Variational Bayesian inference for stochastic processes

Manfred Opper, AI group, TU Berlin

October 13, 2017

Outline

- Probabilistic inference (”inverse problem”)
- Why it is not trivial ...
- Variational Approximation
- Path inference for stochastic differential equations
- Drift estimation
- Outlook

Probabilistic inference

- Observations $y \equiv (y_1, \dots, y_K)$ ("**data**")
- Latent, unobserved variables $x \equiv (x_1, \dots, x_N)$
- Likelihood $p(y|x)$ **forward model**
- Prior distribution $p(x)$

Probabilistic inference

- Observations $y \equiv (y_1, \ldots, y_K)$ ("**data**")
- Latent, unobserved variables $x \equiv (x_1, \ldots, x_N)$
- Likelihood $p(y|x)$ **forward model**
- Prior distribution $p(x)$
- **Inverse problem:** Make predictions on x given observations using **Bayes rule:**

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Probabilistic inference

- Observations $y \equiv (y_1, \dots, y_K)$ ("**data**")
- Latent, unobserved variables $x \equiv (x_1, \dots, x_N)$
- Likelihood $p(y|x)$ **forward model**
- Prior distribution $p(x)$
- **Inverse problem:** Make predictions on x given observations using **Bayes rule:**

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- Easy ?

Not quite ...

- Often easy to write down the posterior of **all** hidden variables

$$p(x_1, \ldots, x_N | \text{data}) = \frac{p(\text{data}|x_1, \ldots, x_N)\, p(x_1, \ldots, x_N)}{p(\text{data})}$$

Not quite ...

- Often easy to write down the posterior of **all** hidden variables

$$p(x_1, \dots, x_N | \text{data}) = \frac{p(\text{data}|x_1, \dots, x_N) p(x_1, \dots, x_N)}{p(\text{data})}$$

- But what we really need are **marginal distributions** eg.

$$p(x_i|\text{data}) =$$

$$\int dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_N \; \frac{p(\text{data}|x_1, \dots, x_N) p(x_1, \dots, x_N)}{p(\text{data})}$$

- and

$$p(\text{data}) = \int dx_1 \dots \dots dx_N \; p(\text{data}|x_1, \dots, x_N) p(x_1, \dots, x_N)$$

Variational approximation

- **Approximate** intractable posterior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

by $q(x)$ which belongs to a family of **simpler tractable** distributions.

Variational approximation

- **Approximate** intractable posterior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

by $q(x)$ which belongs to a family of **simpler tractable** distributions.

- Optimise q by minimising the **Kullback–Leibler divergence** (relative entropy)

$$D_{KL}[q\|p(\cdot|y)] \doteq E_q\left[\ln \frac{q(x)}{p(x|y)}\right] =$$
$$D_{KL}[q\|p] - E_q[\ln p(y|x)] + \ln p(y)$$

Variational approximation

- **Approximate** intractable posterior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

by $q(x)$ which belongs to a family of **simpler tractable** distributions.

- Optimise q by minimising the **Kullback–Leibler divergence** (relative entropy)

$$D_{KL}[q \| p(\cdot|y)] \doteq E_q\left[\ln \frac{q(x)}{p(x|y)}\right] =$$
$$D_{KL}[q\|p] - E_q[\ln p(y|x)] + \ln p(y)$$

- **Minimize** the **variational free energy**

$$\mathcal{F}[q] = D_{KL}[q\|p] - E_q[\ln p(y|x)] \geq -\ln p(y)$$

The variational approximation in statistical physics

(Feynman, Peierls, Bogolubov, Kleinert...)

- Let $p(x|y) = \frac{1}{Z}\, e^{-H(x)}$ and $q(x) = \frac{1}{Z_0}\, e^{-H_0(x)}$
- The variational bound on the free energy is

$$-\ln Z \leq -\ln Z_0 + E_q[H(x)] - E_q[H_0(x)] = \mathcal{F}[q]$$

- Equivalent to first order perturbation theory around H_0
- Well known approximations: Gaussian, factorising ("mean field").

Example: Fimite dim Gaussian variational densities

$$q(\mathbf{x}) = (2\pi)^{-N/2}|\mathbf{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

The variational free energy becomes

$$\mathcal{F}[q] = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{N}{2} - E_q[\log p(\mathbf{y},\mathbf{x})]$$

Example: Fimite dim Gaussian variational densities

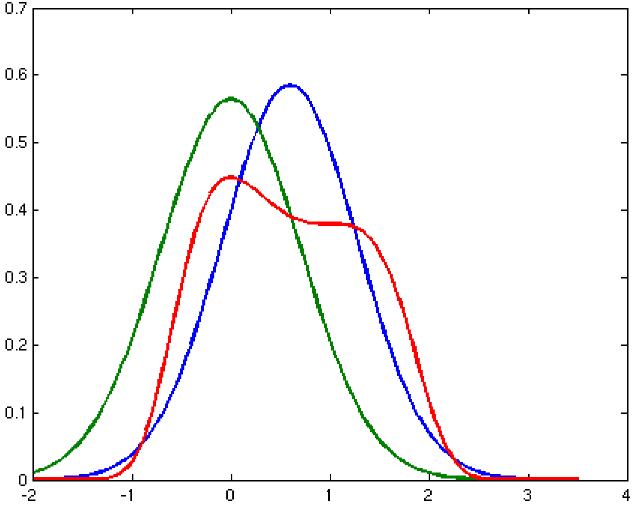
$$q(\mathbf{x}) = (2\pi)^{-N/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right).$$

The variational free energy becomes

$$\mathcal{F}[q] = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{N}{2} - E_q[\log p(\mathbf{y}, \mathbf{x})]$$

Taking derivatives w.r.t. variational parameters

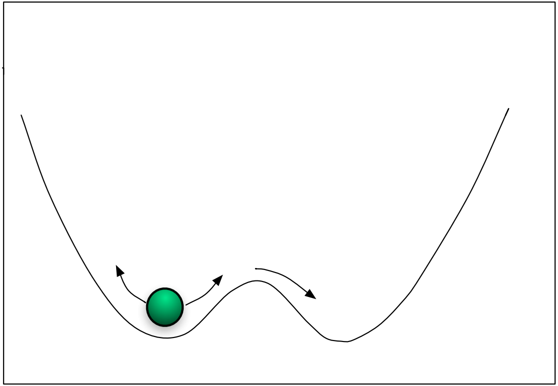
$$\begin{aligned} 0 &= E_q \left[\nabla_{\mathbf{x}} \log p(\mathbf{y}, \mathbf{x})\right] \\ (\boldsymbol{\Sigma}^{-1})_{ij} &= -E_q \left[\frac{\partial^2 \log p(\mathbf{y}, \mathbf{x})}{\partial x_i \partial x_j}\right] \end{aligned}$$


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-2
-1
0
1
2
3
4

Stochastic differential equation

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

E.g. $f_\theta(x) = -\frac{dV_\theta(x)}{dx}$



Prior process: Stochastic differential equations (SDE)

- Mathematicians prefer **Ito** version

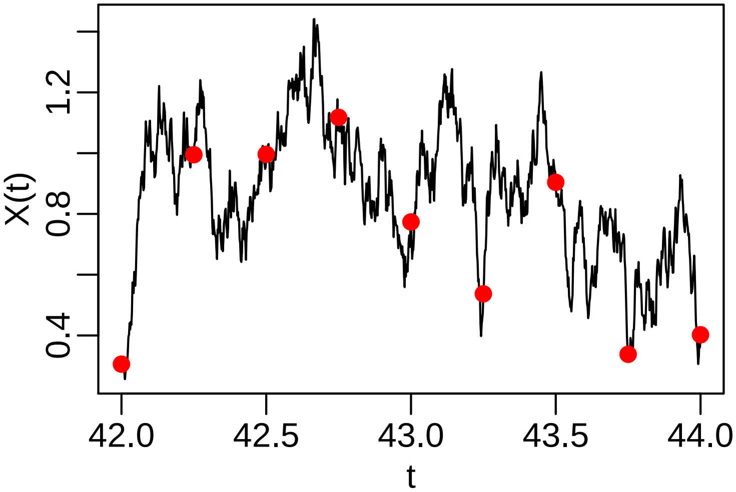
$$dX_t = \underbrace{f(X_t)}_{\text{Drift}}\, dt + \underbrace{D^{1/2}(X_t)}_{\text{Diffusion}} \times \underbrace{dW_t}_{\text{Wiener process}}$$

  for $X_t \in R^d$

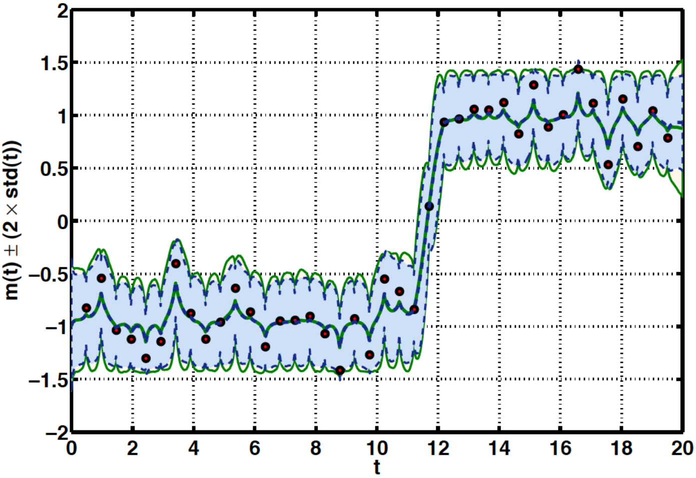
- Limit of discrete time process X_k

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

  ϵ_k i.i.d. Gaussian.



Path with observations.



Inference of unobserved path.

What we would like to do

- **State estimation (smoothing:)** $p[X_t|\{y_i\}_{i=1}^N, \theta]$

What we would like to do

- **State estimation (smoothing:)** $p[X_t | \{y_i\}_{i=1}^N, \theta]$
- Use **Bayes rule** for conditional distribution over **paths** $X_{0:T}$ (∞ dimensional object)

$$p(X_{0:T} | \{y_i\}_{i=1}^N, \theta) = \underbrace{p_{prior}(X_{0:T} | \theta)}_{\text{dynamics}} \quad \underbrace{\prod_{n=1}^N p(y_n | X_{t_n})}_{\text{observation model}} \quad / p(\{y_i\}_{i=1}^N | \theta)$$

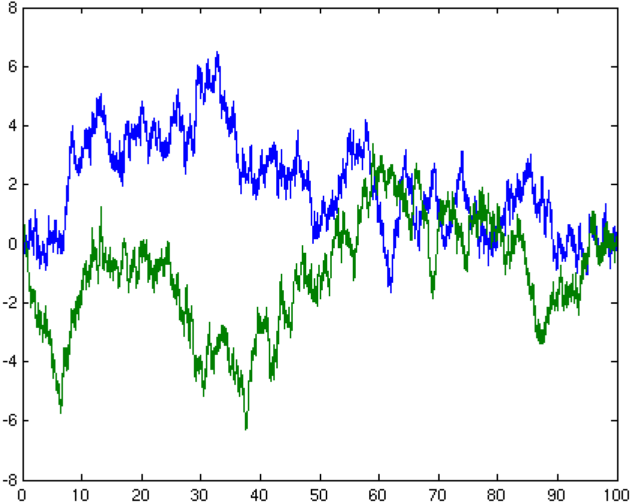
What we would like to do

- **State estimation (smoothing:)** $p[X_t|\{y_i\}_{i=1}^N, \theta]$
- Use **Bayes rule** for conditional distribution over **paths** $X_{0:T}$ (∞ dimensional object)

$$p(X_{0:T}|\{y_i\}_{i=1}^N, \theta) = \underbrace{p_{prior}(X_{0:T}|\theta)}_{\text{dynamics}} \quad \underbrace{\prod_{n=1}^N p(y_n|X_{t_n})}_{\text{observation model}} \quad / p(\{y_i\}_{i=1}^N|\theta)$$

- **Parameter estimation**:
  1. Maximum Likelihood: Maximise $p(\{y_i\}_{i=1}^N|\theta)$ with respect to θ
  2. Bayes: Use prior over parameters $p(\theta)$ to compute $p(\theta|\{y_i\}_{i=1}^N) \propto p(\{y_i\}_{i=1}^N|\theta)p(\theta)$

Example: Process conditioned on endpoint



Wiener process with single, noise free observation $y = X_T = 0$

How to represent path measure ?

- Conditioned process is also Markovian!

How to represent path measure ?

- Conditioned process is also Markovian!
- It fulfils SDE

$$dX_t = g(X_t, t)dt + D^{1/2}(X_t)\, dW_t$$

with a new time dependent drift $g(X_t, t)$ but the **same diffusion** D.

How to represent path measure ?

- Conditioned process is also Markovian!
- It fulfils SDE
$$dX_t = g(X_t, t)dt + D^{1/2}(X_t)\, dW_t$$
with a new time dependent drift $g(X_t, t)$ but the **same diffusion** D.
- Previous example: $g(x, t) = -\frac{x}{T-t}$ for $0 < t < T$.

Change of measure theorem and KL divergence for path probabilities

- Girsanov theorem

$$\frac{dQ}{dP}(X_{0:T}) = \exp\left\{-\int_0^T (f-g)^\top D^{-1/2}\, dB_t + \frac{1}{2}\int_0^T \|f-g\|_D^2\, dt\right\}$$

B_t: Wiener process with respect to Q and
$\|f-g\|_D = f^\top(x,t)D^{-1}g(x,t)$

Change of measure theorem and KL divergence for path probabilities

- Girsanov theorem

$$\frac{dQ}{dP}(X_{0:T}) = \exp\left\{-\int_0^T (f-g)^\top D^{-1/2}\, dB_t + \frac{1}{2}\int_0^T \|f-g\|_D^2\, dt\right\}$$

B_t: Wiener process with respect to Q and
$\|f-g\|_D = f^\top(x,t)D^{-1}g(x,t)$

- Let Q and P be measures over paths for SDEs with drifts $g(X,t)$ and $f(X,t)$ having the **same diffusion** $D(X)$. Then

$$D\left[Q\|P\right] = E_Q \ln\frac{dQ}{dP} = \frac{1}{2}\int_0^T dt\left\{\int dx\; q(x,t)\;\|g(x,t) - f_\theta(x)\|^2\right\}$$

$q(x,t)$ is the marginal density of X_t.

The (full) variational problem

- Minimise variational free energy $\mathcal{F}(Q) =$

$$= \frac{1}{2}\int_0^T \int q(x,t) \left\{ \|g(x,t) - f_\theta(x)\|^2 \; - \sum_i \delta(t - t_i) \ln p(y_i|x) \right\} dx\; dt$$

with respect to the posterior drift $g(x, t)$.

The (full) variational problem

- Minimise variational free energy $\mathcal{F}(Q) =$

$$= \frac{1}{2}\int_0^T \int q(x,t)\left\{\|g(x,t) - f_\theta(x)\|^2 \; - \sum_i \delta(t-t_i)\ln p(y_i|x)\right\} dx\, dt$$

with respect to the posterior drift $g(x,t)$.

- The marginal density $q(x,t)$ and the drift $g(x,t)$ are coupled through the **Fokker - Planck** equation

$$\frac{\partial q(x,t)}{\partial t} = \left\{-\sum_k \partial_k g_k(x) + \frac{1}{2}\sum_{kl}\partial_k\partial_l D_{kl}(x)\right\} q(x,t)$$

Variation leads to forward–backward PDEs: KSP equations (Kushner '62, Stratonovich '60 & Pardoux '82).

The Variational Gaussian Approximation for SDE

(Archambeau, Cornford, Opper & Shawe - Taylor, 2007)

- Approximate (Gaussian) process over paths $X_{0:T}$ induced by linear SDE:
$$dX_t = \{A(t)X_t + b(t)\}\,dt + D^{1/2}dW$$
- Diffusion D must be independent of X !
- Cost function (action) is of the form $\mathcal{F}_\theta[m, S, A, b]$.

The Variational Gaussian Approximation for SDE

(Archambeau, Cornford, Opper & Shawe - Taylor, 2007)

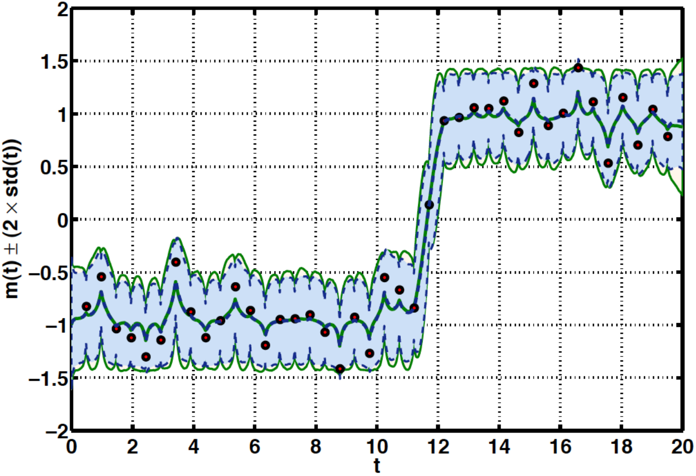
- Approximate (Gaussian) process over paths $X_{0:T}$ induced by linear SDE:
$$dX_t = \{A(t)X_t + b(t)\}\, dt + D^{1/2} dW$$
- Diffusion D must be independent of X !
- Cost function (action) is of the form $\mathcal{F}_\theta[m, S, A, b]$.
- Constraints are evolution eqs. for marginal **mean** $m(t)$ and **covariance** $S(t)$

$$\begin{aligned} \frac{dm}{dt} &= Am + b \\ \frac{dS}{dt} &= AS + SA^\top + D. \end{aligned}$$

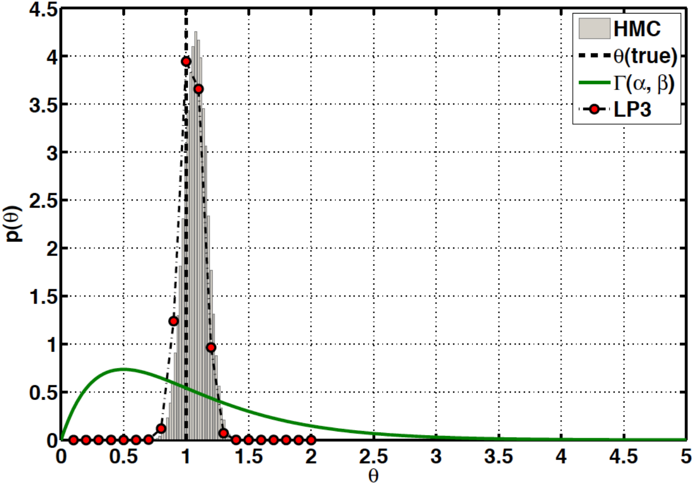
$\rightarrow$ **nonlinear ODEs** instead of PDEs !

Prediction & comparison with hybrid Monte Carlo

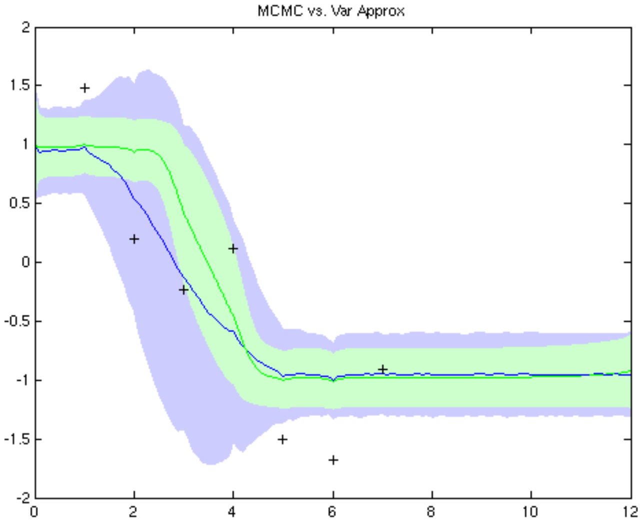
$dX = X(\theta - X^2)dt + \sigma dW$.
$T = 20$, $\theta = 1$, $\sigma^2 = 0.8$ and $N = 40$ observations with noise $\sigma_o^2 = 0.04$.



Posterior for θ



Breakdown for large observation noise



Double well with observation noise $\sigma_o = 0.6$

Variational inference for higher dimensions: Mean field approximation

Action functional (Vrettas, Opper & Cornford, 2015) for mean $m_i(t)$ and variance $s_i(t)$ (compare to Onsager–Machlup)

$$
\mathcal{F}_\theta[q] = \sum_{i=1}^{d} \frac{1}{2\sigma_i^2} \int_0^T E_q\left[\left(\dot{m}_i - f_i(X_t)\right)^2 \right] \, dt
$$

$$
+ \sum_{i=1}^{d} \frac{1}{2\sigma_i^2} \int_0^T \left\{ \frac{(\dot{s}_i - \sigma_i^2)^2}{4 s_i^2} + (\sigma_i^2 - \dot{s}_i) E_q\left[\frac{\partial f_i(X_t)}{\partial X_t^i} \right] \right\} \, dt
$$

$$
- \sum_{j=1}^{n} E_q\left[\ln p(y_j | X_{t_j}) \right]
$$

Variational inference for higher dimensions: Mean field approximation

Action functional (Vrettas, Opper & Cornford, 2015) for mean $m_i(t)$ and variance $s_i(t)$ (compare to Onsager–Machlup)

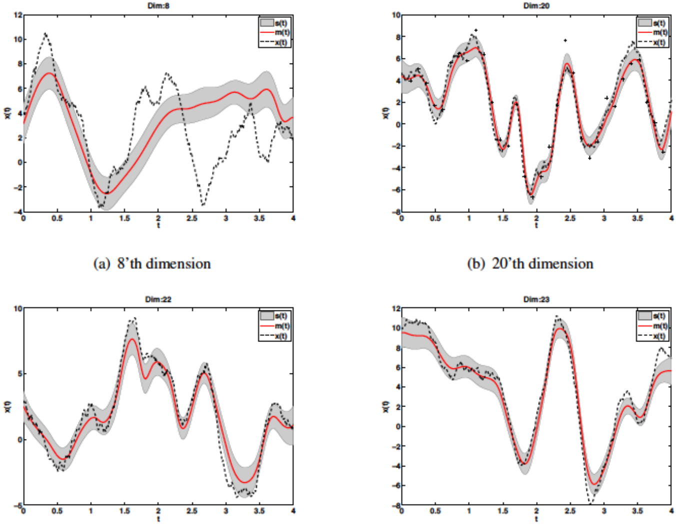
$$\mathcal{F}_\theta[q] = \sum_{i=1}^{d} \frac{1}{2\sigma_i^2} \int_0^T E_q\left[(\dot{m}_i - f_i(X_t))^2\right] \, dt$$

$$+ \sum_{i=1}^{d} \frac{1}{2\sigma_i^2} \int_0^T \left\{ \frac{(\dot{s}_i - \sigma_i^2)^2}{4s_i^2} + (\sigma_i^2 - \dot{s}_i) E_q\left[\frac{\partial f_i(X_t)}{\partial X_t^i}\right] \right\} \, dt$$

$$- \sum_{j=1}^{n} E_q\left[\ln p(y_j | X_{t_j})\right]$$

Test on Lorenz 1998 model: $\mathbf{x} = (x^1, \ldots, x^d)$ with

$$\frac{dx_t^i}{dt} = \left(x^{i+1} - x^{i-2}\right) x^{i-1} - x^i + \theta + \xi^i(t)$$



(a) 8'th dimension

(b) 20'th dimension

(c) 22'nd dimension

(d) 23'rd dimension

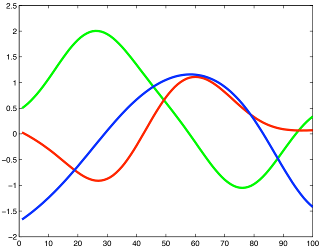
System of 1000 SDE with only 350 components observed.

Nonparametric drift estimation

- Reconsider SDE $dX = f(X)dt + \sigma dW$: Infer the function $f(\cdot)$ under smoothness assumptions from observations of the process X.

Nonparametric drift estimation

- Reconsider SDE $dX = f(X)dt + \sigma dW$: Infer the 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 $K(x, x')$.



Basic idea

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

Basic idea

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

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

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

- Posterior process is also a GP with analytical solution.
- For sparse observations (Δ not small) one needs to impute unobserved path $X_{0:T}$ between observations e.g. within an (approximate) EM–algorithm (Ruttor, Batz, Opper, 2013) .

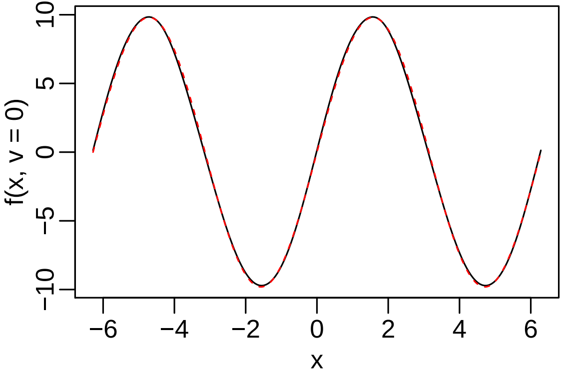
A simple pendulum

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

A simple pendulum

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

$N = 4000$ data points (x, v) with $\Delta t = 0.3$ and known diffusion constant $d = 1$.



Outlook

- Bias of approximation ?

Outlook

- Bias of approximation ? Not easy, because D_{KL} only known up to a constant !

Outlook

- Bias of approximation ? Not easy, because D_{KL} only known up to a constant !
- Get rid of bias by using q as informative proposal within MCMC sampler.

Outlook

- Bias of approximation ? Not easy, because D_{KL} only known up to a constant !
- Get rid of bias by using q as informative proposal within MCMC sampler.
- More general infinite dimensional problems (F. Pinski, G. Simpson, A.M. Stuart, H. Weber, 2015)

Outlook

- Bias of approximation ? Not easy, because D_{KL} only known up to a constant !
- Get rid of bias by using q as informative proposal within MCMC sampler.
- More general infinite dimensional problems (F. Pinski, G. Simpson, A.M. Stuart, H. Weber, 2015)
- Inference for SDE beyond Gaussian approximation (T.Sutter, A. Ganguly and Heinz Koeppl, 2016). Allows for state dependent diffusion.

Many thanks to my collaborators:

Dan Cornford & Michail Vrettas (Aston U)
Andreas Ruttor, Florian Stimberg, Philipp Batz (TUB)