different methods over different time intervals.

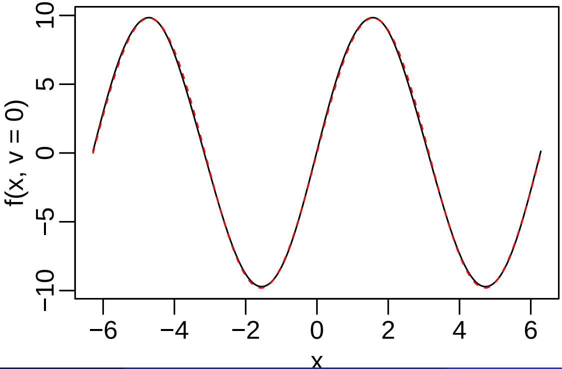
Example: A simple pendulum

$$
\begin{aligned}
dX &= V dt, \\
dV &= \frac{-\gamma V + mgl \sin(X)}{ml^2} dt + d^{1/2} dW_t,
\end{aligned}
$$

Example: A simple pendulum

$$dX = V dt,$$
$$dV = \frac{-\gamma V + mgl\sin(X)}{ml^2} dt + d^{1/2} dW_t,$$

$N = 4000$ data points (x, v) with $\Delta t_{\text{obs}} = 0.3$ and known diffusion constant $d = 1$. GP with periodic kernel.



Drift estimation for SDE using empirical distribution

- For $X \in R^d$ consider SDE

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t$$

- Try to estimate the **drift function** $g(\cdot)$ given only empirical distribution of (noise free) observations

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$$

Drift estimation for SDE using empirical distribution

- For $X \in R^d$ consider SDE

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t$$

- Try to estimate the **drift function** $g(\cdot)$ given only empirical distribution of (noise free) observations

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$$

- Possible, if drift has the form

$$f(x) = g(x) + A(x)\nabla\psi(x)$$

where A and g are known functions.

Generalised score functional

- Define

$$\varepsilon[\psi] = \int \left\{ \frac{1}{2} \|\nabla \psi(x)\|_A^2 + \mathcal{L}_g^\dagger \psi(x) \right\} p(x) dx$$

with

$$\mathcal{L}_g^\dagger \psi(x) = g(x) \cdot \nabla \psi(x) + \frac{1}{2} \sum_{ij} D_{ij}(x) \frac{\partial^2 \psi(x)}{\partial x^{(i)} \partial x^{(j)}}$$

and $\|g(x)\|_A^2 = g(x)^\top \cdot A(x) g(x)$ and $D(x) \doteq \sigma(x)\sigma(x)^\top$.

Generalised score functional

- Define

$$\varepsilon[\psi] = \int \left\{ \frac{1}{2} \|\nabla\psi(x)\|_A^2 + \mathcal{L}_g^\dagger \psi(x) \right\} p(x) dx$$

with

$$\mathcal{L}_g^\dagger \psi(x) = g(x) \cdot \nabla\psi(x) + \frac{1}{2} \sum_{ij} D_{ij}(x) \frac{\partial^2 \psi(x)}{\partial x^{(i)} \partial x^{(j)}}$$

and $\|g(x)\|_A^2 = g(x)^\top \cdot A(x) g(x)$ and $D(x) \doteq \sigma(x)\sigma(x)^\top$.

- $\frac{\delta\varepsilon[\psi]}{\delta\psi} = 0$ yields **stationary Fokker–Planck equation**

$$\mathcal{L}_g p(x) - \nabla \cdot (A(x)\psi(x)p(x)) = 0$$

where $\mathcal{L}_g$ is the adjoint of $\mathcal{L}_g^\dagger$

$$\mathcal{L}_g p(x) = \nabla \cdot \left[-f(x)p(x) + \frac{1}{2} \nabla \cdot (D(x)p(x)) \right] \; .$$

- $A = g = 0$: Score matching (Hyvärinen, 2005), (Sriperumbudur et al, 2014) for density estimation.

- Consider 'pseudo' log–likelihood

$$\sum_{i=1}^{n} \left\{ \frac{1}{2} \|\nabla \psi(x_i)\|_A^2 + \mathcal{L}_g^{\dagger} \psi(x_i) \right\}$$

where $x_i \doteq X(t_i)$, $i = 1, \ldots, n$ is sample from the stationary density $p(\cdot)$

- Consider 'pseudo' log–likelihood

$$\sum_{i=1}^{n} \left\{ \frac{1}{2} \|\nabla \psi(x_i)\|_A^2 + \mathcal{L}_g^\dagger \psi(x_i) \right\}$$

where $x_i \doteq X(t_i)$, $i = 1, \ldots, n$ is sample from the stationary density $p(\cdot)$

- Combine with a GP prior over ψ yields **estimator for drift of the SDE** if this is of the form

$$f(x) = g(x) + A(x)\nabla\psi(x)$$

where A and g are given (Batz, Ruttor, Opper 2016).

Langevin dynamics

- Consider 2nd order SDE

$$dX_t = V_t dt, \qquad dV_t = (F(x) - \lambda v)dt + \sigma(X_t, V_t)dW_t.$$

and set $A_{xx} = A_{xv} = 0$ and $A_{vv} = I$, $f_x = v$, $g_v = -\lambda v$ (known)

Langevin dynamics

- Consider 2nd order SDE

$$dX_t = V_t dt, \qquad dV_t = (F(x) - \lambda v)dt + \sigma(X_t, V_t)dW_t.$$

and set $A_{xx} = A_{xv} = 0$ and $A_{vv} = I$, $f_x = v$, $g_v = -\lambda v$ (known)

- The condition

$$f_v(x, v) = -\lambda v + \nabla_v \psi(x, v) = -\lambda v + F(x)$$

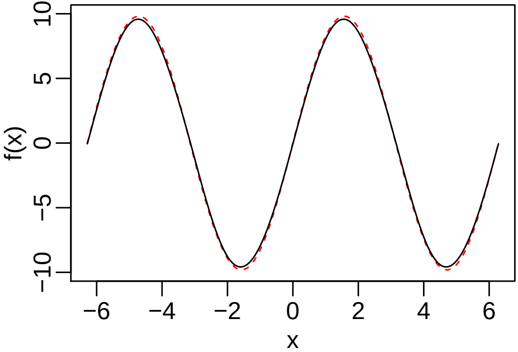
is fulfilled with

$$\psi(x, v) = v \cdot F(x)$$

and allows for arbitrary $F(x)$. Use GP prior over F.

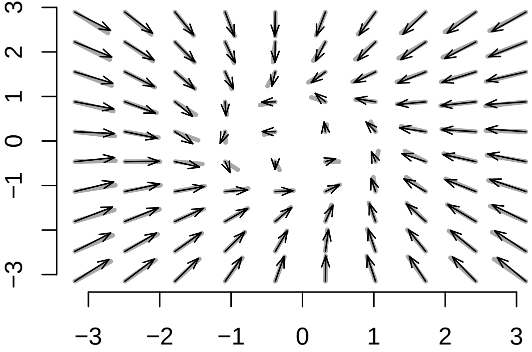
Example 1:

Periodic model:
$F(x) = a \sin x$, $D_v = (\sigma \cos(x))^2$ with $n = 2000$ observations, time lag $\tau = 0.25$



Example 2:

Non–conservative drift model:
$F^{(1)}(x) = x^{(1)}(1 - (x^{(1)})^2 - (x^{(2)})^2) - x^{(2)}$
$F^{(2)}(x) = x^{(2)}(1 - (x^{(1)})^2 - (x^{(2)})^2) - x^{(1)}$
with friction λ known, but D unknown, $n = 2000$ observations.



The case $A = D$: Likelihood for densely observed path

$$-\ln L(X_{0:T}|g) = \frac{1}{2}\int \left\{||f(X_t)||^2_{D^{-1}} dt - 2\langle f(X_t), dX_t\rangle\right\} + \text{const}$$

with $\langle u, v\rangle \doteq u^\top D^{-1} v$.

The case $A = D$: Likelihood for densely observed path

$$-\ln L(X_{0:T}|g) = \frac{1}{2}\int \left\{||f(X_t)||^2_{D^{-1}} dt - 2\langle f(X_t), dX_t\rangle\right\} + \text{const}$$

with $\langle u, v\rangle \doteq u^\top D^{-1} v$. Assume $f = g + D\nabla\psi$ and apply Ito formula

$$= \frac{1}{2}\int_0^T \{\nabla\psi \cdot D\, \nabla\psi\, dt + 2g \cdot \nabla\psi\, dt - 2\nabla\psi \cdot dX_t\}$$

The case $A = D$: Likelihood for densely observed path

$$-\ln L(X_{0:T}|g) = \frac{1}{2}\int \left\{||f(X_t)||^2_{D^{-1}} dt - 2\langle f(X_t), dX_t\rangle\right\} + \text{const}$$

with $\langle u, v\rangle \doteq u^\top D^{-1} v$. Assume $f = g + D\nabla\psi$ and apply Ito formula

$$= \frac{1}{2}\int_0^T \{\nabla\psi \cdot D\ \nabla\psi\, dt + 2g \cdot \nabla\psi\, dt - 2\nabla\psi \cdot dX_t\}$$

$$= \frac{1}{2}\int_0^T \left\{\|\nabla\psi(X_t)\|^2_D + \mathcal{L}_g^\dagger \psi(X_t)\right\} dt - \psi(X_T) + \psi(X_0)$$

Outlook

- Beyond EM: Full variational Bayesian approximation
- Estimation of diffusion from sparse data
- Quality of sparse GP approximation ?
- Langevin dynamics with unobserved velocities ?

Many thanks to my collaborators:

Andreas Ruttor, Philipp Batz

funded by EU–STREP project

CompLACS
Composing Learning for Artificial Cognitive Systems