

Gaussian Process Approximations to Stochastic Differential Equations

Manfred Opper



joint work with:

Cédric Archambeau (UCL)

Dan Cornford (Aston U)

John Shawe–Taylor (UCL)

Overview

- The Gaussian Variational Method (finite case)
- Application to Stochastic differential equations (infinite case)
- Lagrange - function
- A Hamiltonian formulation for the 'potential case'

The Variational Method in Bayesian Modeling

- Approximate intractable posterior $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) \rightarrow q(\mathbf{x})$ minimising $KL[q||p]$ where $q(\cdot) \in \mathcal{F} =$ family of **tractable** distributions.
- Most applications: $\mathcal{F} =$ **factorising**. Advantage: Free form optimisation possible!
- Only few applications with $\mathcal{F} =$ Gaussian distributions. (see e.g. Barber & Bishop (1998), Seeger (2000), Honkela & Valpola (2005)). **Too many parameters?**

Gauss-Var: The finite \mathbf{D} case

Let \mathbf{y} be observations and \mathbf{x} latent parameters. Approximate posterior

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{y}, \boldsymbol{\theta})},$$

by a **tractable density** $q(\mathbf{x})$ using the Variational bound

$$-\log p(\mathbf{y}, \boldsymbol{\theta}) = \mathcal{F}(q, \boldsymbol{\theta}) - KL [q||p] \leq \mathcal{F}(q, \boldsymbol{\theta})$$

with the **variational free energy**

$$\mathcal{F}(q, \boldsymbol{\theta}) = -H[q] - \langle \log p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \rangle_q$$

Gaussian variational densities

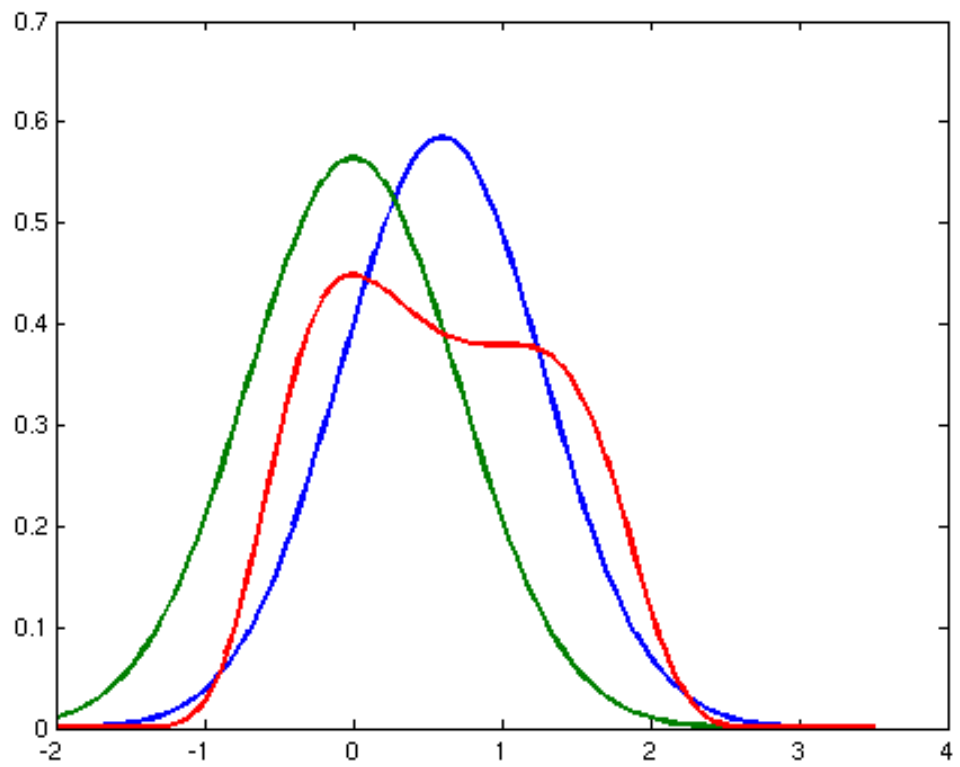
$$q(\mathbf{x}) = (2\pi)^{-N/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

The variational free energy becomes

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

Setting $\nabla \mathcal{F}(q, \boldsymbol{\theta}) = 0$, we obtain

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\mu}} \langle \log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_q \\ \boldsymbol{\Sigma}^{-1} &= -2 \nabla_{\boldsymbol{\Sigma}} \langle \log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_q = - \left\langle \frac{\partial^2 \log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta})}{\partial \mathbf{x}^T \partial \mathbf{x}} \right\rangle_q \end{aligned}$$



GPs with factorising likelihood

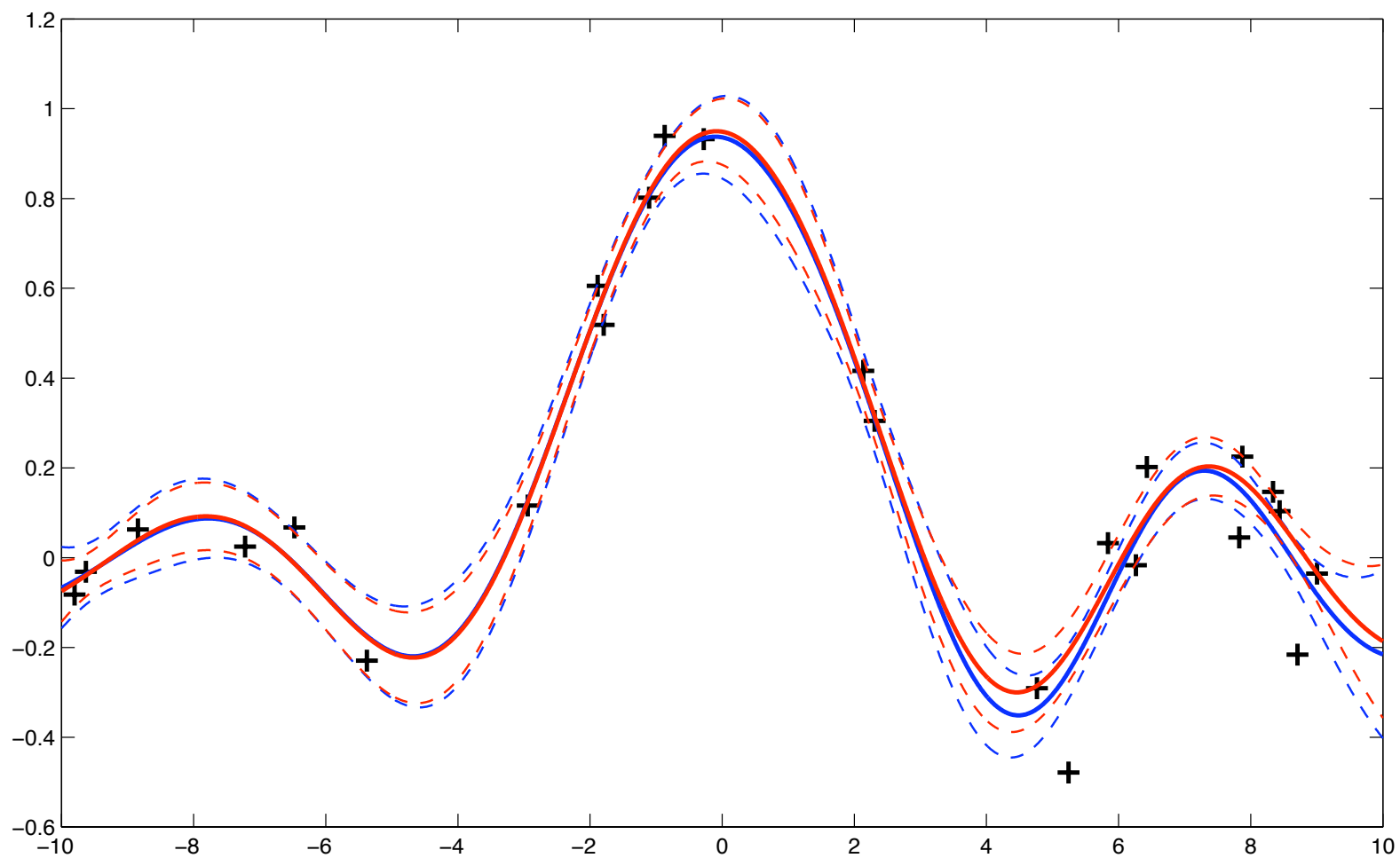
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z_0} \exp \left(- \sum_n V_n(y_n, x_n) - \frac{1}{2} \mathbf{x}^\top \mathbf{K}^{-1} \mathbf{x} \right),$$

Covariance

$$\boldsymbol{\Sigma}^{-1} = \mathbf{K}^{-1} + \text{diag} \left\langle \frac{\partial^2 V_n}{\partial x_n^2} \right\rangle_q$$

is parametrised by N elements!

$$V_n(y_n, x_n) = \lambda|y_n - x_n| \text{ (Laplace noise).}$$



The infinite case: Stochastic differential equations

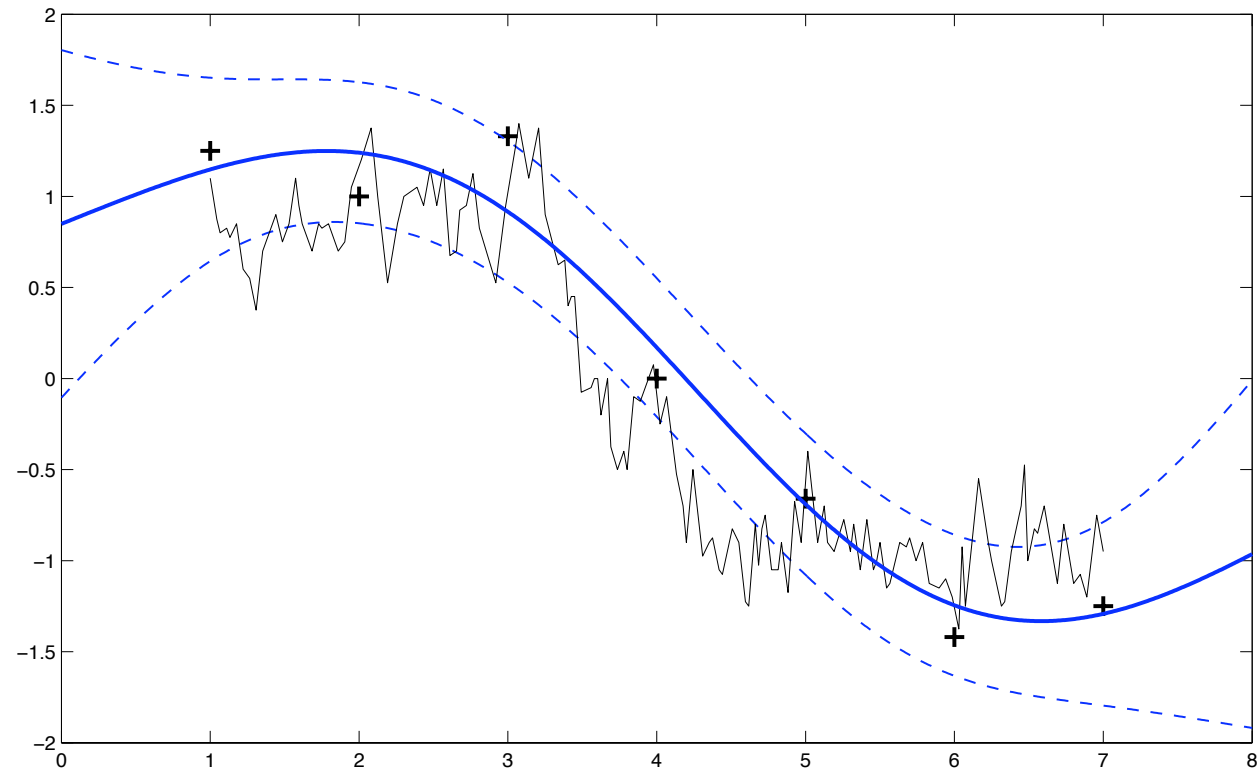
$\{y_n\}_{n=1}^N$: noisy observations of hidden process $\mathbf{x}(t)$ at times $\{t_n\}_{n=1}^N$.

Time evolution: (Ito) stochastic differential equation (SDE):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \sigma d\mathbf{W}$$

$d\mathbf{W}(t)$ is a *Wiener process*.

Goal: Predict latent path & uncertainty



The prior measure

SDE: limit of a discrete time process \mathbf{x}_k

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

Δt = time increment and $\boldsymbol{\epsilon}_k$ is a sequence of i.i.d. $\boldsymbol{\epsilon}_k = \mathcal{N}(0, \mathbf{I})$.

Probability density over **discrete time paths** $\mathbf{x}_{1:K} = (\mathbf{x}_1, \dots, \mathbf{x}_K)$

$$p_{sde}(\mathbf{x}_{1:K}) \propto \prod_{k=1}^{K-1} \exp \left[-\frac{1}{2\sigma^2 \Delta t} \left(\mathbf{x}_{k+1} - \mathbf{x}_k - \mathbf{f}(\mathbf{x}_k) \Delta t \right)^2 \right]$$

Taking the limit & Fancy notation

Convergence to **Wiener measure** μ for $\Delta t \rightarrow 0$ and $K\Delta t = T$

$$\prod_{k=1}^{K-1} \frac{1}{(2\pi\Delta t)^{1/2}} \exp\left[-\frac{1}{2\sigma^2\Delta t} (\mathbf{x}_{k+1} - \mathbf{x}_k)^2\right] \prod_i d\mathbf{x}_i \rightarrow d\mu$$

and

$$\frac{dp_{sde}}{d\mu} = \exp\left[\frac{1}{\sigma^2} \int_0^T \mathbf{f}^T d\mathbf{x} - \frac{1}{2\sigma^2} \int_0^T \|\mathbf{f}\|^2 dt\right]$$

where $\int_0^T \mathbf{f}^T d\mathbf{x}$ is an **Ito - integral**.

The **Stratonovich** form is

$$\frac{dp_{sde}}{d\mu} = \exp\left[\frac{1}{\sigma^2} \int_0^T \mathbf{f}^T \circ d\mathbf{x} - \frac{1}{2\sigma^2} \int_0^T \|\mathbf{f}\|^2 dt - \frac{1}{2} \int_0^T \nabla \mathbf{f} dt\right]$$

The posterior measure

Given the observations, the posterior measure is

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

where $\mathbf{y}_{1:N} \equiv (\mathbf{y}_1, \dots, \mathbf{y}_N)$ and $Z = p(\mathbf{y}_{1:N})$.

The Likelihood is (for simplicity)

$$p(\mathbf{y}_n | \mathbf{x}(t_n)) = \mathcal{N}(\mathbf{x}(t_n), \sigma_o^2),$$

Variational approximation

Approximate p_{post} by Gaussian measure q minimising $KL [q || p_{post}]$

- Posterior is Markovian! → choose **Markovian** q
- Noise variance σ^2 is the same (otherwise $KL = \infty$).
- Gaussianity → linear SDE:

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

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

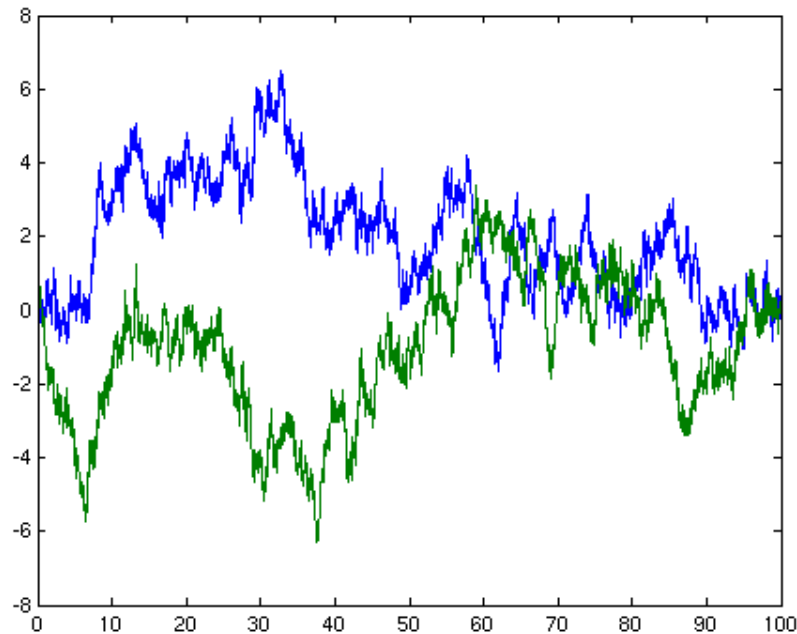
$\mathbf{A}(t)$ and $\mathbf{b}(t)$ must be optimised.

Markovian posterior

Single, noise free observation $y = x(t = T) = 0$

Wiener process:

$$\mathbf{f}_{prior}(x) = 0 \rightarrow \mathbf{f}_{post}(x, t) = -\frac{x}{T-t} \text{ for } 0 < t < T.$$



Ornstein - Uhlenbeck process:

$$\mathbf{f}_{prior}(x) = -\gamma x \rightarrow \mathbf{f}_{post}(x, t) = -\gamma \left(\frac{2}{e^{2\gamma(T-t)} - 1} + 1 \right) x$$

The Kullback Leibler (KL) divergence

Let p_1 and p_2 be measures over paths for SDEs with drifts $\mathbf{f}_1(\mathbf{x}, t)$ and $\mathbf{f}_2(\mathbf{x}, t)$ with **same diffusion** σ^2 . Then

$$KL [p_1 \| p_2] = \int dp_1 \ln \frac{dp_1}{dp_2} = \frac{1}{2\sigma^2} \int_0^T \langle \|\mathbf{f}_1(\mathbf{x}, t) - \mathbf{f}_2(\mathbf{x}, t)\|^2 \rangle_{p_1, t} dt$$

The expectation at the right is over the **marginal density** of \mathbf{x} at time t w.r.t. p_1 .

The KL divergence cont'd

$$KL [q||p_{post}] = \int dq \ln \frac{dq}{dp} = \int_0^T \mathcal{L}(t) dt + \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathbf{Q}| + \ln Z,$$

with

$$\mathcal{L}(t) = \mathcal{L}_{sde}(t) + \mathcal{L}_{obs}(t)$$

and

$$\begin{aligned} \mathcal{L}_{sde}(t) &= \frac{1}{2\sigma^2} \left\langle \|\mathbf{f}_1(\mathbf{x}) - \mathbf{f}_L(\mathbf{x}, t)\|^2 \right\rangle_{qt}, \\ \mathcal{L}_{obs}(t) &= \frac{1}{2\sigma_o^2} \sum_{n=1}^N \left\langle \|\mathbf{y}_n - \mathbf{x}(t_n)\|^2 \right\rangle_{qt_n} \delta(t - t_n), \end{aligned}$$

$KL [q||p_{post}] \geq 0$ yields an upper bound on $-\ln Z = -\ln p(\mathbf{y}_{1:N})$.

Consistency

Establish consistency between marginals

$q_t(\mathbf{x}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ and conditionals $p(\mathbf{x}_{t+\Delta t}|\mathbf{x}_t)$

defined through $\mathbf{A}(t)$ and $\mathbf{b}(t)$.

$$\begin{aligned}\frac{d\mathbf{m}}{dt} &= -\mathbf{A}(t)\mathbf{m} + \mathbf{b}(t) \\ \frac{d\mathbf{S}}{dt} &= -\mathbf{A}(t)\mathbf{S} - \mathbf{S}\mathbf{A}^\top(t) + \sigma^2\mathbf{I}.\end{aligned}$$

The covariance kernel of the GP

$\mathbf{K}(t, t') = \langle \mathbf{x}(t)\mathbf{x}^T(t') \rangle_q - \mathbf{m}(t)\mathbf{m}(t')$ is the solution to the ODE

$$\left(-\partial_t^2 - (\mathbf{A} - \mathbf{A}^T)\partial_t - \frac{d\mathbf{A}}{dt} + \mathbf{A}^T\mathbf{A} \right) \mathbf{K}(t, t') = \sigma^2\delta(t - t')$$

Lagrange function

Include constraints using Lagrange multipliers $\Psi(t)$ and $\lambda(t)$

$$\mathcal{A} = \int_0^T \left\{ \mathcal{L}(t) - \text{tr} \left\{ \Psi^\top(t) \left(\frac{d\mathbf{S}}{dt} + \mathbf{A}\mathbf{S} + \mathbf{S}\mathbf{A}^\top - \sigma^2\mathbf{I} \right) \right\} - \lambda^\top(t) \left(\frac{d\mathbf{m}}{dt} + \mathbf{A}\mathbf{m} - \mathbf{b} \right) \right\} dt$$

Variational equations by functional derivatives of \mathcal{A} with respect to $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{S}(t)$ and $\mathbf{m}(t)$.

The full solution ($D = 1$)

Compute the drift g for the posterior process. Perform variation on

$$\mathcal{A} = \int_0^T \left\{ \mathcal{L}(t) - \lambda(x, t) \left(\partial_t + \partial_x g - \frac{\sigma^2}{2} \partial_x^2 \right) q(x, t) \right\} dt$$

$q(x, t)$ is the marginal at time t . We find (equivalent to Eyink (2002)):

$$\begin{aligned} g &= f + \partial_x \ln r && \text{drift} \\ \left(\partial_t + f \partial_x + \frac{\sigma^2}{2} \partial_x^2 \right) r(x, t) &= 0 && \text{backward eq.} \\ r^+ &= r^- \exp \left[\frac{1}{2\sigma_o^2} \langle (y_i - x(t_i))^2 \rangle_q \right] && \text{jump} \end{aligned}$$

Variational Equations

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} - (\Psi + \Psi^\top) \mathbf{S} - \lambda \mathbf{m}^\top = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} + \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}} - (\mathbf{A}^\top \Psi + \Psi \mathbf{A}) + \frac{d\Psi}{dt} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} - \mathbf{A}^\top \lambda + \frac{d\lambda}{dt} = 0$$

We can express *variational parameters* $\mathbf{A}(t)$ and $\mathbf{b}(t)$ by Lagrange multipliers $\Psi(t)$ and $\lambda(t)$.

$$\mathbf{A}(t) = - \left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right\rangle_{q_t} + \sigma^2 (\Psi + \Psi^\top),$$

$$\mathbf{b}(t) = \langle \mathbf{f}(\mathbf{x}) \rangle_{q_t} + \mathbf{A}(t) \mathbf{m} - \sigma^2 \lambda(t),$$

Smoothing algorithm

Make initial guess for $\mathbf{A}(t), \mathbf{b}(t)$. Repeat until sufficient accuracy is achieved:

1. Solve moment equations

$$\begin{aligned}\frac{d\mathbf{m}}{dt} &= -\mathbf{A}(t)\mathbf{m} + \mathbf{b}(t) \\ \frac{d\mathbf{S}}{dt} &= -\mathbf{A}(t)\mathbf{S} - \mathbf{S}\mathbf{A}^\top(t) + \sigma^2\mathbf{I}.\end{aligned}$$

forward in time.

2. With $\mathbf{m}(t)$ and $\mathbf{S}(t)$ found for $0 \leq t \leq T$, solve **backward** in time with $\Psi(T) = \mathbf{0}$ and $\lambda(T) = \mathbf{0}$:

$$\begin{aligned}\frac{d\Psi}{dt} &= (\Psi^\top + \Psi)\mathbf{A} - \frac{\partial \mathcal{L}_{sde}}{\partial \mathbf{S}}, \\ \frac{d\lambda}{dt} &= \mathbf{A}^\top \lambda - \frac{\partial \mathcal{L}_{sde}}{\partial \mathbf{m}}.\end{aligned}$$

3. For observations use jump-conditions:

$$\begin{aligned}\Delta\Psi &= -\frac{1}{2\sigma_o^2} \mathbf{I} \\ \Delta\lambda &= \frac{1}{2\sigma_o^2} (\mathbf{y}_n - \mathbf{m}(t_n))\end{aligned}$$

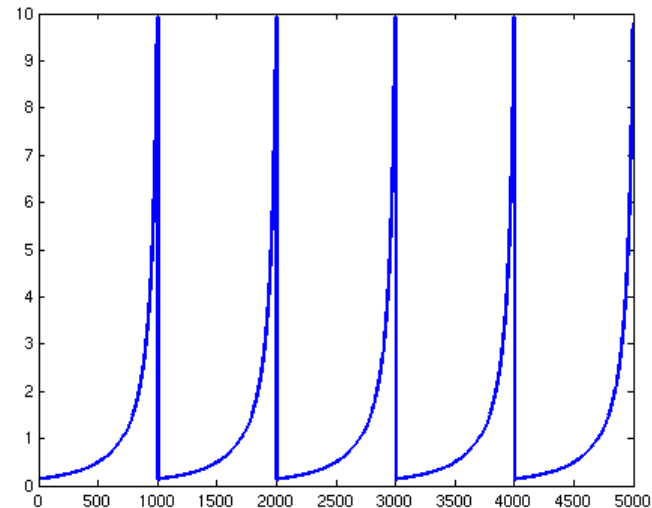
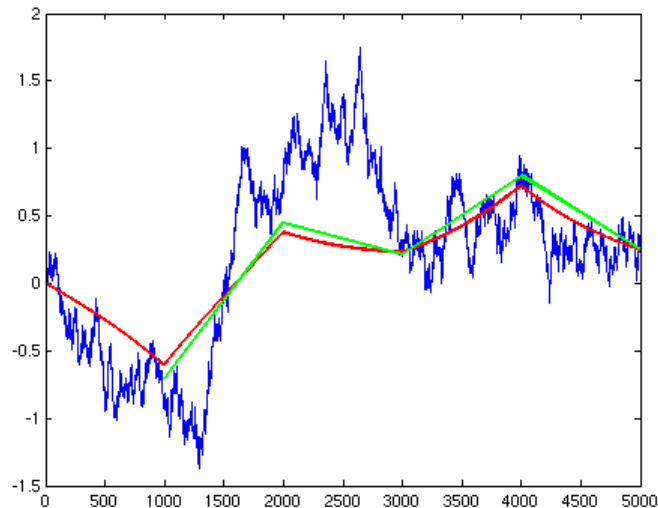
Ornstein-Uhlenbeck process

$$dx = -\gamma x dt + \sigma^2 dW,$$

Prior process is GP with the nonstationary kernel (fixed initial condition $x(0)$)

$$K(t, t') = \frac{\sigma^2}{2\gamma} \left\{ e^{-\gamma|t-t'|} - e^{-\gamma(t+t')} \right\}.$$

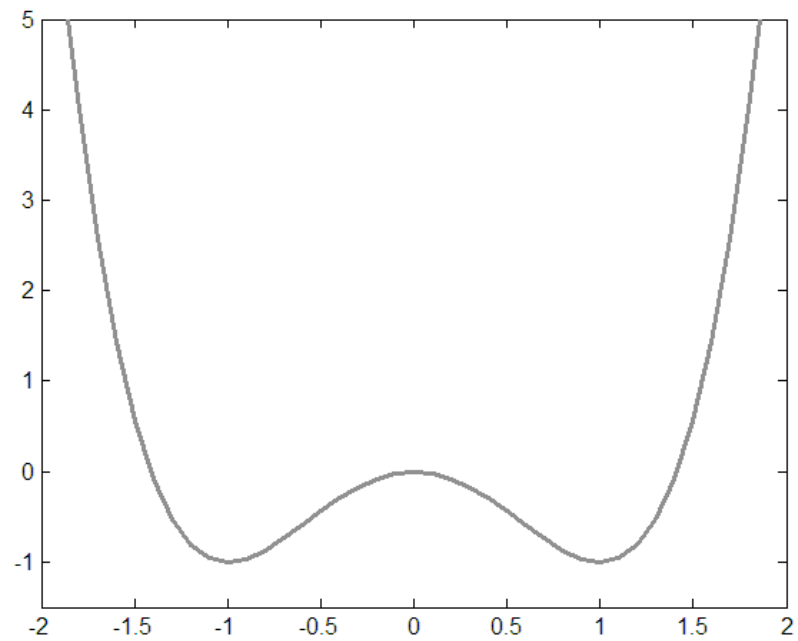
left: 5 observations with predictions **right:** $\Psi(t)$



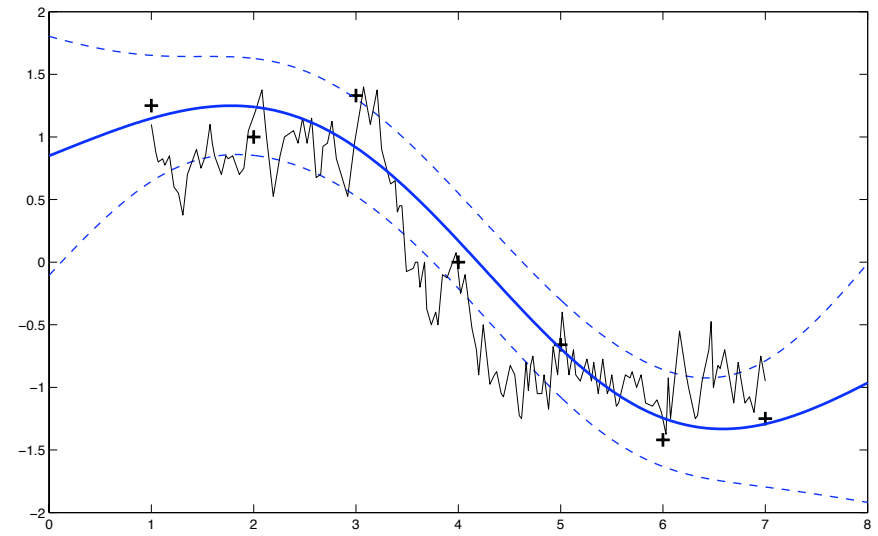
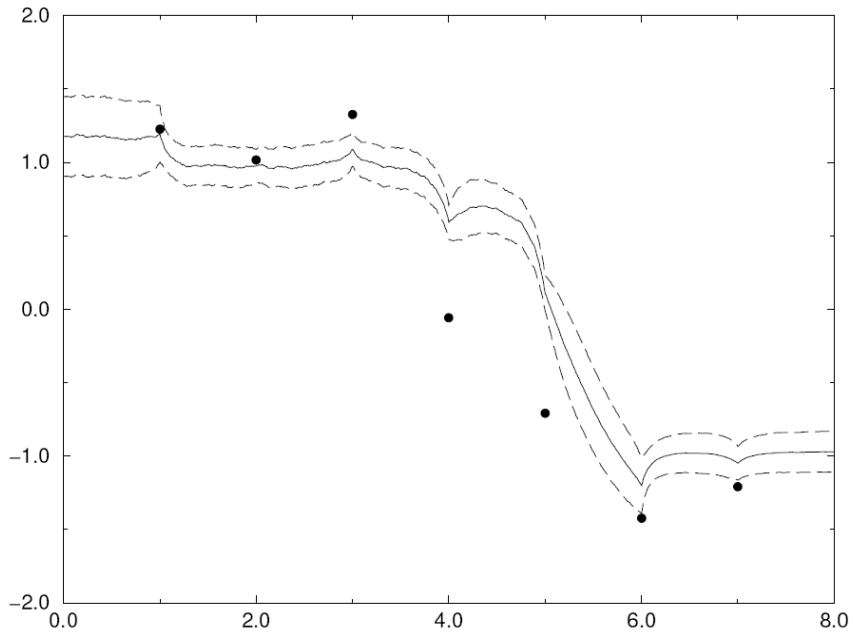
Motion in double-well potential

$$dx = 4x(1 - x^2)dt + \sigma^2 dW.$$

Potential



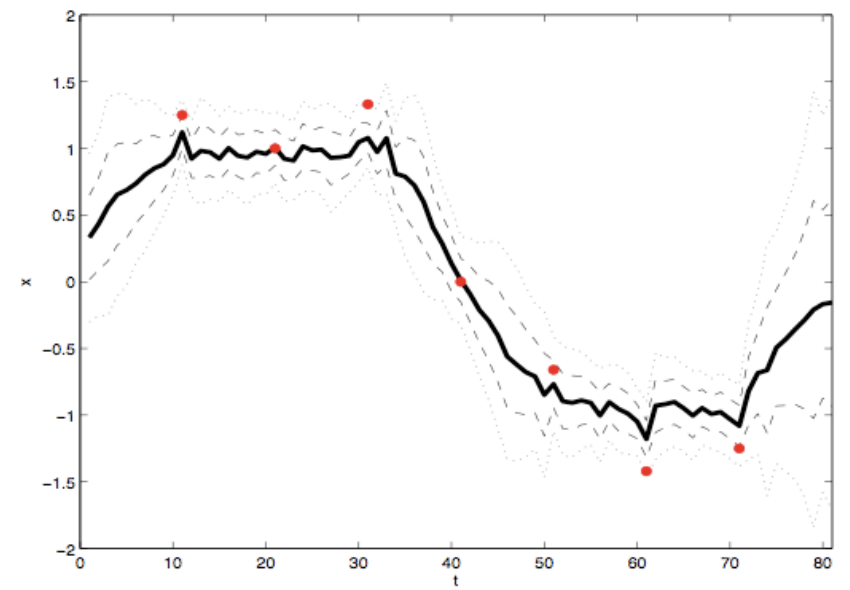
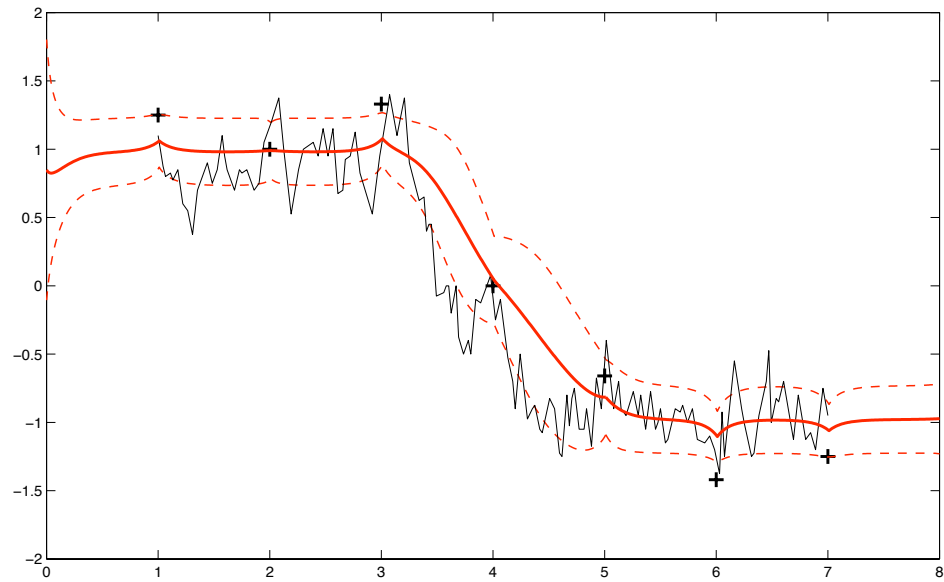
Prediction with ensemble Kalman smoother and GP (RBF kernel)

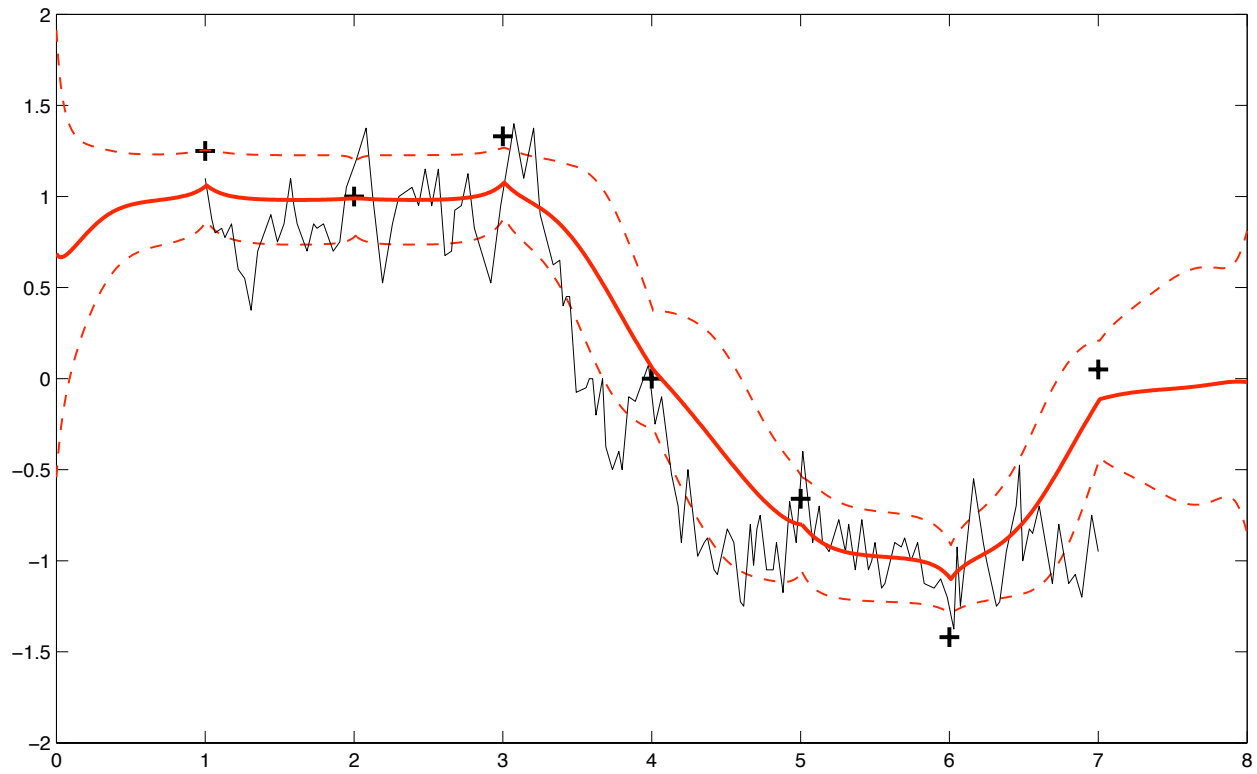


ODEs for Lagrange multipliers

$$\begin{aligned}\frac{d\psi}{dt} &= -8\psi(1 - 3m^2 - 6s^2) - \frac{144}{\sigma^2}\{2m^2 + s^2\}s^2 + 12m\lambda + 2\psi^2\sigma^2, \\ \frac{d\lambda}{dt} &= -4(1 - 3m^2 - 3s^2)\lambda - \frac{288}{\sigma^2}ms^4 + 48\psi ms^2.\end{aligned}$$

Variational result and comparison to MCMC





A Hamiltonian approach for the 'potential' case

$$\mathcal{A}_{sde} = \int_0^T \mathcal{L}_{sde} dt \quad \text{with} \quad \mathcal{L}_{sde} = \frac{1}{2\sigma^2} \langle \|\mathbf{f} + \mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \rangle$$

Try to express \mathbf{A} and \mathbf{b} by \mathbf{S} and \mathbf{m} using

$$\frac{d\mathbf{m}}{dt} = -\mathbf{A}\mathbf{m} + \mathbf{b} \quad \frac{d\mathbf{S}}{dt} = -\mathbf{A}\mathbf{S} - \mathbf{S}\mathbf{A}^T + \sigma^2\mathbf{I}$$

Possible if $\mathbf{f}(\mathbf{x}) = -\nabla U(\mathbf{x})$ and after some calculations

$$2\sigma^2\mathcal{L} = \left(\frac{d\mathbf{m}}{dt}\right)^2 + \langle \|\mathbf{f}\|^2 \rangle + \sigma^2 \langle \nabla \mathbf{f} \rangle + 2\frac{d\langle U \rangle}{dt} + \text{trace}(\mathbf{A}^2\mathbf{S})$$

We need to eliminate \mathbf{A}^2 .

The 1 - D case

Set $\mathbf{S} \rightarrow s^2$ (variance). Then

$$\frac{ds}{dt} = -\alpha s + \frac{\sigma^2}{2s} \quad \rightarrow \quad \alpha = \frac{\sigma^2}{2s^2} - \frac{1}{s} \frac{ds}{dt}$$

Thus the action is

$$\mathcal{A} = \frac{1}{2\sigma^2} \int_0^T \left\{ \left(\frac{dv}{dt} \right)^2 + \left(\frac{dm}{dt} \right)^2 - V_{eff}(m, s) \right\} dt + \text{data and surface terms}$$

with

$$V_{eff}(m, s) = - \left(\frac{\sigma^4}{4s^2} + \langle f^2(x) \rangle + \sigma^2 \langle f'(x) \rangle \right)$$

This shouldn't be too surprising:

The 'path probabilities' of an SDE are related to the Wiener measure via

$$\ln \frac{dp}{d\mu} = \frac{1}{\sigma^2} \int_0^T \mathbf{f}^\top \circ d\mathbf{x} - \frac{1}{2\sigma^2} \int_0^T \|\mathbf{f}\|^2 dt - \frac{1}{2} \int_0^T \nabla \mathbf{f} dt$$

(*Stratonovich* form). Hence, if $\mathbf{f} = \nabla U$

$$\ln \frac{dp}{d\mu} = \frac{1}{\sigma^2} (U(0) - U(T)) - \frac{1}{2\sigma^2} \int_0^T \|\mathbf{f}\|^2 dt - \frac{1}{2} \int_0^T \nabla \mathbf{f} dt$$

The Hamilton equations

Between observations we have the Hamiltonian flow

$$\begin{aligned} 2\frac{d^2m}{dt^2} &= -\frac{\partial V_{eff}(m, s)}{\partial m} \\ 2\frac{d^2s}{dt^2} &= -\frac{\partial V_{eff}(m, s)}{\partial s} \end{aligned}$$

Data and surface terms

$$\mathcal{A}_{data} = \frac{1}{2\sigma_0^2} \sum_i \left\{ s^2(t_i) + (y_i - m(t_i))^2 \right\} - \frac{1}{2} \ln \frac{s(T)}{s(0)} + \frac{1}{\sigma^2} (\langle U(T) \rangle - \langle U(0) \rangle)$$

lead to jump conditions for momenta

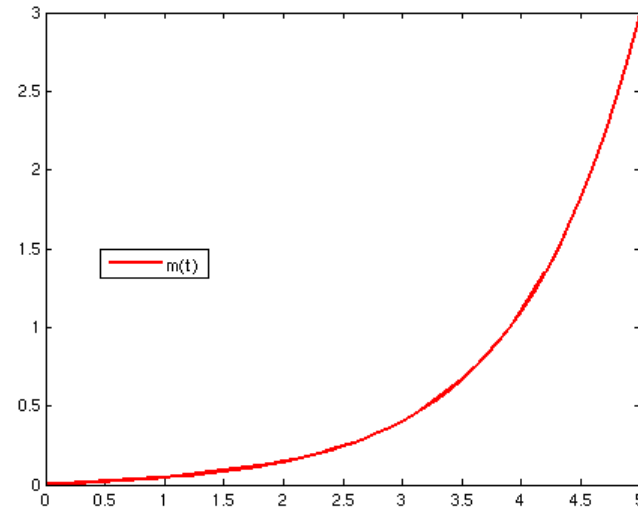
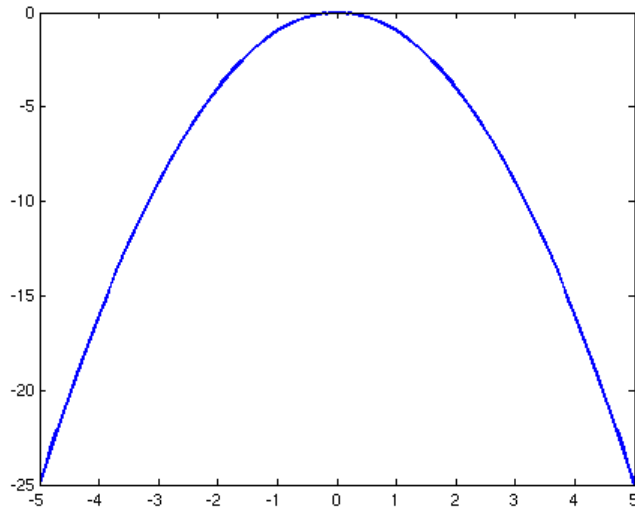
$$\begin{aligned} \frac{dm}{dt}(t_i^+) - \frac{dm}{dt}(t_i^-) &= \frac{m(t_i) - y_i}{2\sigma_0^2} \\ \frac{ds}{dt}(t_i^+) - \frac{ds}{dt}(t_i^-) &= \frac{1}{2\sigma_0^2} s(t_i) \end{aligned}$$

Ornstein - Uhlenbeck Process

$$f = -\gamma x \rightarrow V_{eff}(m, s) = -\frac{\sigma^4}{4s^2} - \gamma^2 s^2 - \gamma^2 x^2 + \sigma^2 \gamma$$

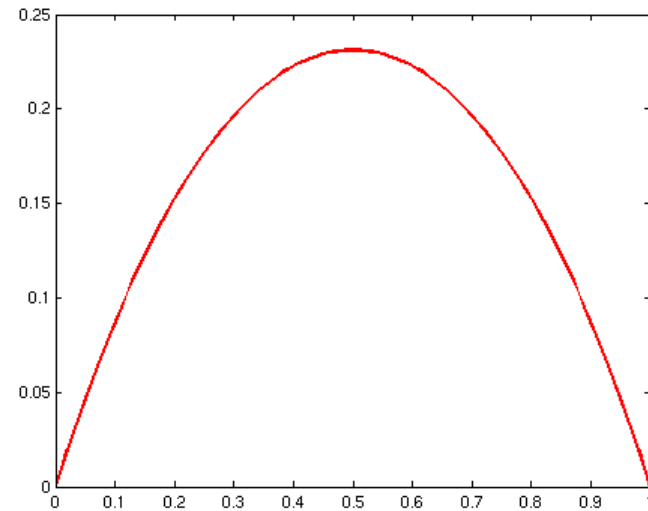
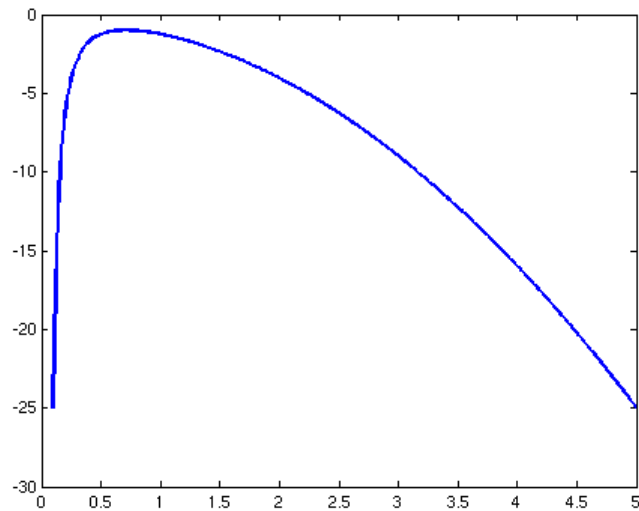
The potential separates in m and s !

Potential $V_{eff}(m)$ and **Prediction** $m(t) = y \frac{\sinh(\gamma t)}{\sinh(\gamma T)}$ for perfect observation $x(T = 5) = 3$.



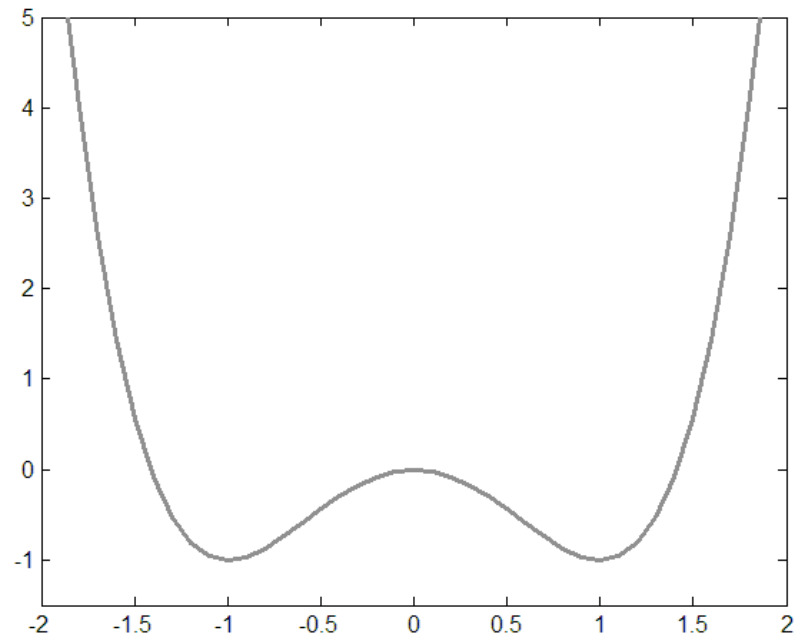
Potential $V_{eff}(s)$ and predictive variance $s^2(t)$ when $y = x(1)$ perfectly observed:

$$s^2(t) = \frac{\sigma^2}{2\gamma \sinh(\gamma)} (\cosh(\gamma) - \cosh(2\gamma t - \gamma))$$



Motion in double-well potential

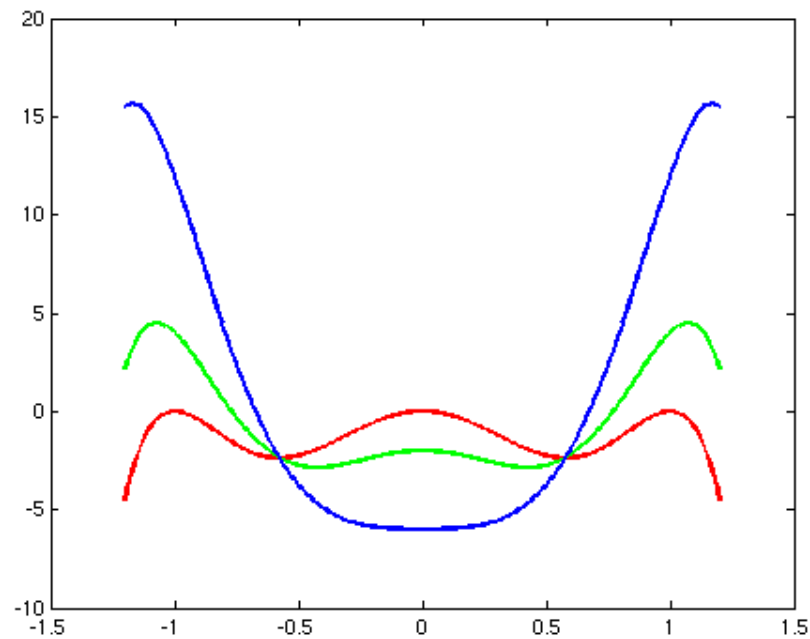
$f(x) = 4x(1 - x^2)$ with potential $U(x)$



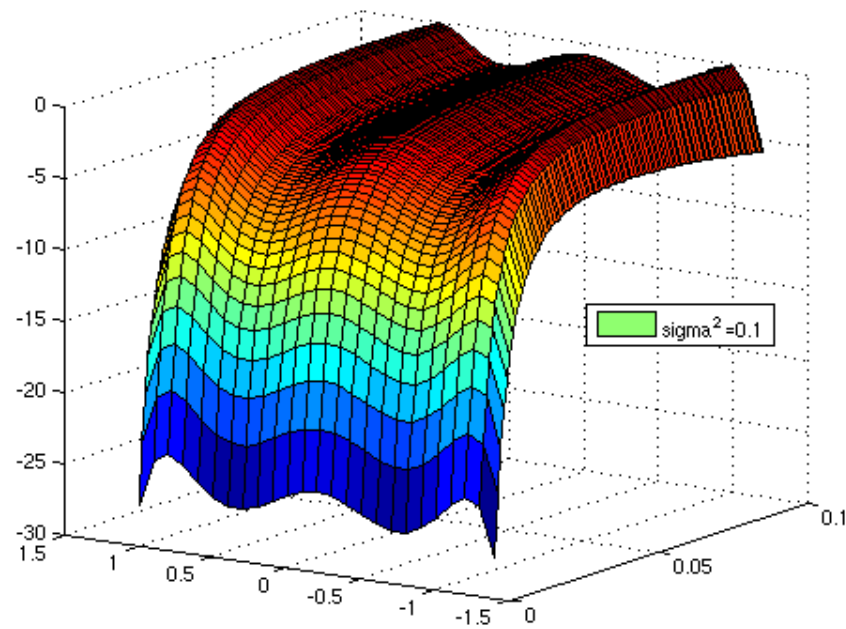
Effective potential

$$V_{eff}(m, s) = -\frac{\sigma^4}{4s^2} - (16m^6 - 32m^4 + 16m^2) - (240m^4 - 192m^2 + 16)s^2 - (720m^2 - 96)s^4 - 240s^6 - \sigma^2(4 - 12m^2 - 12s^2)$$

The nonsingular part of $V_{eff}(m, s = 0)$ for $\sigma^2 = 0, 0.5, 1.5$:



$V_{eff}(m, s)$ for $\sigma^2 = 0.1$



What about $D > 1$

- It is possible to solve

$$\frac{d\mathbf{S}}{dt} = -\mathbf{A}(t)\mathbf{S} - \mathbf{S}\mathbf{A}^T + \sigma^2\mathbf{I}.$$

for \mathbf{A} , when \mathbf{A} is symmetric: linear system with $D(D + 1)/2$ unknowns.

- For $\mathbf{A} \neq \mathbf{A}^T$ there are more unknowns than equations!

Future work

- Convergent & efficient numerical approaches
- Good suboptimal variational *ansätze*.
- Multiplicative noise
- Other processes
- PAC Bayesian bounds
- Importance sampling
- Perturbative corrections

With some more effort

Apply Girsanov to $f = -\mathbf{A}\mathbf{x}$:

$$\ln \frac{dP}{d\mu} = \frac{1}{2\sigma^2} \int \mathbf{x}^T \left(\frac{d\mathbf{A}}{dt} - \mathbf{A}^T \mathbf{A} + (\mathbf{A} - \mathbf{A}^T) \partial_t \right) \mathbf{x} dt \\ - \frac{1}{2} \mathbf{x}^T \mathbf{A}^T \mathbf{x} \Big|_0^T + \frac{1}{2} \int_0^T \text{trace}(\mathbf{A}) dt$$

Suggest approximations where e.g. $\frac{d\mathbf{A}}{dt} - \mathbf{A}^T \mathbf{A} = \text{constant!}$