Approximate inference

for continuous time Markov processes

Manfred Opper

TU Berlin, Dept of Computer Science

Collaborators:

- Andreas Ruttor & Florian Stimberg (TU Berlin)
- Cédric Archambeau & John Shawe–Taylor (UCL)
- Dan Cornford, Yuan Shen & Michail Vrettas (Aston)
- Guido Sanguinetti & Andrea Ocone (Edinburgh)

Ito stochastic differential equations

for state $X_t \in R^d$

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

Limit of discrete time process X_k

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

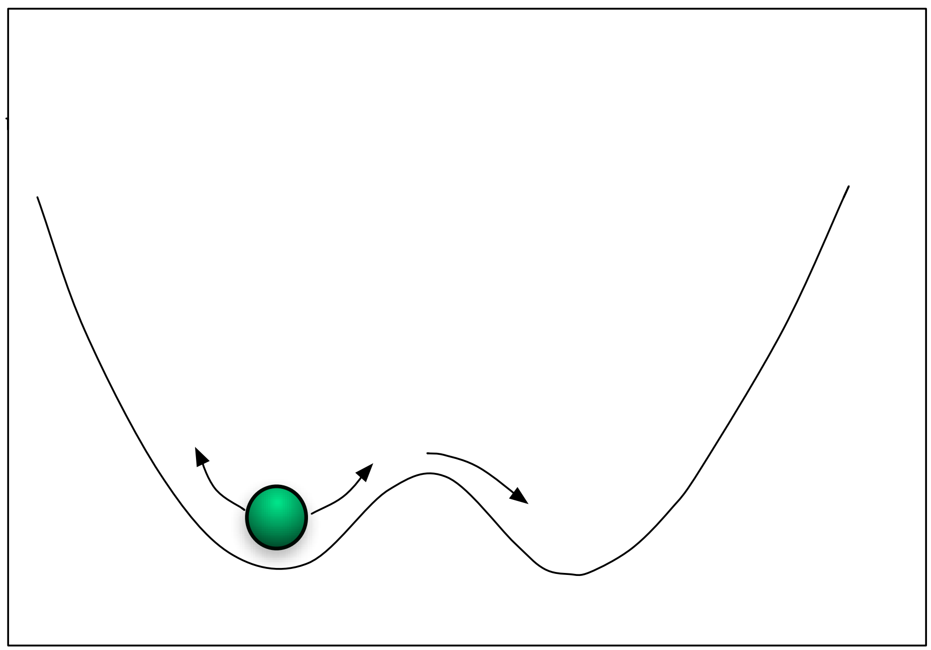
ϵ_k i.i.d. Gaussian.

Overview

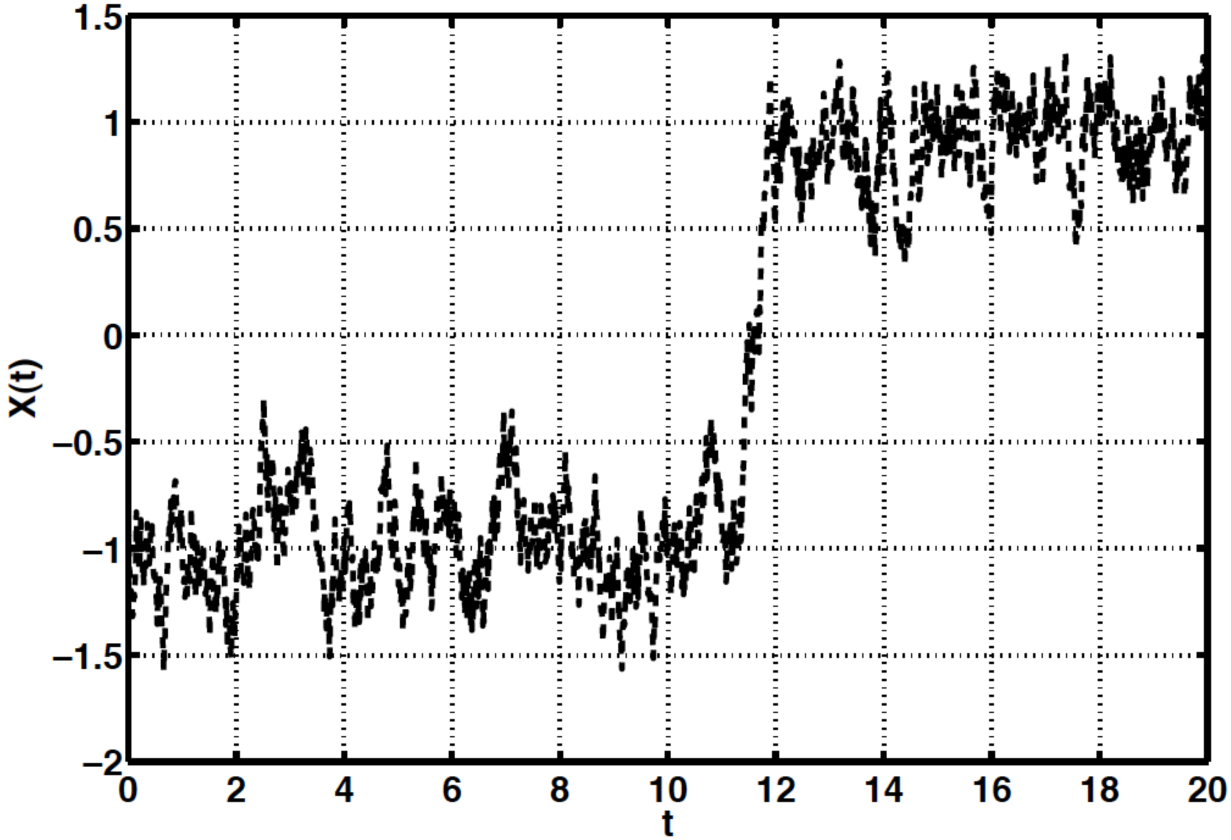
- Inference for stochastic dynamics
- Variational approximation in machine learning and physics
- Formulation for probabilities over paths
- Results for low dimensional models
- Hybrid models
- Nonparametric approach to drift estimation

Motion in double-well potential

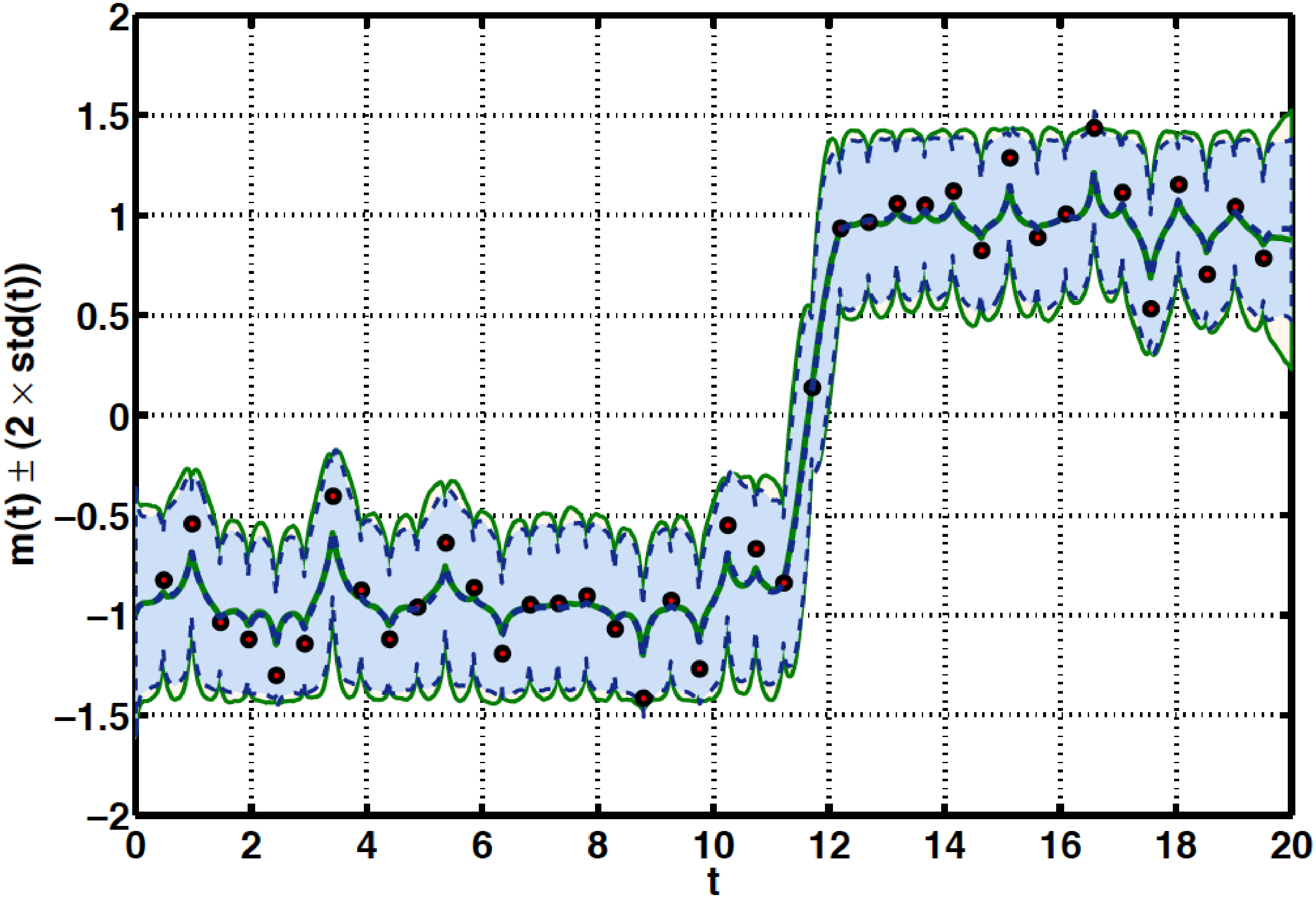
$$dX = X(\theta - X^2)dt + \sigma dW.$$



A sample path might look like this



Optimal state prediction



Jump Processes

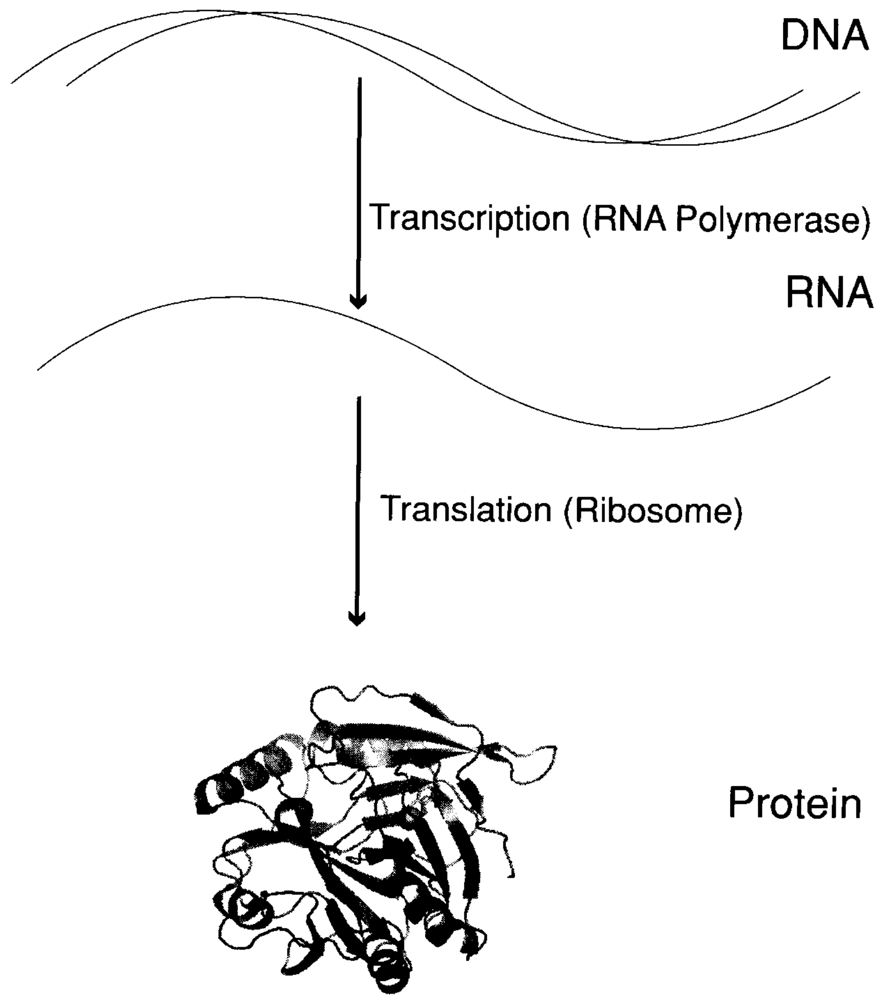
Assume that X_t jumps between discrete states.

Short time behaviour of transition kernel defined by **transition rate** f:

$$
\begin{aligned}
P_{t+\Delta t,t}(x'|x) &\simeq f_\theta(y|x,t)\Delta t \text{ for } x' \neq x \\
P_{t+\Delta t,t}(x|x) &\simeq 1 - \sum_{z\neq x} f_\theta(z|x,t)\Delta t
\end{aligned}
$$

for $\Delta t \to 0$.

Gene expression



Simple model of autoregulatory network

On molecular level kinetics is stochastic: Simple autoregulatory network:

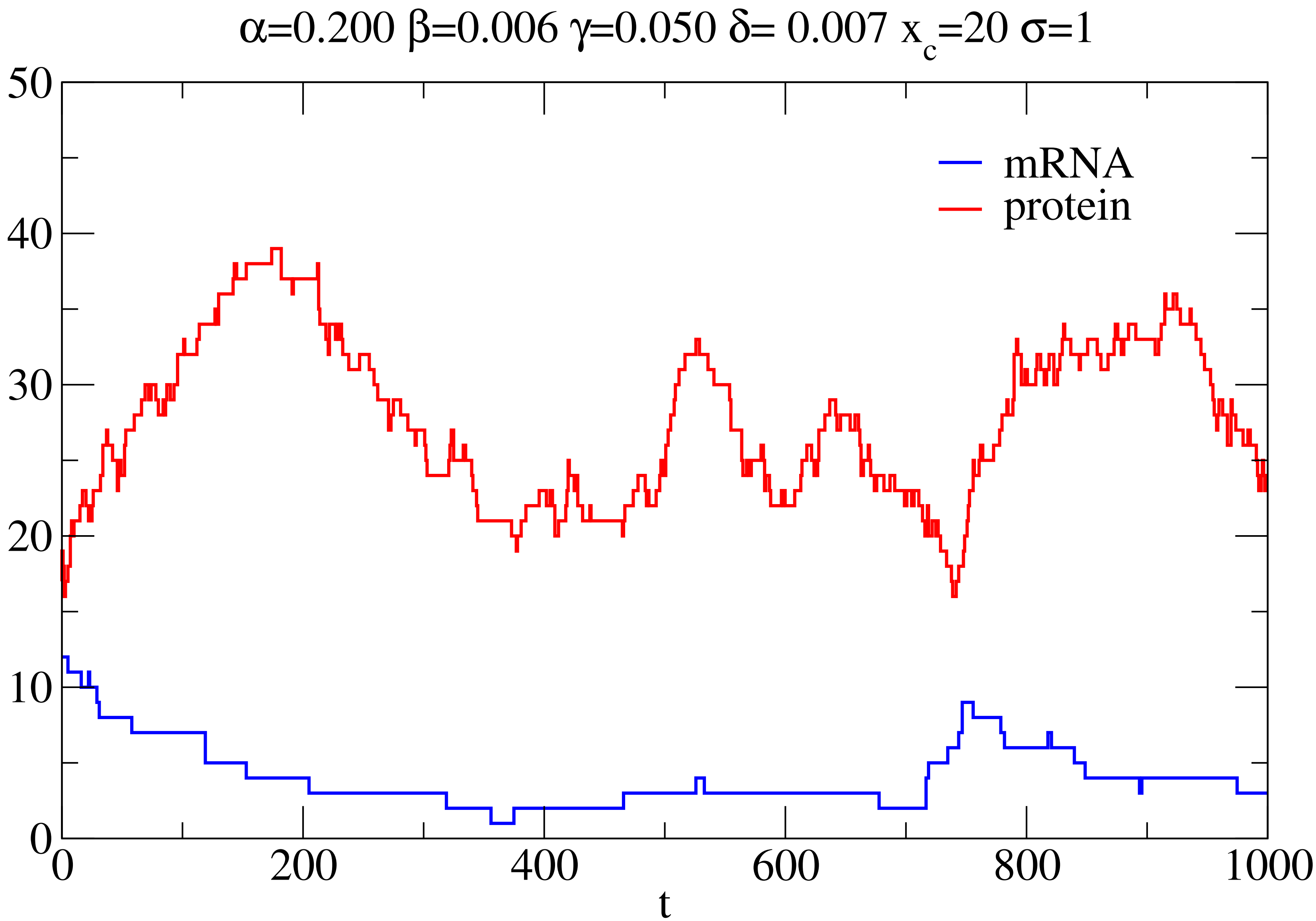
2 interacting molecules: mRNA and a Protein

Number of mRNA and Protein molecules: X_{mRNA}, X_{Prot}

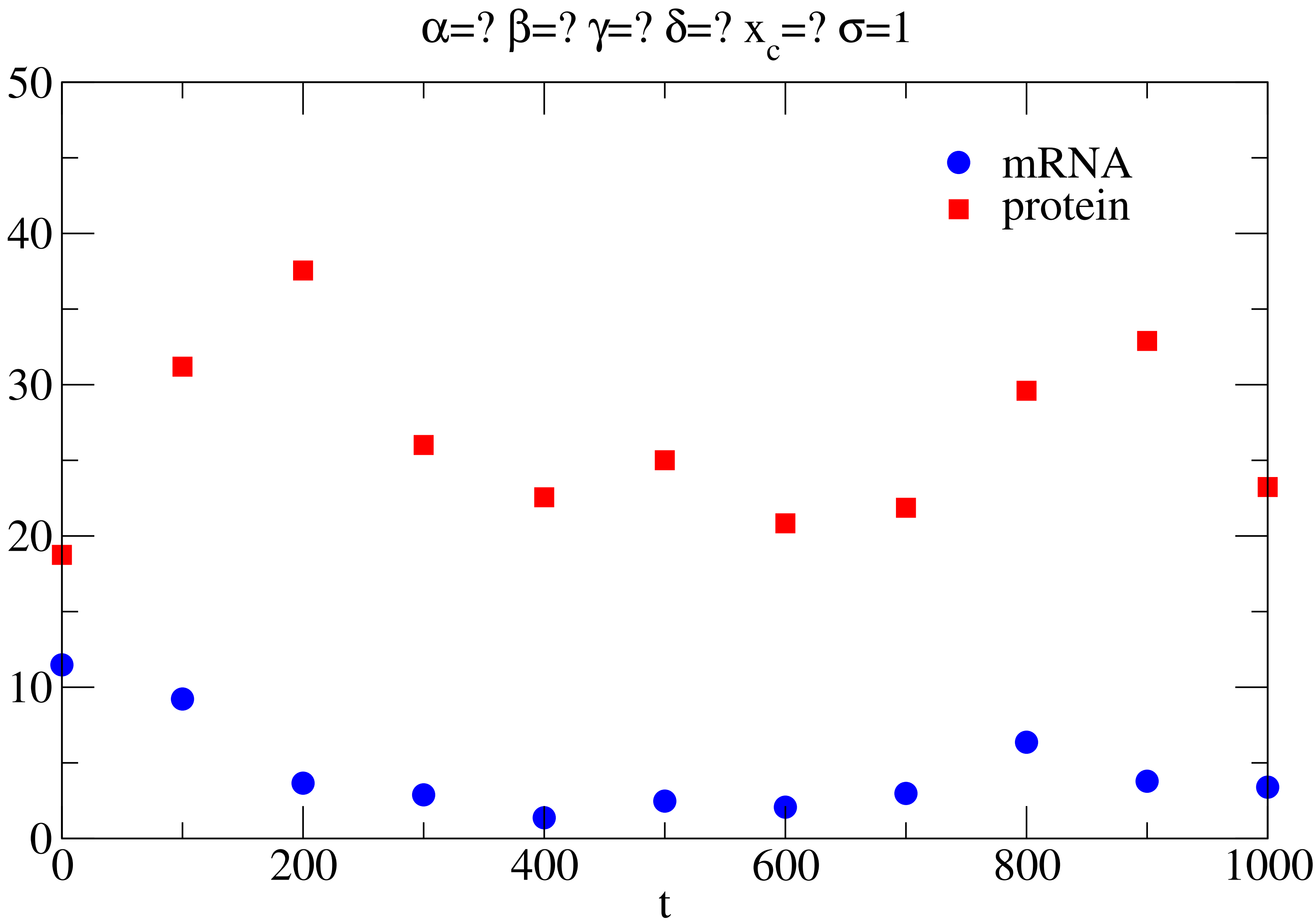
$\mathbf{X} = (X_{\text{mRNA}}, X_{\text{Prot}})$

$$
\begin{array}{rcl@{\quad}c@{\quad}l@{\quad}l}
X_{\text{Prot}} & \rightarrow & X_{\text{Prot}} + 1 & : & \text{with Rate} & \gamma X_{\text{mRNA}}\,, \\
X_{\text{Prot}} & \rightarrow & X_{\text{Prot}} - 1 & : & \text{with Rate} & \delta X_{\text{Prot}}\,, \\
X_{\text{mRNA}} & \rightarrow & X_{\text{mRNA}} - 1 & : & \text{with Rate} & \beta X_{\text{mRNA}}\,, \\
X_{\text{mRNA}} & \rightarrow & X_{\text{mRNA}} + 1 & : & \text{with Rate} & \alpha(1 - \alpha_c \Theta(X_{\text{Prot}} - \theta_c))\,,
\end{array}
$$

where $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$.



Simulation of process



(Noisy) observations at discrete times.

Inference Problems

Given **noisy observations** $\{y_i\}_{i=1}^N \equiv y_1, \ldots, y_N$ of **hidden process** X_{t_i} at times $t_i \le T$ for $i = 1, \ldots, N$.

- Estimate X_t for $0 \le t \le T$ (**smoothing**).

- Estimate **system parameters** θ contained in drift f and diffusion Σ.

Obvious ? questions

- Can't we treat this just as a **discrete time** HMM ?

  Yes, but ....

- Isn't there some simple **forward backward** algorithm ?

  Yes, but ....

- Can't you just discretize in time and run an MCMC sampler ?

  Yes, but ....

What we would like to do

- **State estimation:**

  Use **Bayes rule** for conditional distribution over **paths** $X_{0:T}$ (∞ dimensional objects)

$$p(X_{0:T}|\{y_i\}_{i=1}^N, \theta) = \frac{p_{prior}(X_{0:T}|\theta)}{p(\{y_i\}_{i=1}^N|\theta)} \prod_{n=1}^N p(y_n|X_{t_n})$$

  to compute **state prediction** $E[X_t|\{y_i\}_{i=1}^N, \theta]$

- **Parameter estimation**:

  1. Maximum Likelihood: Maximise $p(\{y_i\}_{i=1}^N|\theta)$ with respect to θ

  2. Bayes: Use a prior $p(\theta)$ to compute $p(\theta|\{y_i\}_{i=1}^N) \propto p(\{y_i\}_{i=1}^N|\theta)p(\theta)$

Path integral representation

of the parameter likelihood (assuming additive noise)

$$p(\{y_i\}_{i=1}^{N}|\theta) =$$

$$\int \mathcal{D}[X_t]\ \mathrm{exp}\left[-\int_0^T \left\{\frac{1}{2\sigma_\theta^2}\left(\frac{dX_t}{dt} - f(X_t)\right)^2 - \sum_n \delta(t-t_n)\ln p(y_n|X_t)\right\}\ dt\right]$$

with **Onsager–Machlup** type action. One needs to be a bit careful about the correct interpretation of '$\frac{dX_t}{dt}$' and integrals.

The variational approximation in statistical physics

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

Let $p(x) = \frac{1}{Z}\, e^{-H(x)}$ and $q(x) = \frac{1}{Z_0}\, e^{-H_0(x)}$

- The variational bound on the free energy is

$$-\ln Z \leq -\ln Z_0 + \langle H(x) \rangle_0 - \langle H_0(x) \rangle_0 \equiv \mathcal{F}[q]$$

  $\mathcal{F}[q]$ is the variational free energy.

- Approximation for free energies often better than the quality of H_0 suggests.

- Choices for H_0

    1. Gaussian approximations for path integrals (e.g. Polaron problem)

    2. Mean field approximations (factorising distributions)

- Look for a formulation that can easily be applied to a variety of systems without bothering too much about details of path integral formulations.

The variational approximation (reformulation)

- We would like to approximate intractable distribution

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

by a $q(x)$ which belongs to a family of **simpler tractable** distributions (e.g. factorising = mean field, or Gaussian densities).

- The variational free energy is

$$\begin{aligned}\mathcal{F}(q) &= D[q\|p(\cdot|y)] - \ln p(y) \\ &= D[q\|p_{prior}] - \int q(x) \ln p(y|x)\, dx \\ &\geq -\ln p(y)\end{aligned}$$

- The relative entropy (Kullback–Leibler divergence) is

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

Approximate maximum likelihood estimate

Assume model depends on parameter θ. The free energy inherits the dependency.

Let $q^*(\theta) = \text{argmin}_q \mathcal{F}_\theta(q)$. Since

$$-\ln p(y|\theta) \leq \mathcal{F}_\theta(q^*(\theta))$$

we can minimise $\mathcal{F}_\theta(q^*)$ wrt θ to get an approximate maximum likelihood estimate.

How to choose the measure q for stochastic differential equations ?

- Process conditioned on data is Markovian!

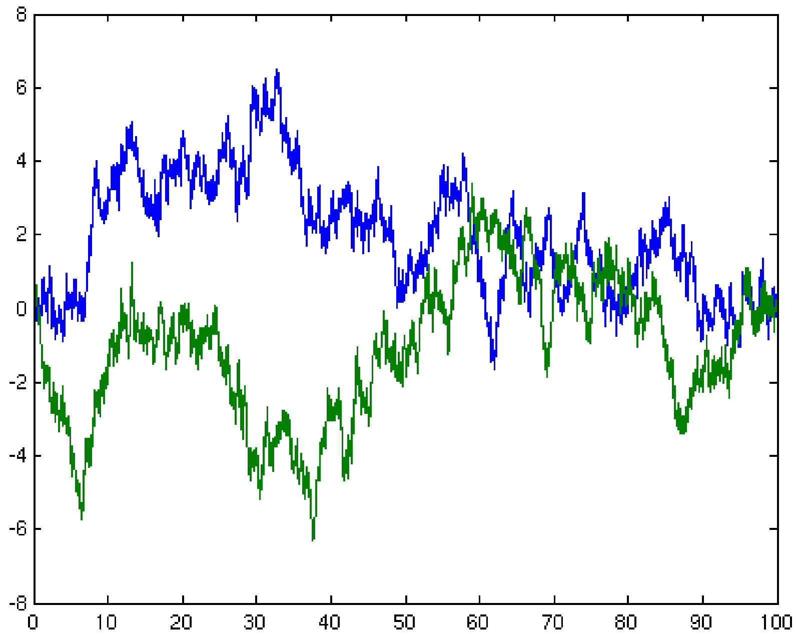
- It fulfils SDE

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

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

Example

Wiener process with single, noise free observation $y = x(t = T) = 0$



Posterior drift $g(x,t) = -\frac{x}{T-t}$ for $0 < t < T$.

KL divergence for path probabilities

Use representation of joint density in term of conditionals and the Markov property (assuming $q_0(x) = p_0(x)$) and work with time discretization $t_{k+1} - t_k = \Delta t$.

$$
\begin{aligned}
D\left[q\|p\right] &= \int dx_{0:T}\, q(x_{0:T}) \ln \frac{q(x_{0:T})}{p(x_{0:T})} \\
&\approx \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) \int dx'\, q_{t_{k+1},t_k}(x'|x) \ln \frac{q_{t_{k+1},t_k}(x'|x)}{p_{t_{k+1},t_k}(x'|x)} \\
&= \sum_{k=0}^{K-1} \int dx\, q_{t_k}(x) D\left[q_{t_{k+1},t_k}(\cdot|x) \| p_{t_{k+1},t_k}(\cdot|x)\right]
\end{aligned}
$$

in terms of transition and marginal probabilities.

We know that short time transition probability

is approximately Gaussian

$$p_{t+\Delta t,t}(x'|x) \propto \exp\left[-\frac{1}{2\Delta t}\left\|x' - x - f(x)\Delta t\right\|^2\right]$$

as $\Delta t \to 0$,

with the squared norm $\|F\|^2 = F^\top \Sigma^{-1} F$.

Then for small Δt

$$D\left[q_{t_{k+1},t_k}(\cdot|x)\|p_{t_{k+1},t_k}(\cdot|x)\right] \approx \frac{1}{2}\|g(x,t) - f(x)\|^2 \Delta t$$

The relative entropy for Stochastic Differential Equations

Let q and p be measures over paths for SDEs with drifts $g(X,t)$ and $f(X,t)$ with **same diffusion** $\Sigma(X)$. Then

$$D\left[q \| p\right] = \frac{1}{2}\int_0^T dt \left\{\int dx\ q_t(x)\ \left\|g(x,t) - f_\theta(x)\right\|^2\right\}$$

$q_t(x)$ is the marginal density of X_t.

Change of measure approach

$$D[Q\|P] = E_Q \ln \frac{dQ}{dP}$$

Girsanov's change of measure theorem results in the following Radon-Nikodym derivative:

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

where B is a Wiener process with respect to Q.

The variational problem (Diffusion)

Minimise variational free energy

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

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

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

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

Variation leads to forward backward PDEs.

The Variational Gaussian Approximation for SDEs

- Approximate (Gaussian) process over paths $X_{0:T}$ induced by linear SDE:

$$dX_t = \{A(t)X_t + b(t)\}\, dt + \Sigma^{1/2} dW$$

- Diffusion Σ must be independent of X !

- Relative entropy is of the form $\mathcal{F}_\theta[m, S, A, b]$.

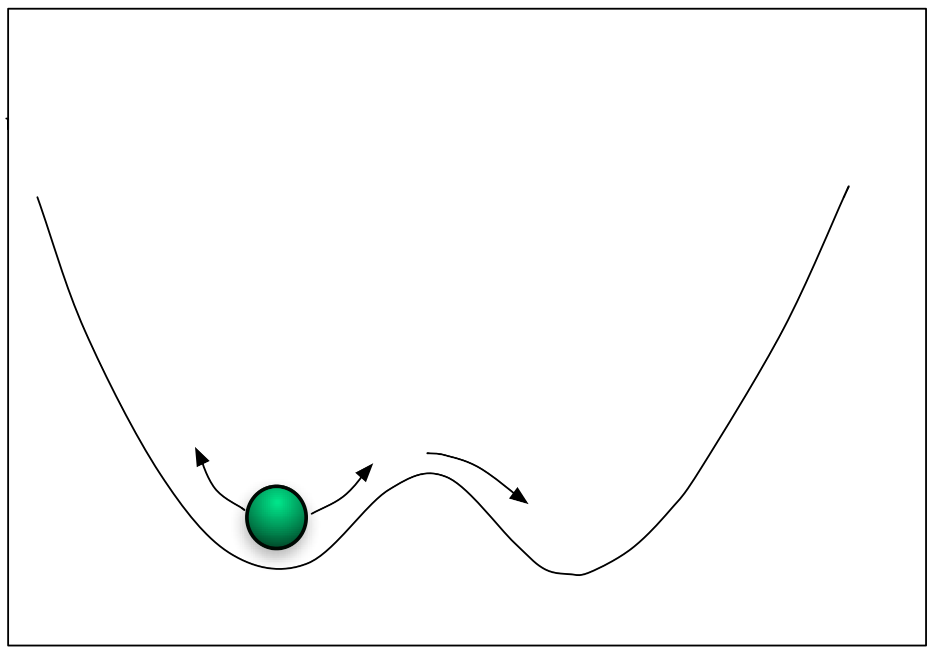
- Constraints are evolution eqs. for marginal **mean** $m(t)$ and **covariance** $S(t)$

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

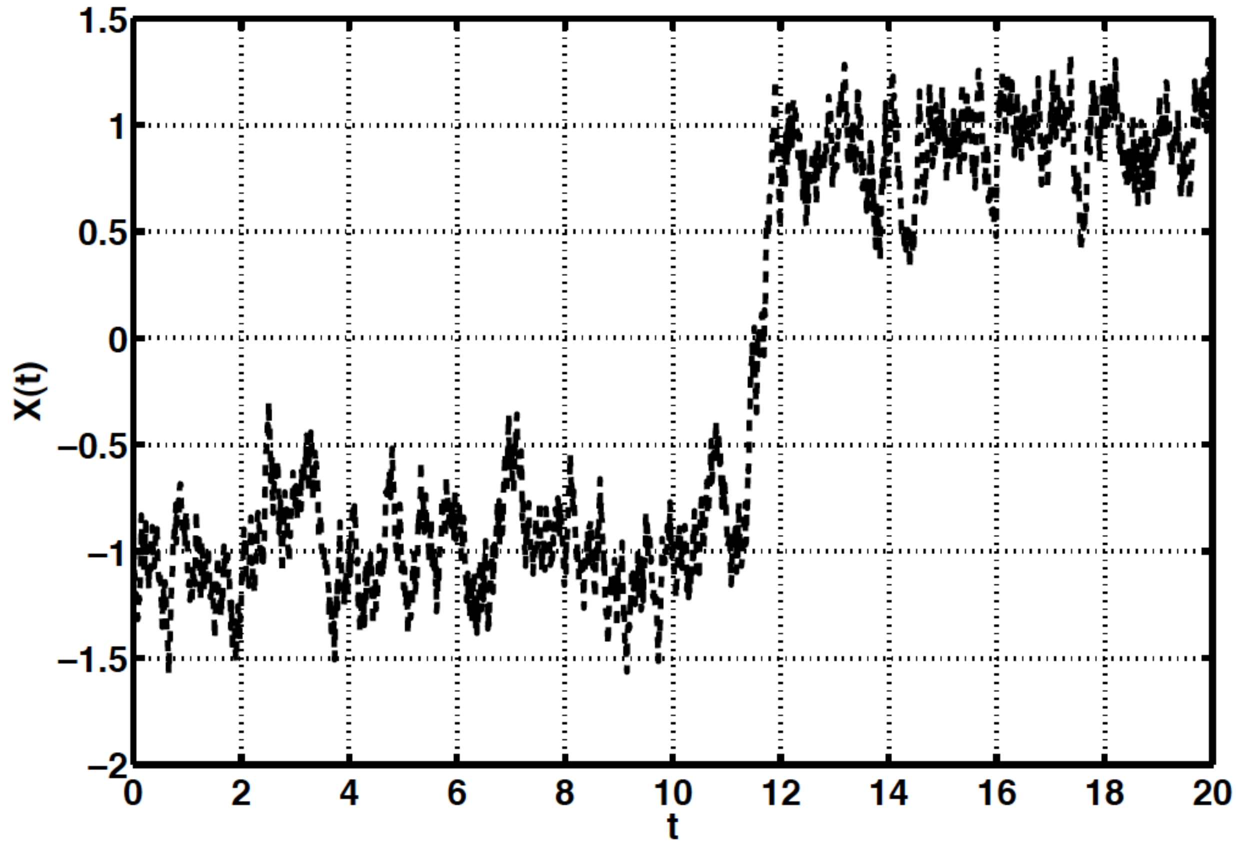
$\rightarrow$ **nonlinear ODEs** instead of PDEs !

Example: Motion in double-well potential

$$dX = X(\theta - X^2)dt + \sigma dW.$$

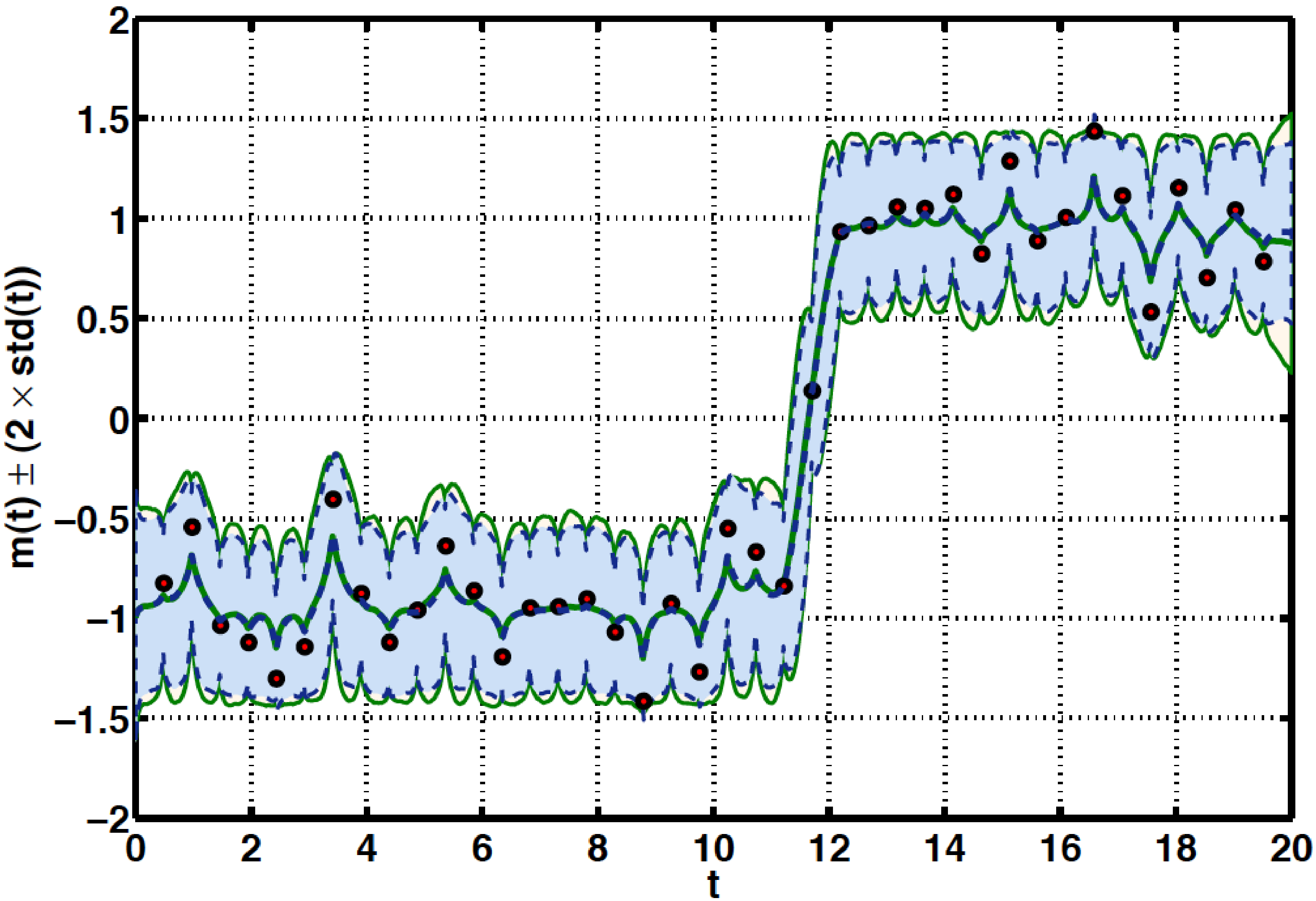


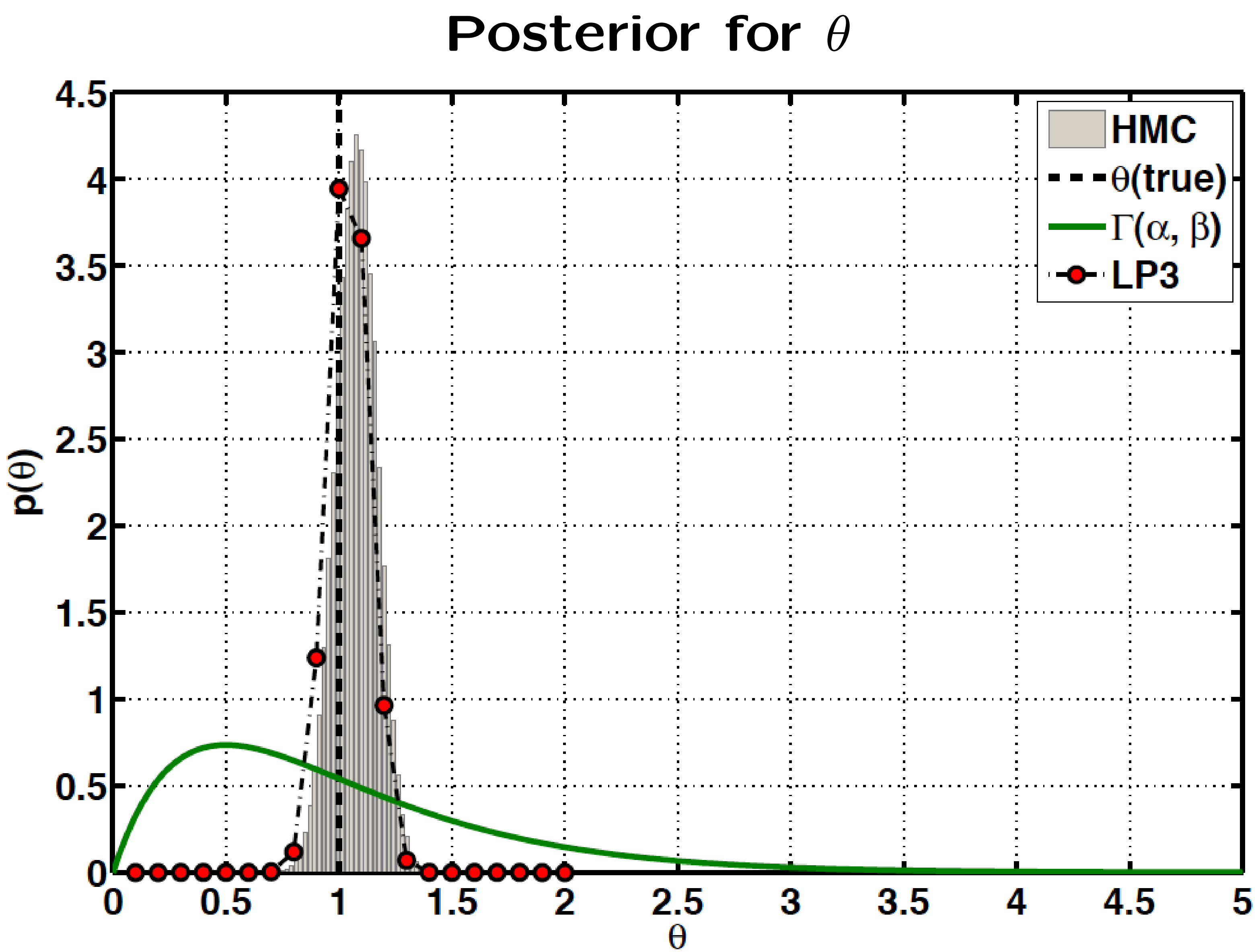
A trajectory



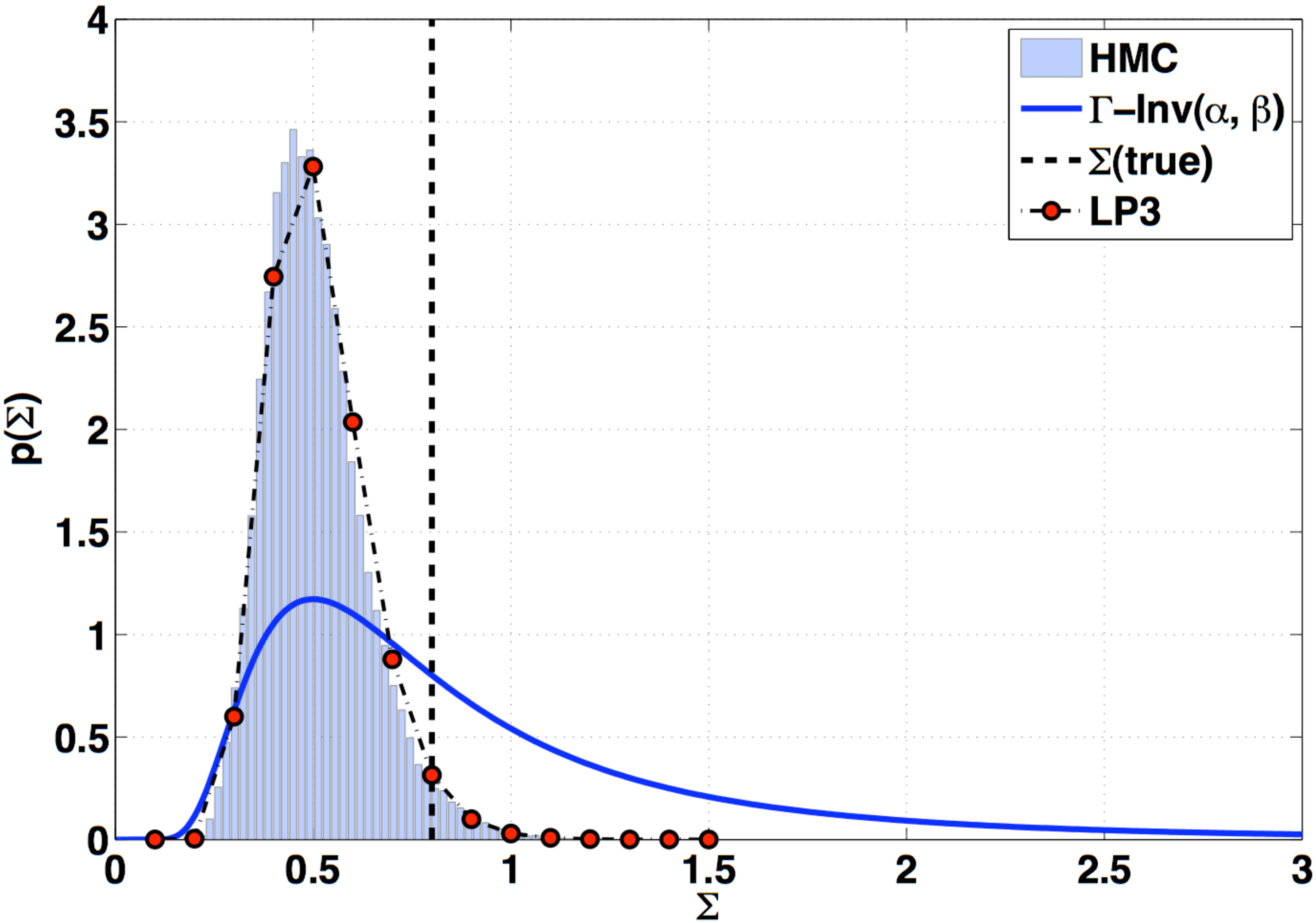
Prediction & comparison with hybrid Monte Carlo

$T = 20$, $\theta = 1$, $\sigma^2 = 0.8$ with $N = 40$ observations with noise $\sigma_o^2 = 0.04$. Fixed initial conditions.



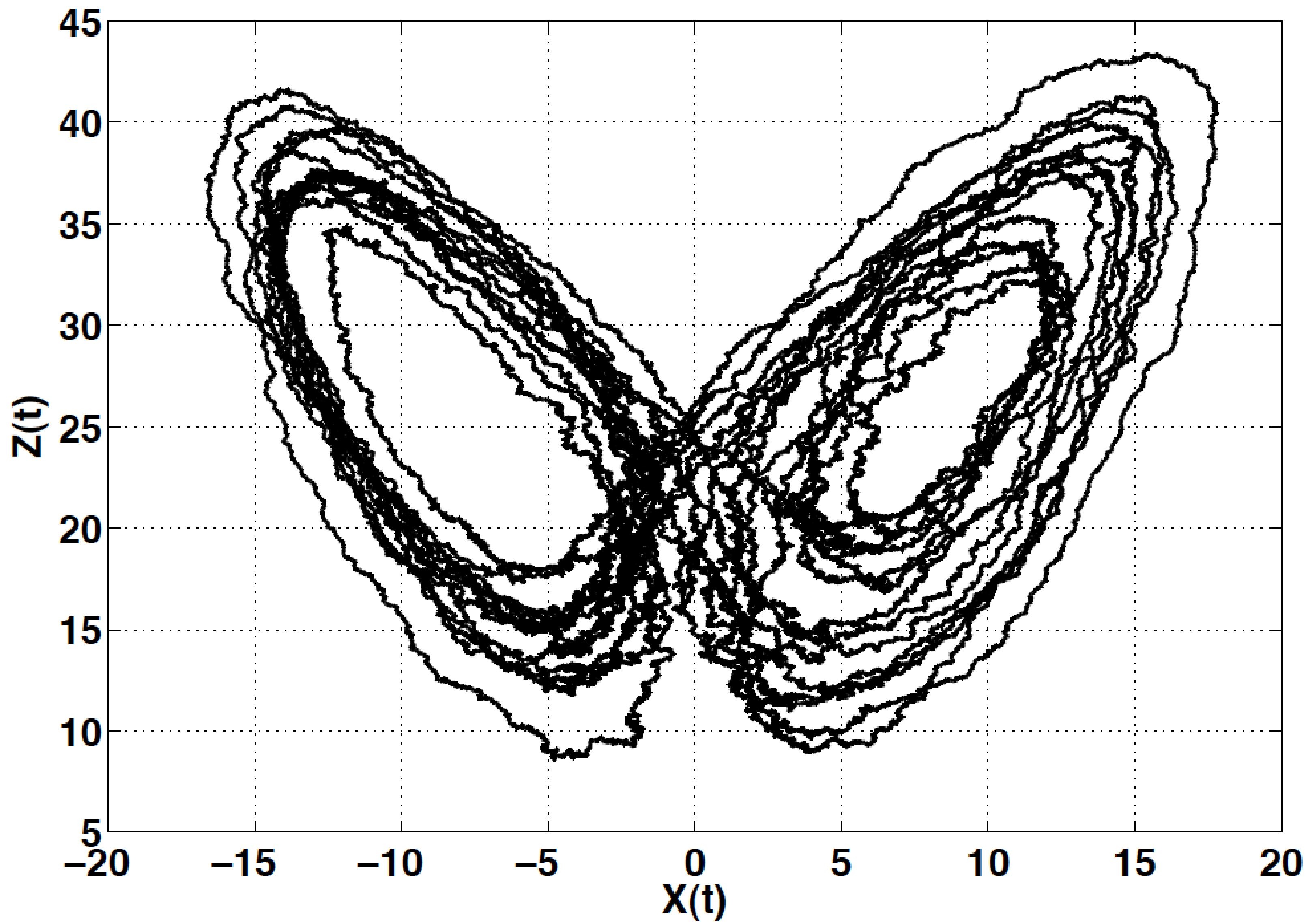

Posterior for θ
HMC
θ(true)
Γ(α, β)
LP3
p(θ)
θ
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Posterior for σ



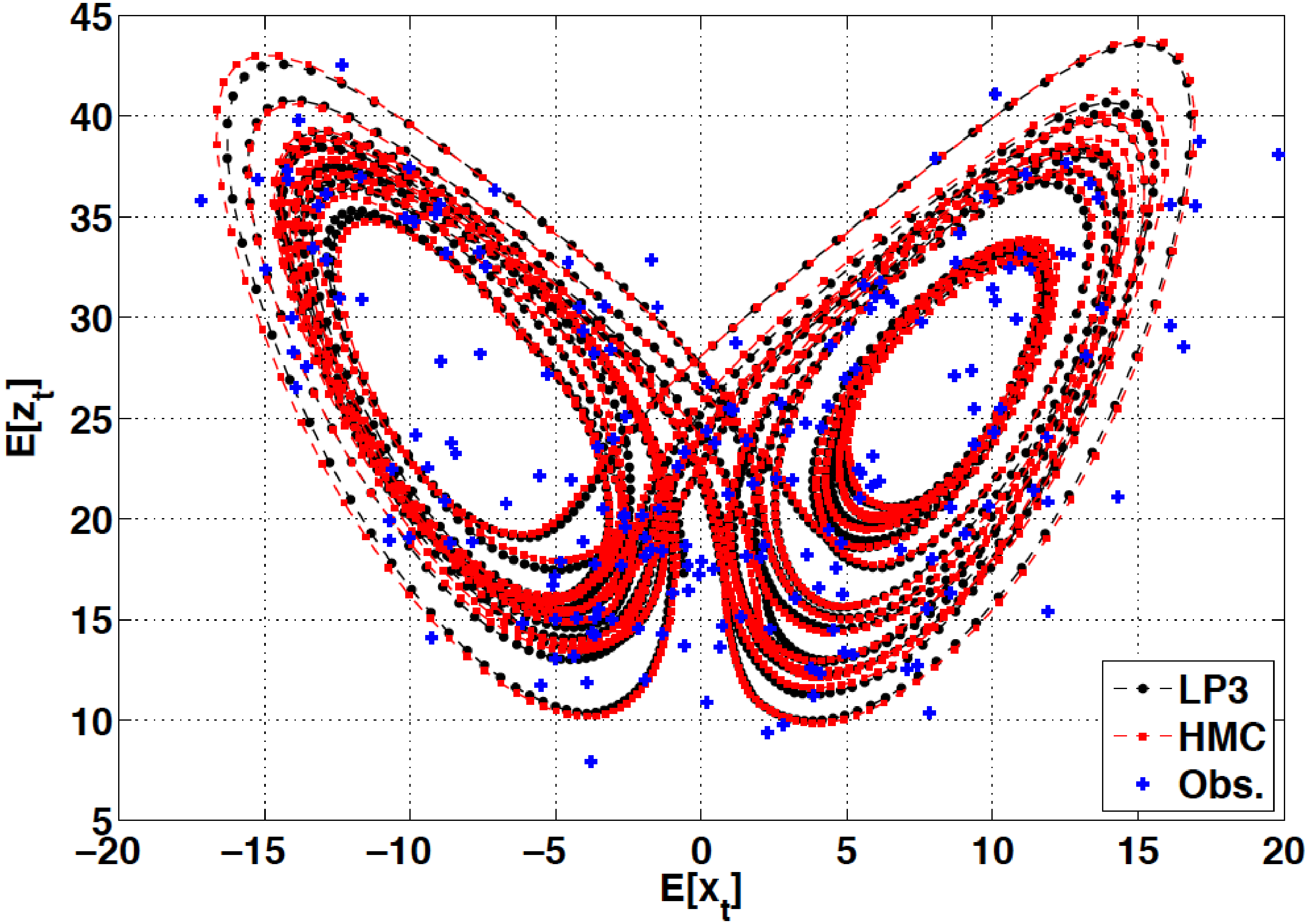
Lorenz 1963

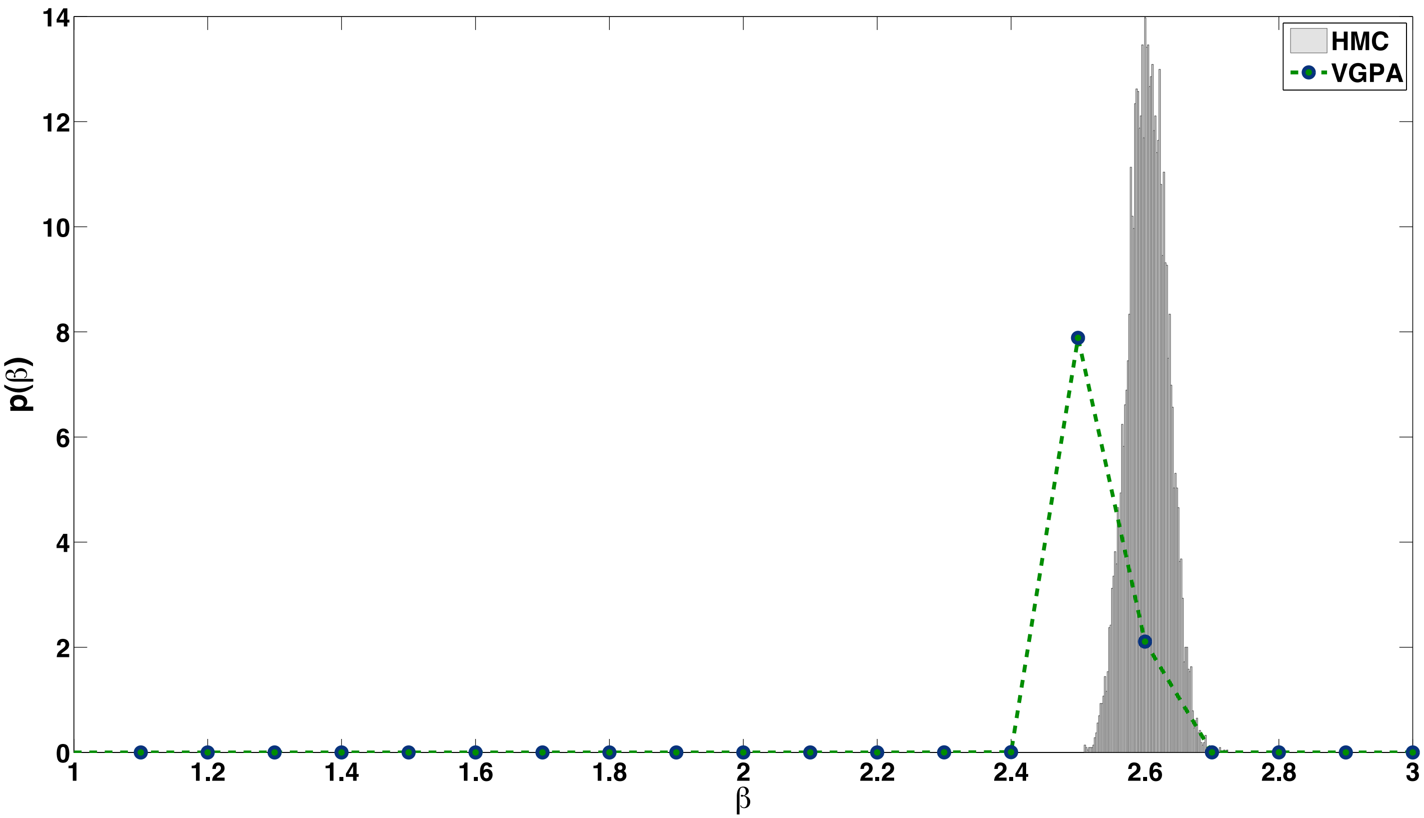
$$
\begin{aligned}
dx_t &= \sigma(y_t - x_t)dt + \sqrt{\Sigma^x}dW^x \\
dy_t &= (\rho x_t - y_t - x_t z_t)dt + \sqrt{\Sigma_y}dW^y \\
dz_t &= (x_t y_t - \beta z_t)dt + \sqrt{\Sigma_z}dW^z
\end{aligned}
$$



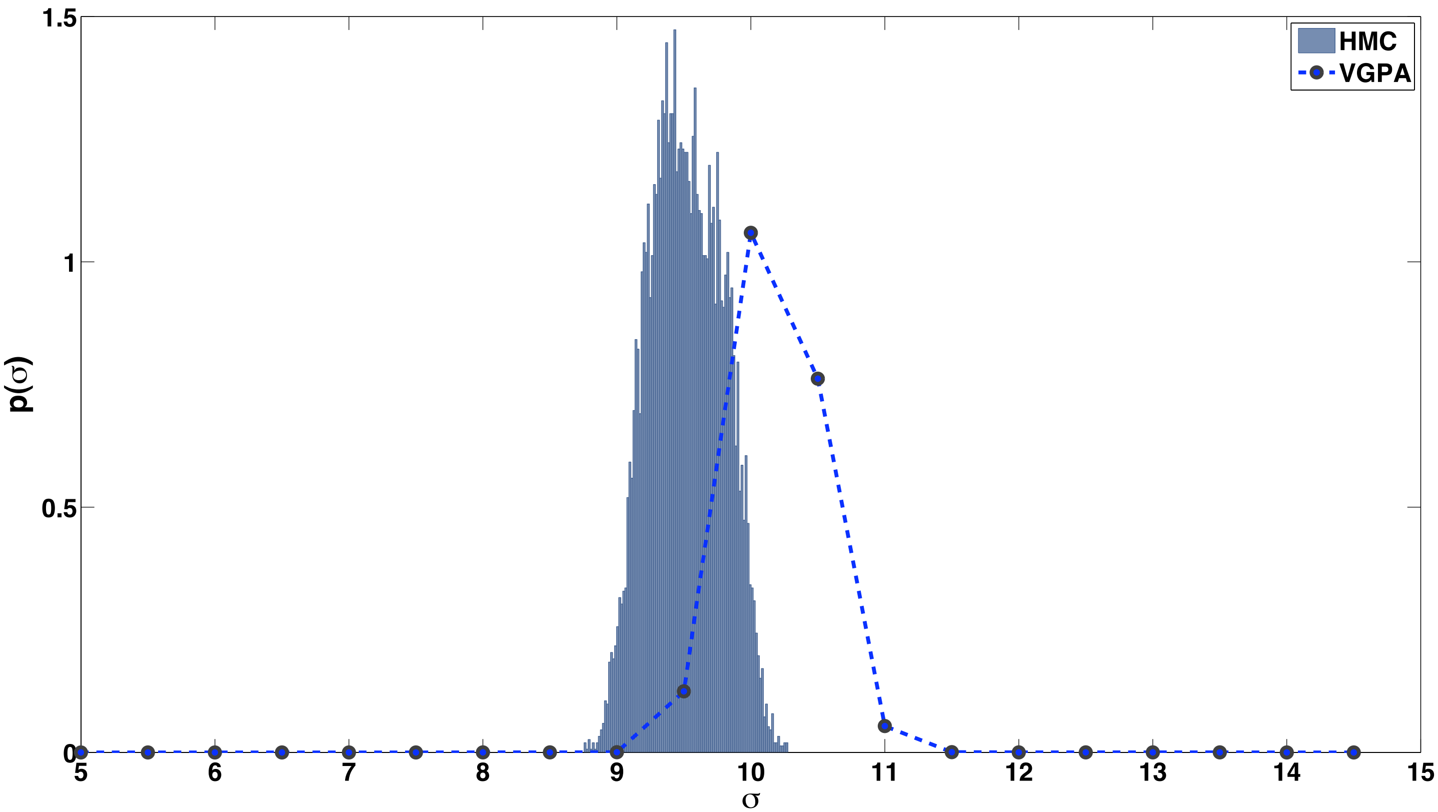
45
40
35
30
25
20
15
10
5
Z(t)
−20
−15
−10
−5
0
5
10
15
20
X(t)

Prediction and comparison with hybrid HMC





HMC
VGPA
p(β)
β
0
2
4
6
8
10
12
14
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3



1.5
1
0.5
0
5
6
7
8
9
10
11
12
13
14
15
p(σ)
σ
HMC
VGPA

More dimensions: Mean field approximation

Approximate further by assuming that processes for different dimensions are independent.

Covariance $S(t) \to \mathrm{Diag}(s_1(t), \ldots, s_D(t))$

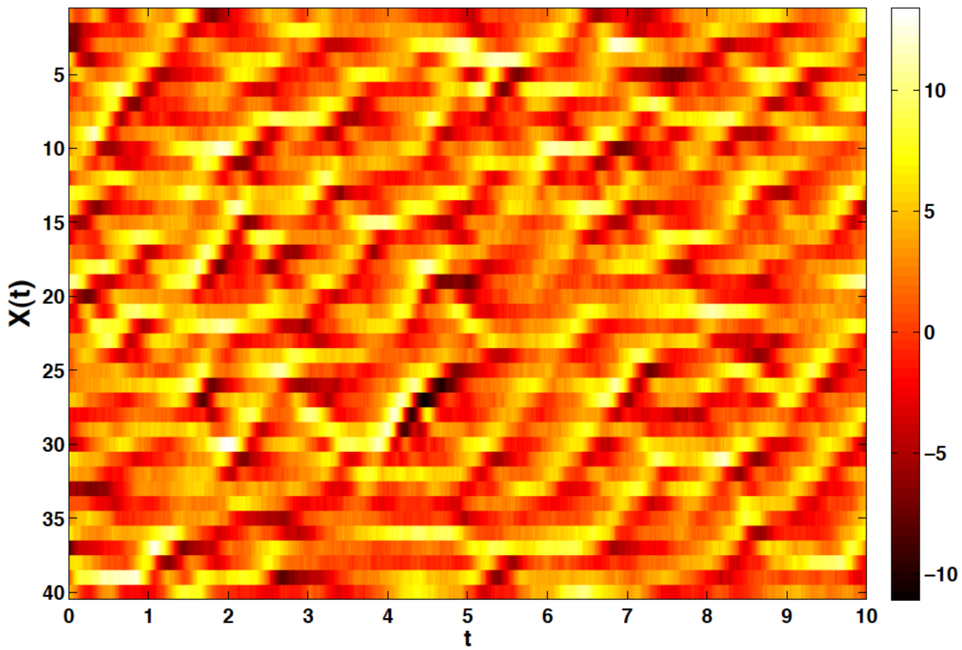
$$
\begin{aligned}
\mathcal{F}_\theta(q) = \sum_{i=1}^{D} \frac{1}{2\sigma_i^2} \int_0^T E_q\left[(\dot{m}_i - f_i(X_t))^2\right]\, dt \\
+ \sum_{i=1}^{D} \frac{1}{2\sigma_i^2} \int_0^T \left\{ \frac{(\dot{s}_i - \sigma_i^2)^2}{4 s_i^2} + (\sigma_i^2 - \dot{s}_i) E_q\left[\frac{\partial f_i(X_t)}{\partial X_t^i}\right] \right\} dt \\
- \sum_{j=1}^{n} E_q\left[\ln p(y_j | X_{t_j})\right]
\end{aligned}
$$

Lorenz 1998 model:

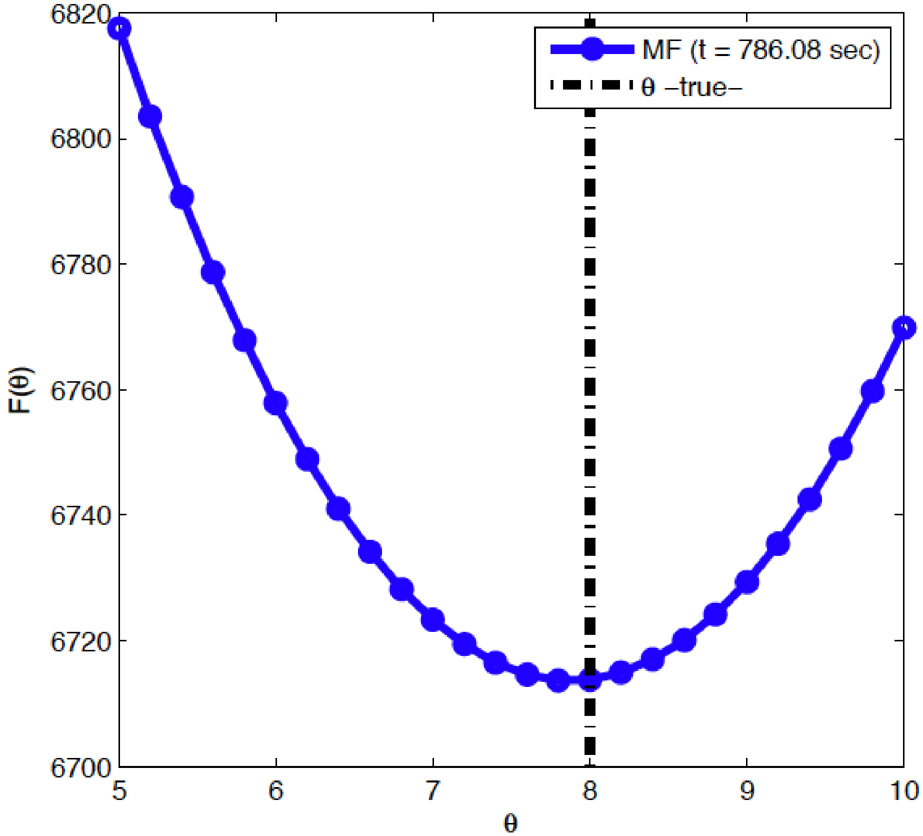
$x = (x^1, \ldots, x^{40})$ with drift

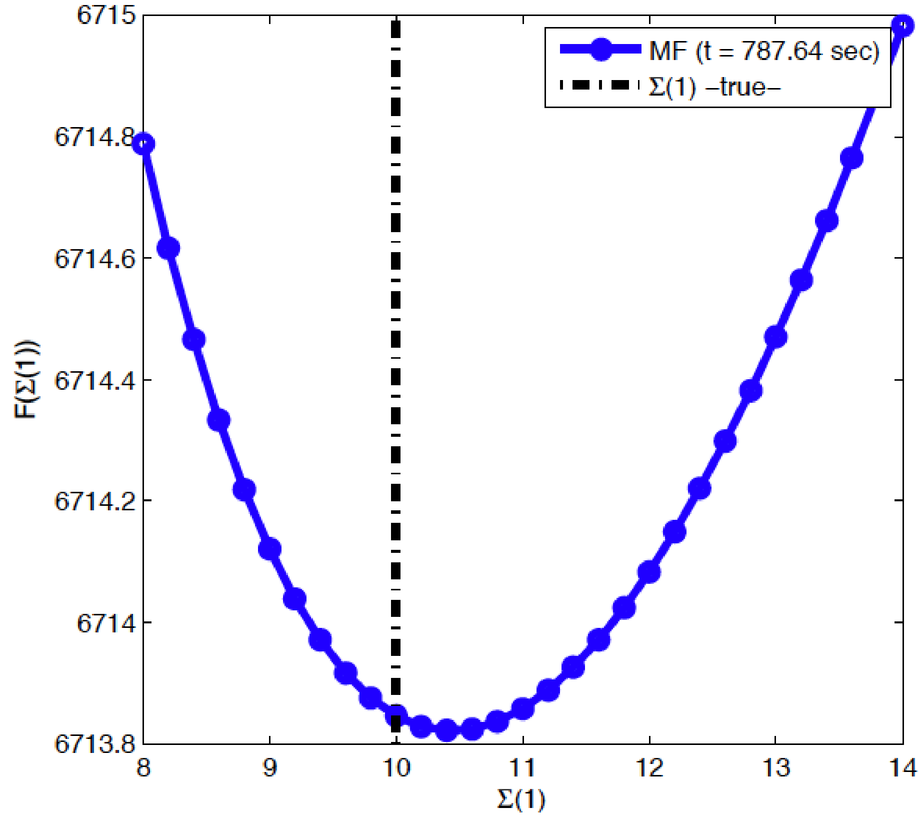
$$f_i(x_t) = \left(x_t^{i+1} - x_t^{i-2}\right) x_t^{i-1} - x_t^i + \theta$$

$\sigma^2 = 5$ and $N = 90$ observations.



Likelihoods





The relative entropy for Markov jump processes

Assume transition rates $g(x'|x,t)$ and $f(x'|x,t)$

$$KL\,[q||p] =$$

$$\int_0^T dt\,\sum_x q_t(x)\sum_{x':x'\neq x}\left\{g(x'|x,t)\ln\frac{g(x'|x,t)}{f(x'|x,t)}+f(x'|x,t)-g(x'|x,t)\right\}$$

Mean field approximation

Multivariate states $X(t) = (X_1(t), \ldots, X_d(t))$

Exact inference: Linear ODEs in S^d variables

Variational approximation: Optimise in family of factorising measures, i.e. of the type

$q(X[0:T]) = \prod_{i=1}^{d} q_i(X_i[0:T])$

Linear ODEs in Sd variables.

(Sanguinetti & Opper, 2008, Cohn, El–Hay, Friedman & Kupferman, 2010)

Lotka Volterra

Comparison with MCMC

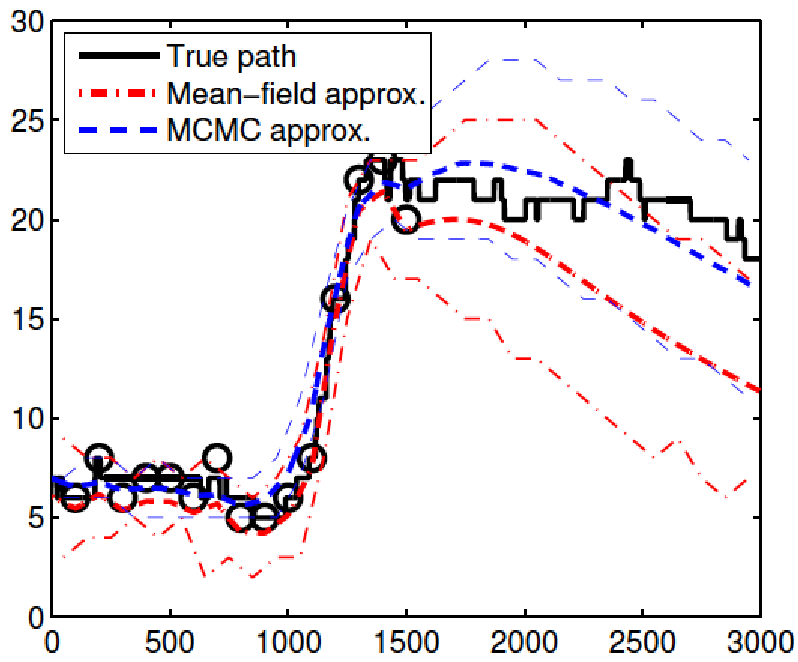


Figure 1: Posterior (mean and 90% confidence intervals) over predator paths (observations (circles) only until 1500).

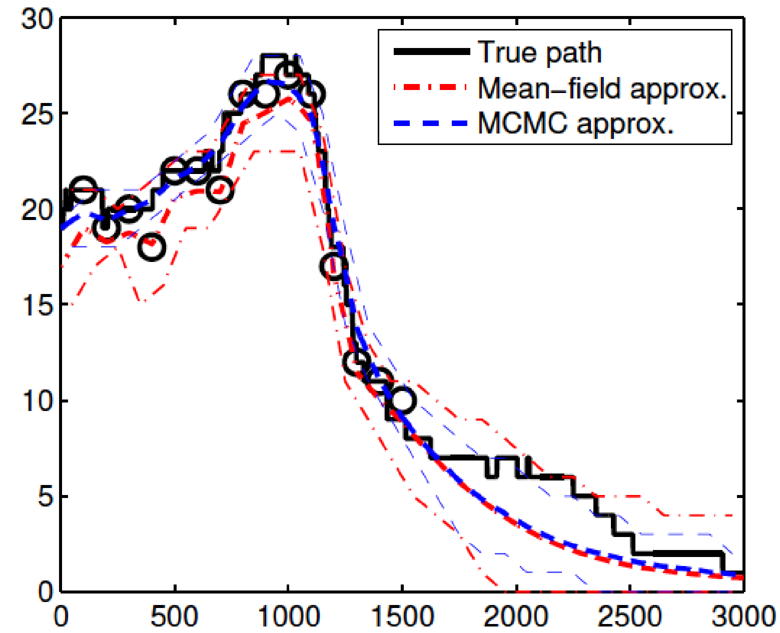
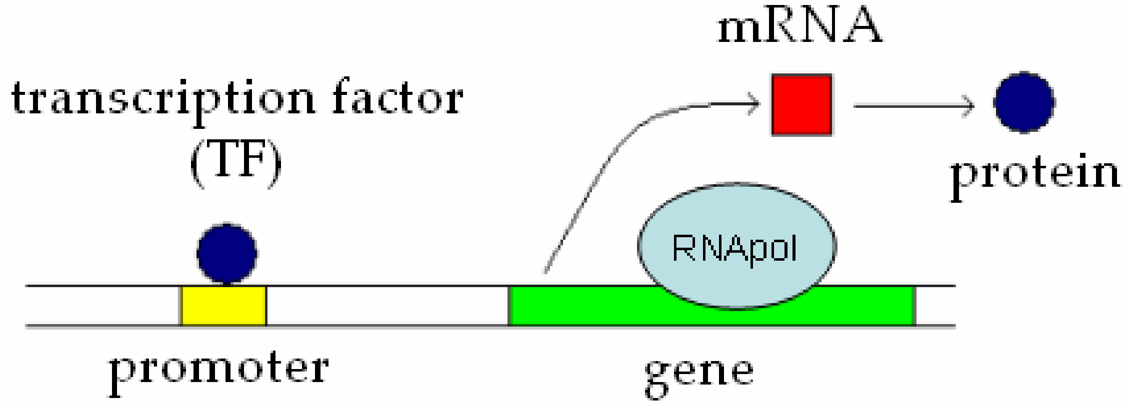


Figure 2: Posterior (mean and 90% confidence intervals) over prey paths (observations (circles) only until 1500).

from (V Rao & Y W Teh, 2011)

Hybrid models: Inference of transcriptional regulation



- Transcription factors regulate genes by binding to specific sites.

- Hard to measure transcription factor activity directly. Inference must be based on measurement of mRNA concentration of target genes.

- Big networks: Clustering of expression profiles or Factor analysis

Small subnetworks:

- More detailed dynamical model (Barenco et al)

$$\frac{dx_i}{dt} = -\lambda_i x_i(t) + b_i + A_i \mu(t)$$

which takes sensitivity and degradation into account.

- Try predictions on TF activity $\mu(t)$ and learn parameters using measurements of mRNA concentration of target genes:

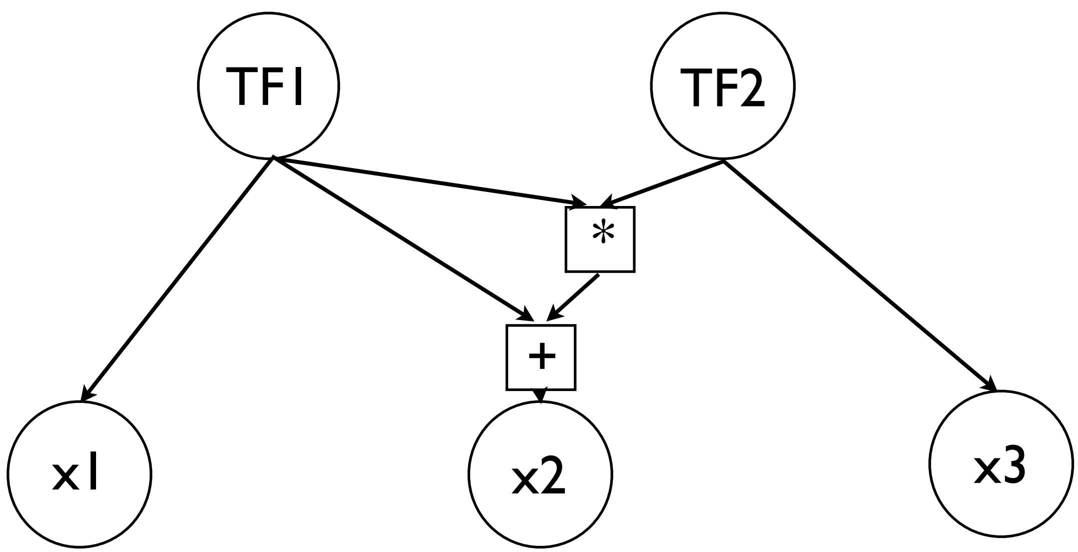
$$y_{ik} = x_i(t_k) + \text{noise}$$

.

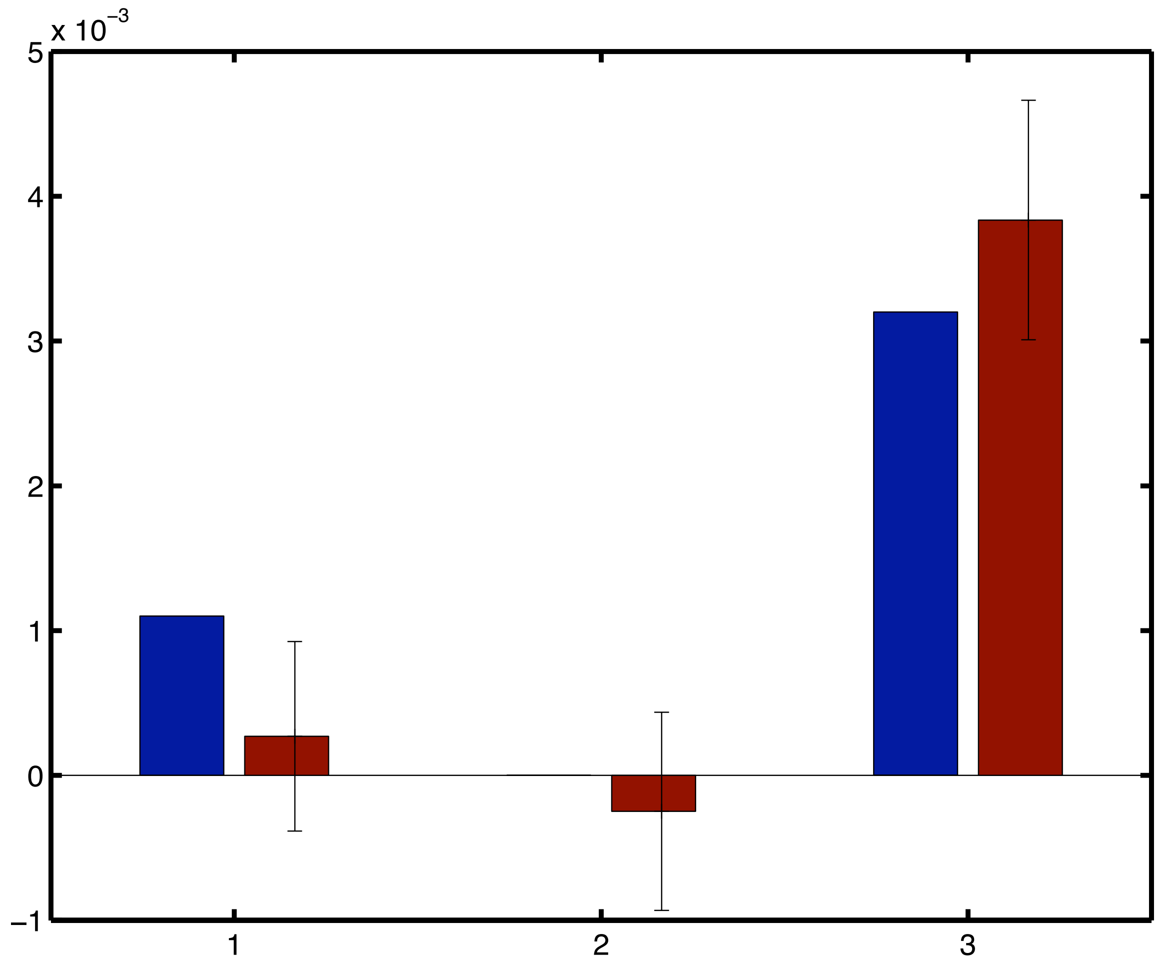
- Assume switching process $\mu(t) \in \{0, 1\}$ and $\mu \to 1 - \mu$ with rates $f_\pm$ modeled by **telegraph process** (Sanguinetti, Ruttor, Archambeau, Opper 2009)

Multiple (2) transcription factors (toy model)

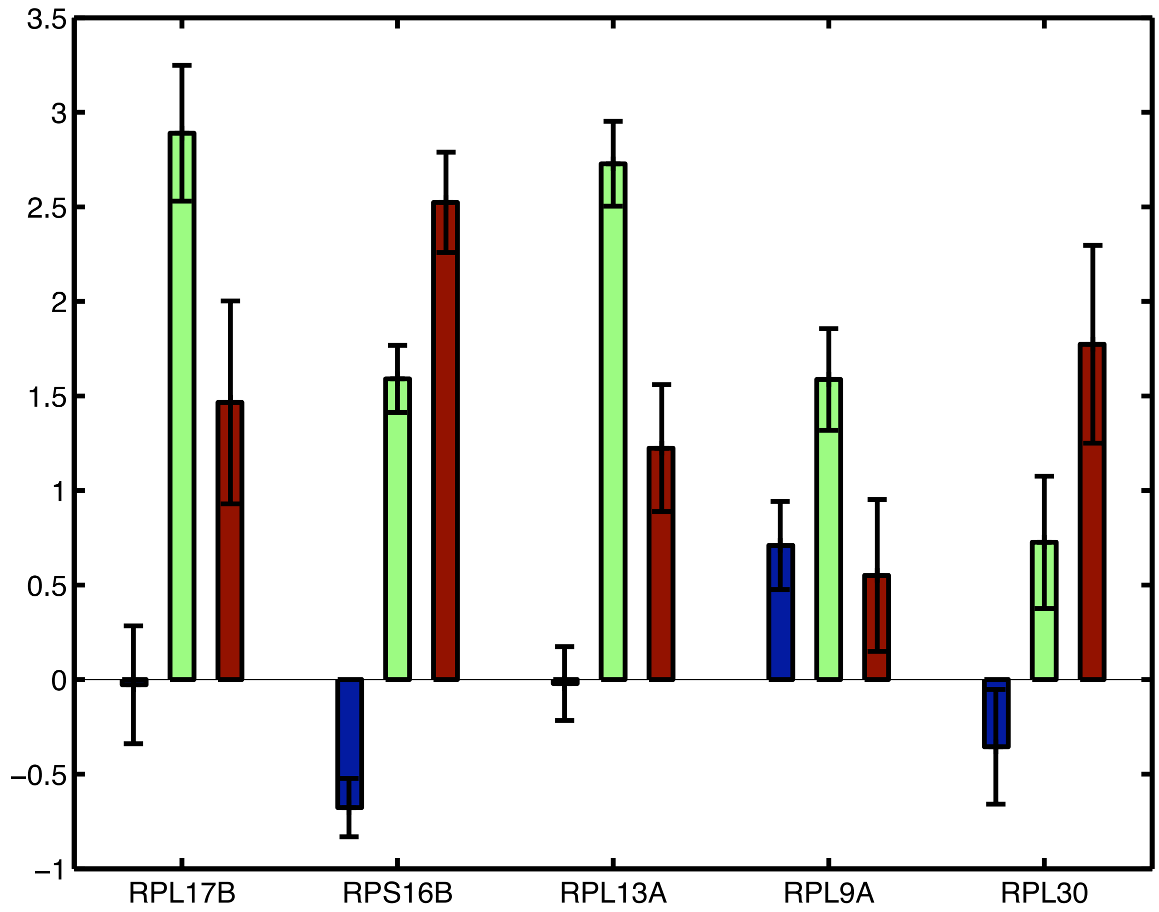
$$\frac{dx_i}{dt} = -\lambda_i x_i(t) + A_1^i \mu_1(t) + A_2^i \mu_2(t) + A_{12}^i \mu_1(t)\mu_2(t) + b_i$$



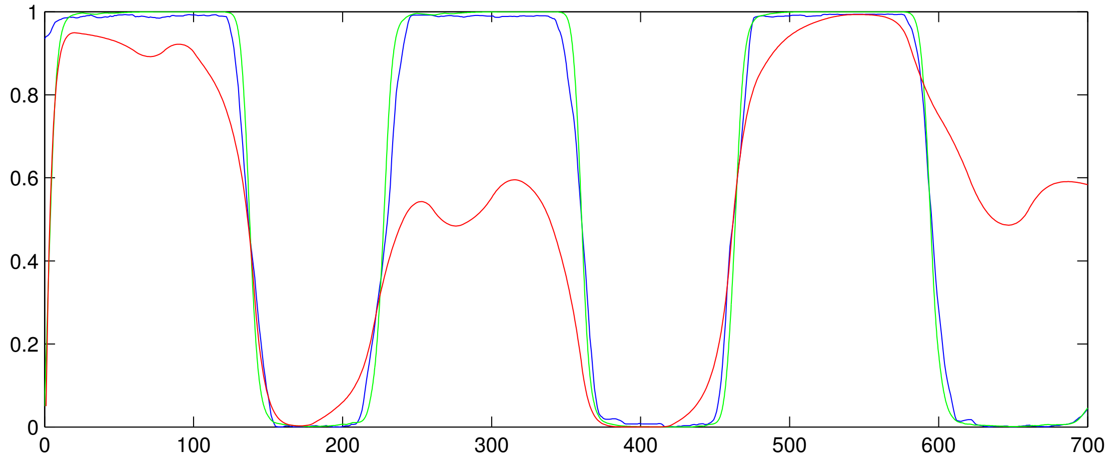
Parameter inference for A_1^2 , A_2^2 and A_{12}^2.

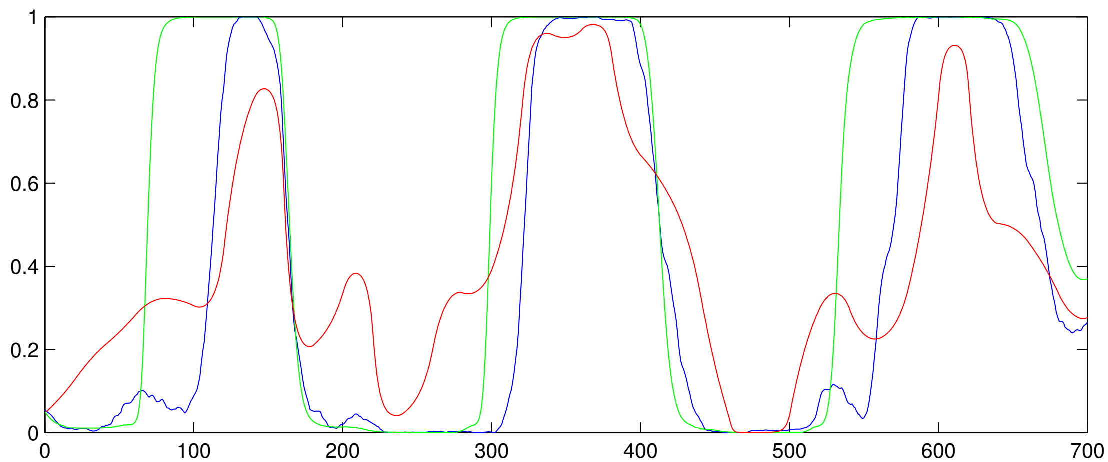


A_1^i, A_2^i and A_{12}^i for 5 target genes.



Prediction of activity of transcription factors **FHL1** and **RAP1** (Microarray data from yeast metabolic cycle). Comparison to MCMC





(blue: MCMC, green: Variational upper, red: Var lower bound)

Feed–forward–loop

(A Ocone & G Sanginetti, 2011)

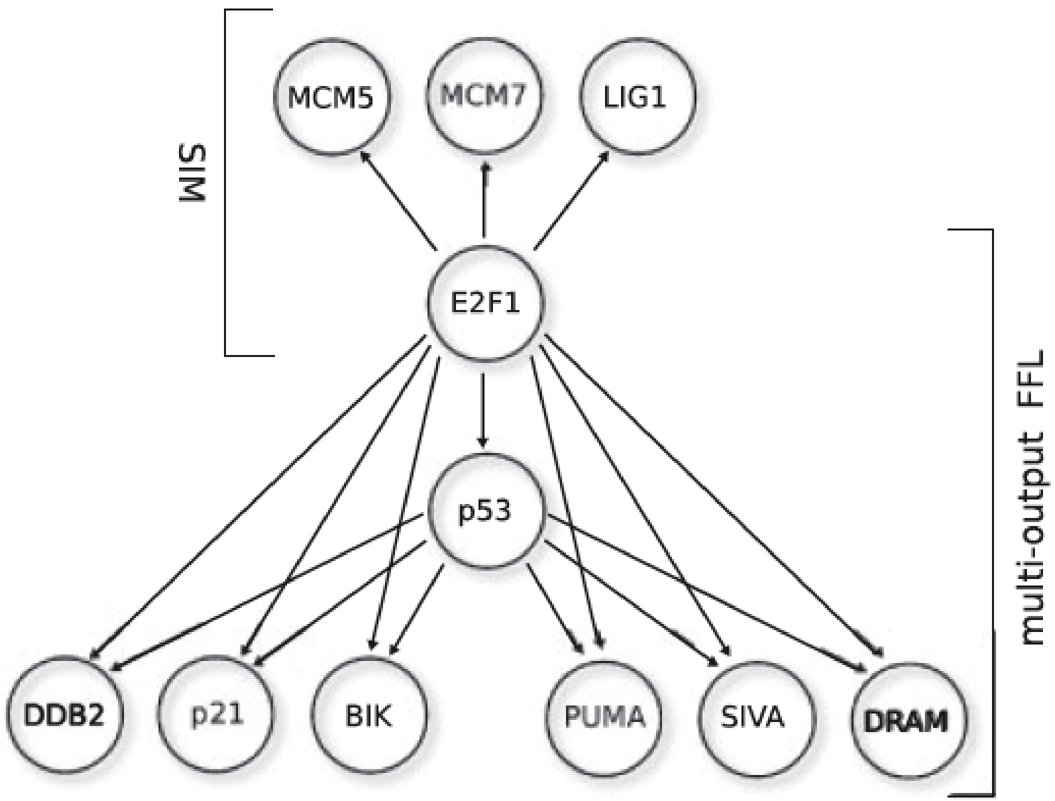
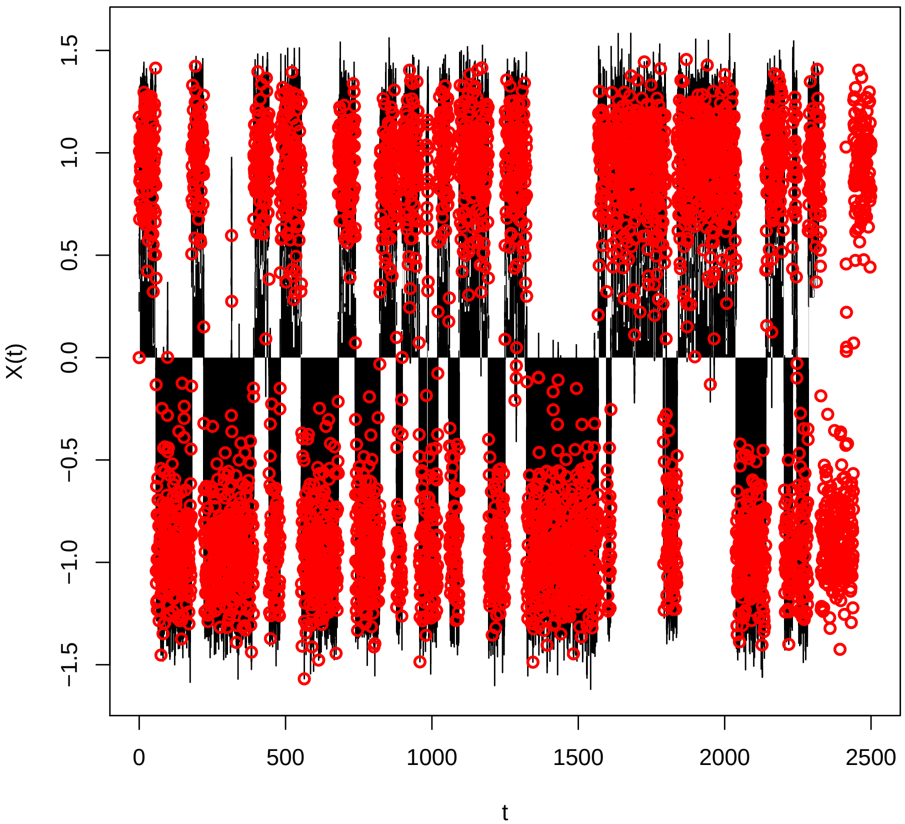
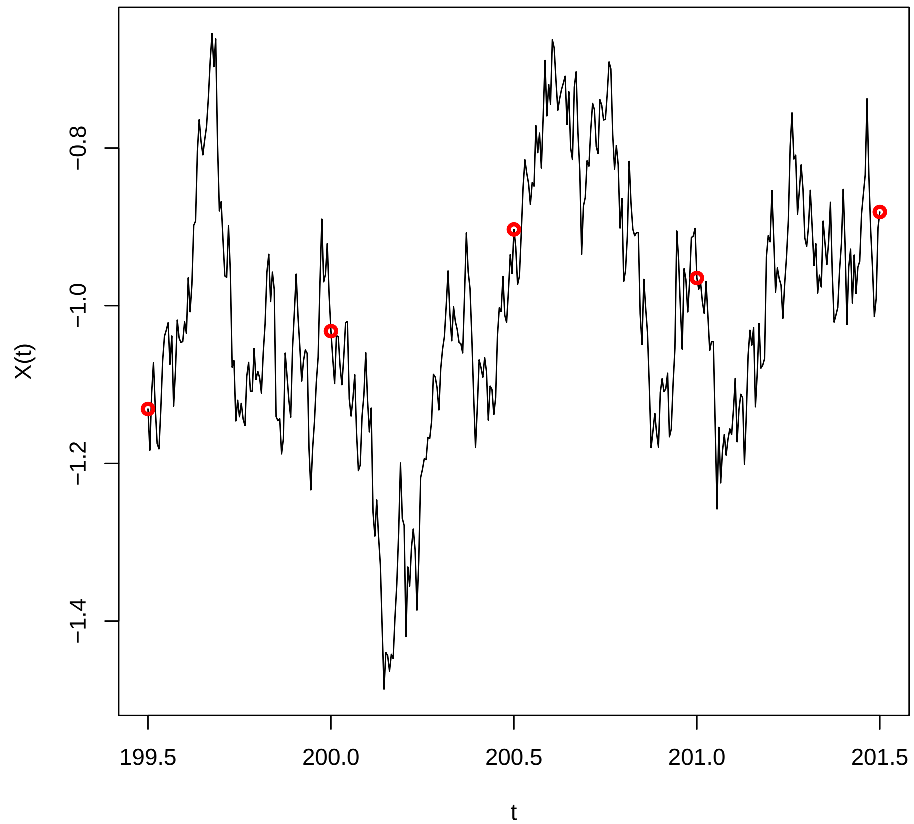


Fig. 3. p53 network architecture. E2F1 is the master TF, p53 is the target TF and both regulate target genes *DDB2, p21, BIK, PUMA, SIVA, DRAM*. Target genes of the only E2F1 (*MCM5, MCM7, LIG1*) have been included.

Nonparametric estimation of drift function

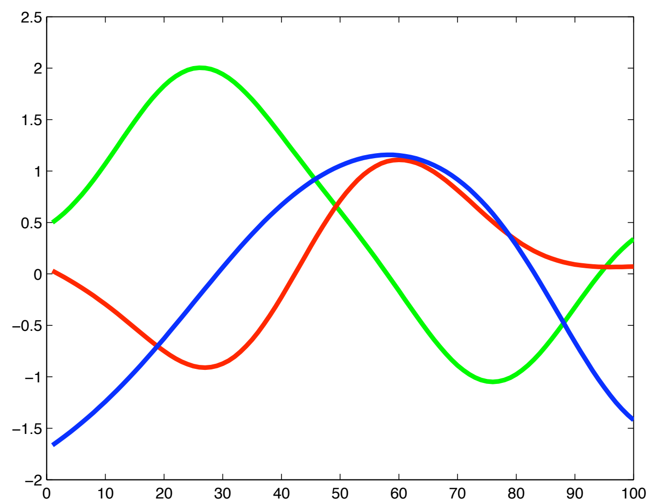




Assume that data are generated from $dX_t = f(X_t)dt + \sigma dW_t$.
Could we directly predict $f(x)$?

Yes, if we use a Gaussian Process prior

distribution $p(\mathbf{f})$ over functions $f(\cdot)$.



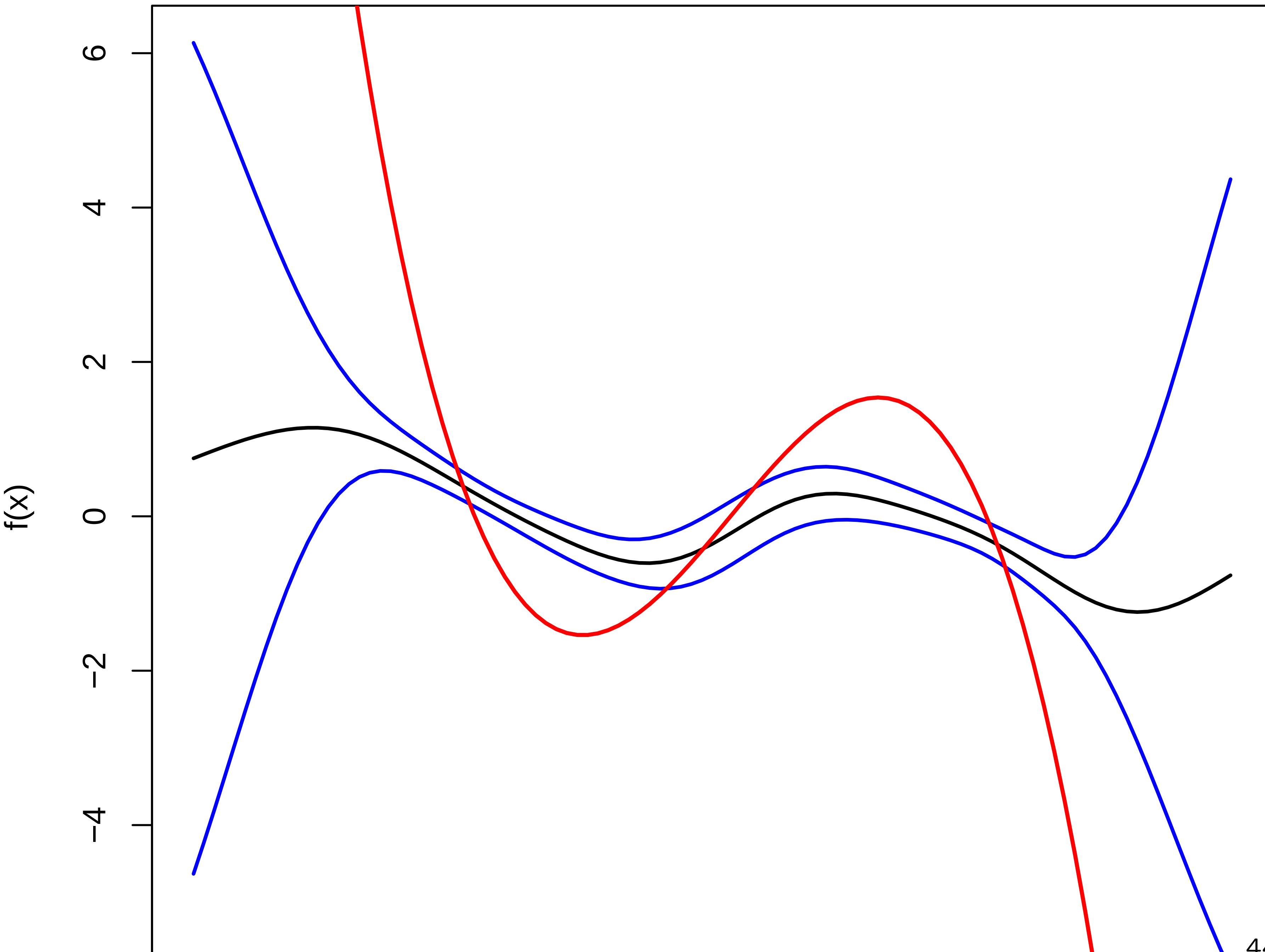
- If we have 'continuous time' samples $\rightarrow$ posterior process is Gaussian (Papaspiliopulis et al, 2011).

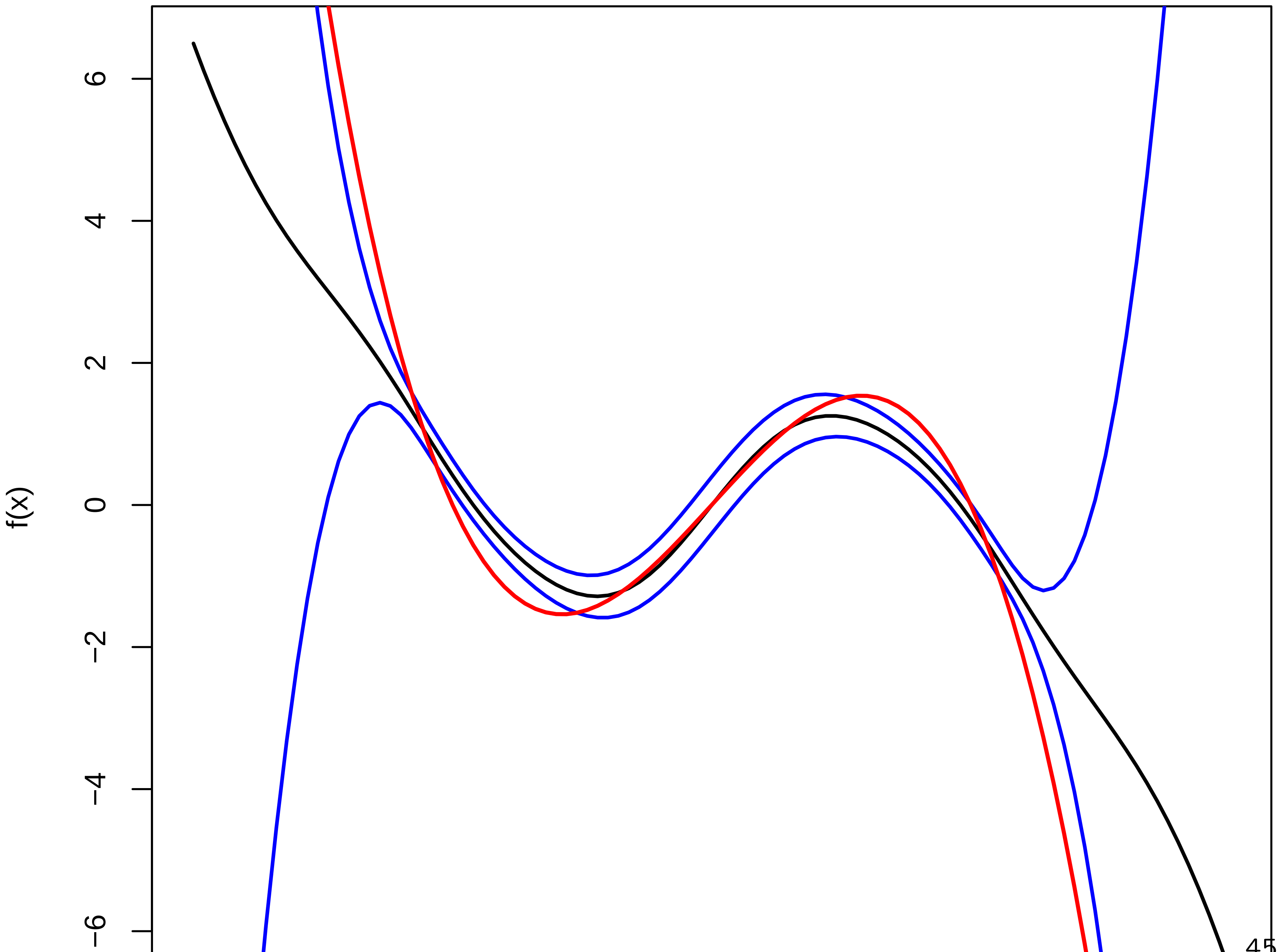
$$p(\mathbf{f}|X_{0:T}) \propto p(\mathbf{f}) L(X_{0:T}|\mathbf{f})$$

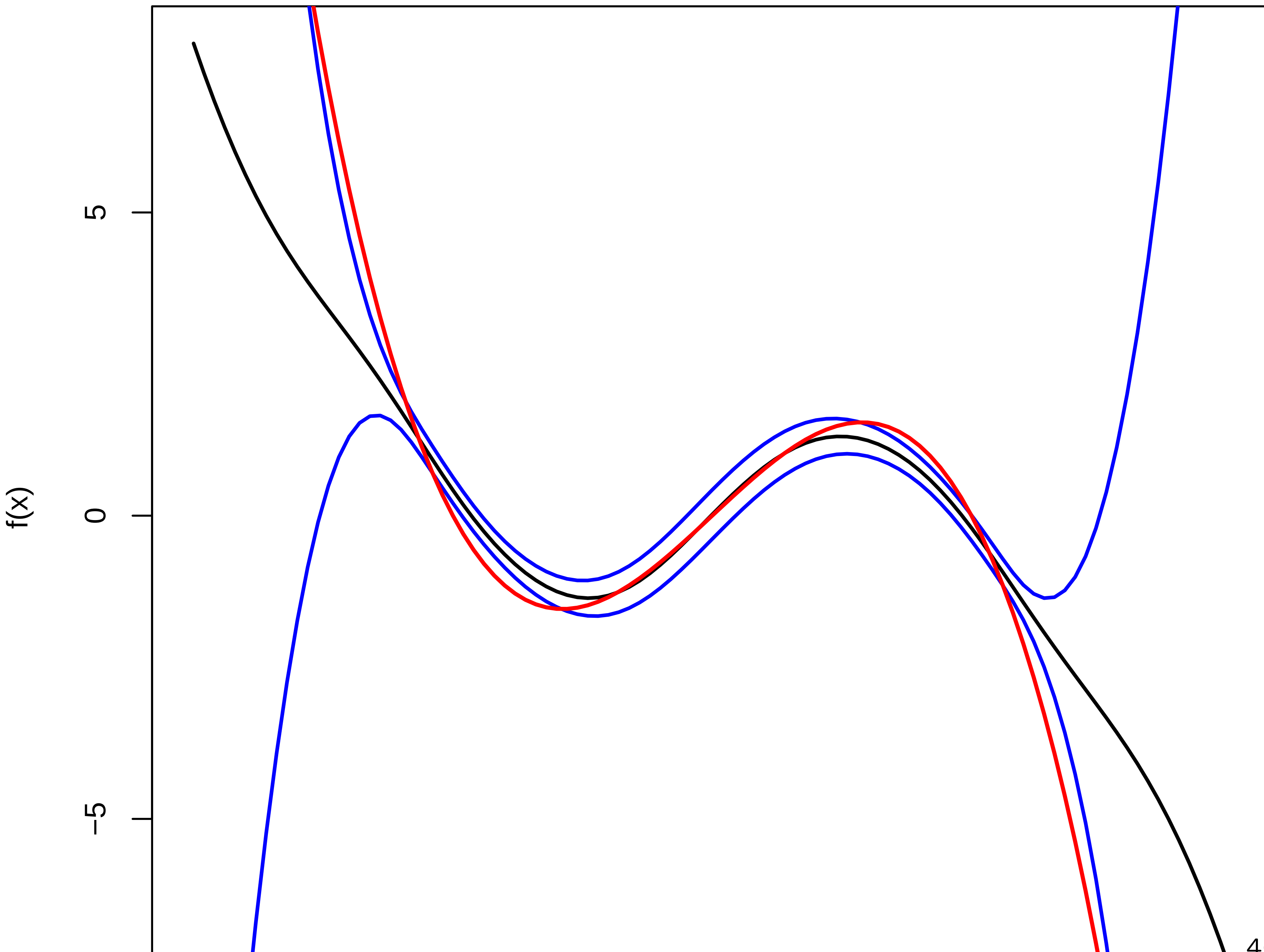
with the path 'likelihood'

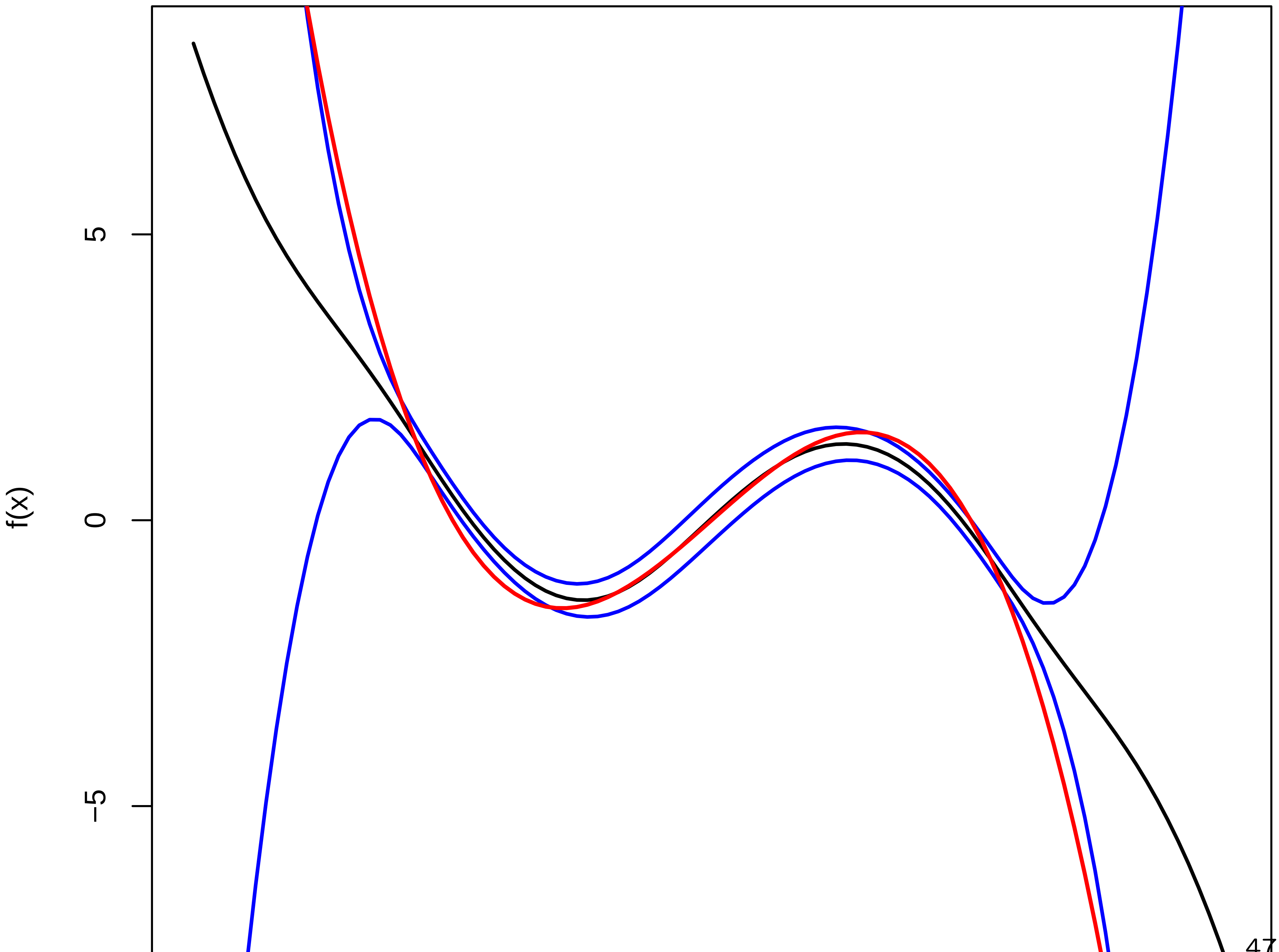
$$L(X_{0:T}|\mathbf{f}) = \exp\left[-\frac{1}{2\sigma^2}\sum_t f^2(X_t)\Delta t + \frac{1}{\sigma^2}\sum_t f(X_t)\left(X_{t+\Delta t} - X_t\right)\right]$$

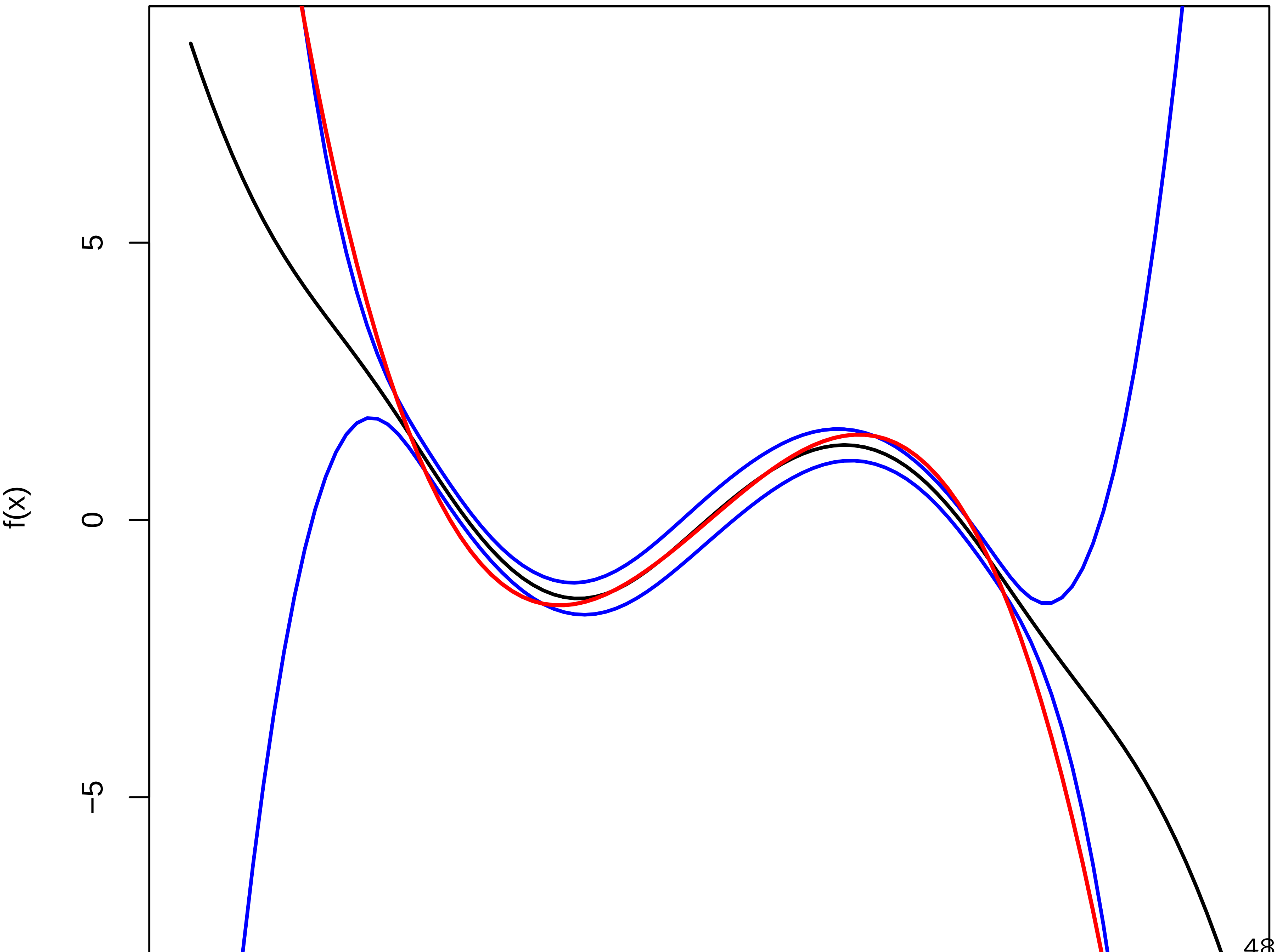
- For **sparse samples** use EM algorithm which cycles between **approximate estimations of latent path** $X_{0:T}$ between observations and recomputing $f(\cdot)$.

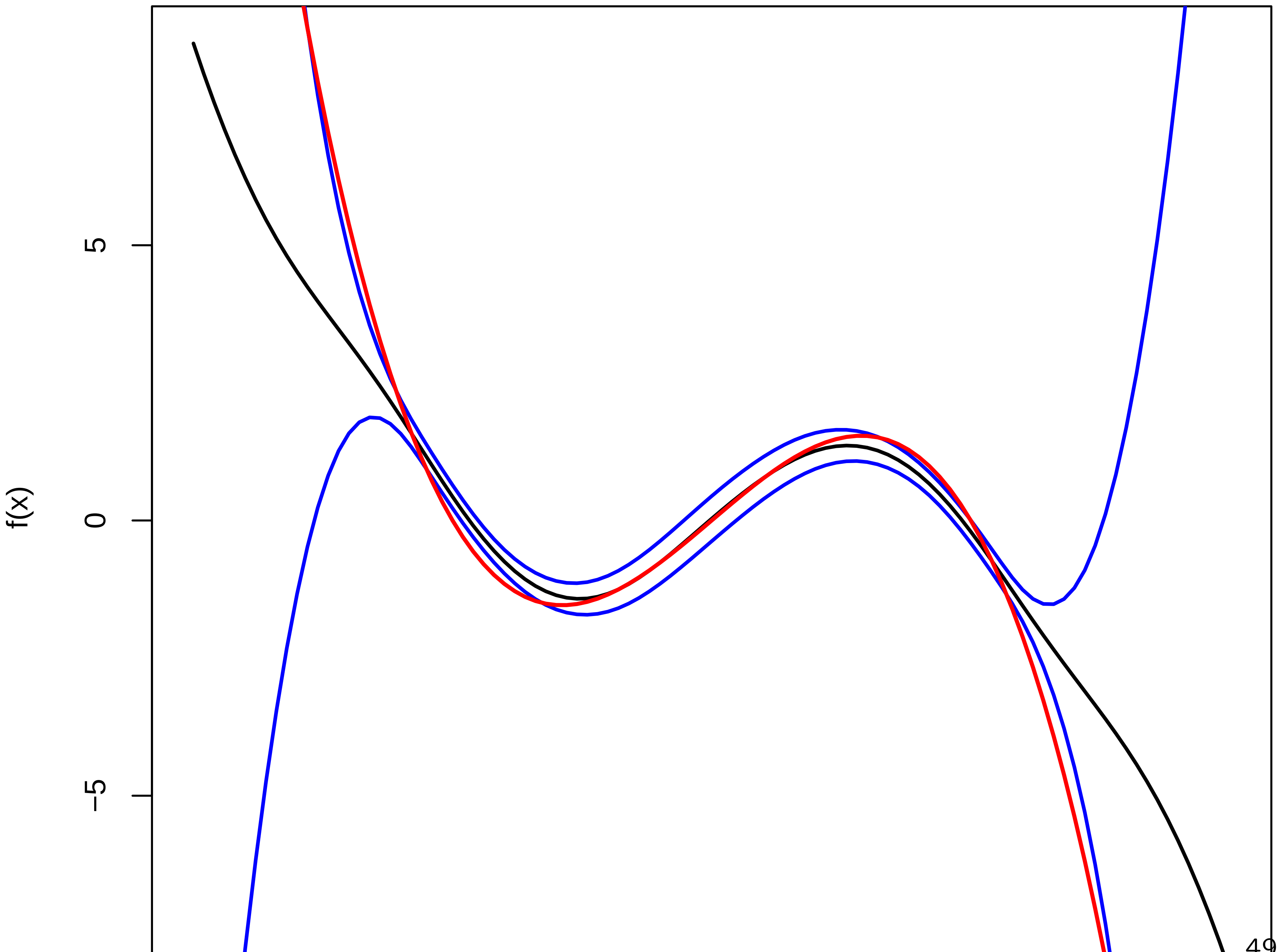

f(x)
6
4
2
0
−2
−4

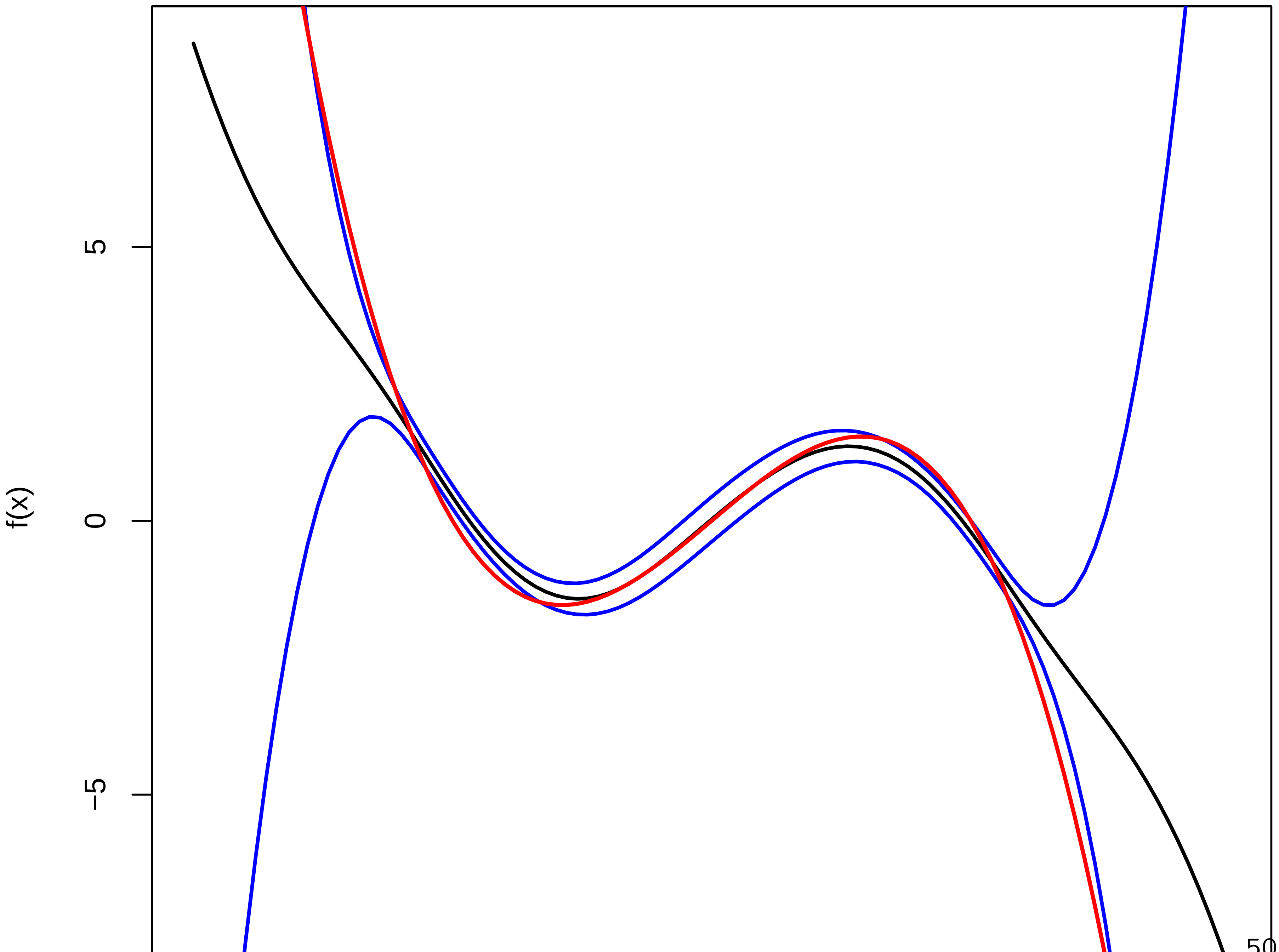

f(x)
6
4
2
0
−2
−4
−6
45

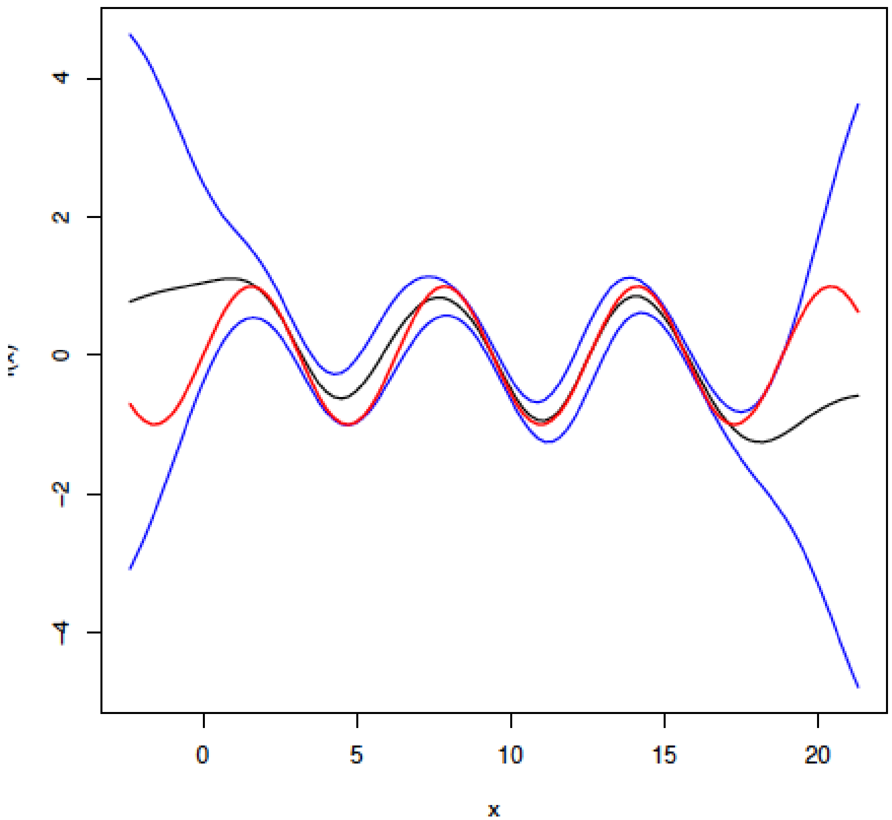

f(x)
5
0
−5


f(x)
5
0
−5

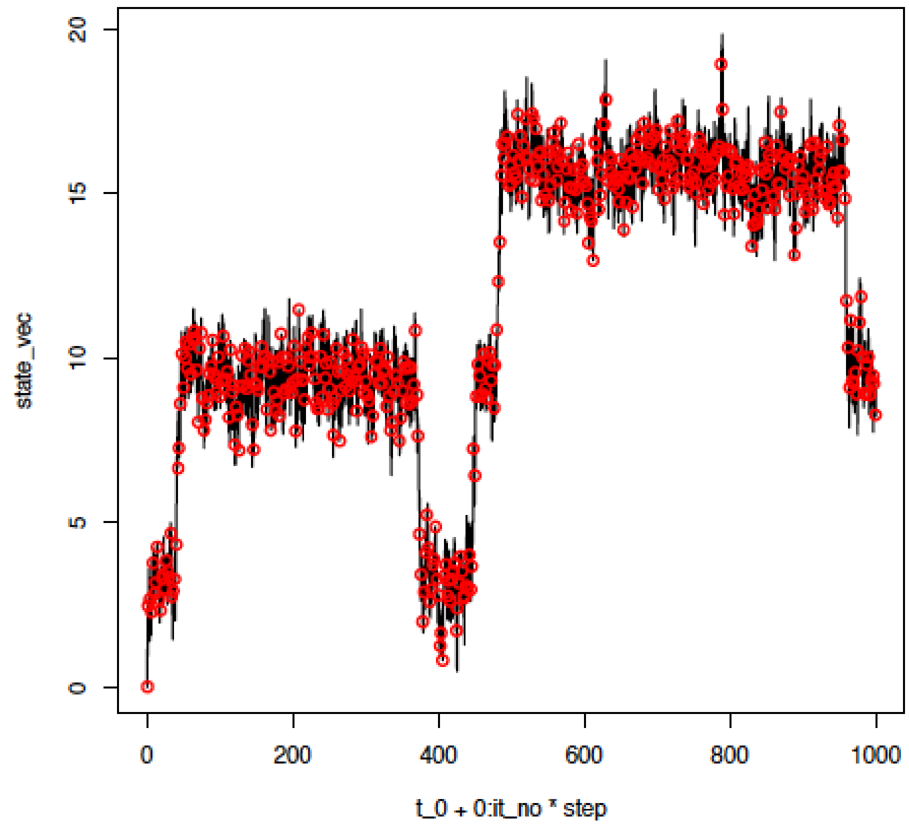

f(x)
5
0
−5


f(x)
5
0
−5


f(x)
5
0
−5
50



True path



Present & Future work

- Large systems: Simpler classes of approximations, eg. parametric forms for large covariance matrices (projections, low rank representations ?)

- Perturbative corrections (estimate for error)

- State dependent noise.

- Nonparametric estimation of drift $f(x)$ for models with detailed balance

- Combination with optimal stochastic control

Publications

Gaussian Process Approximations of Stochastic Differential Equations, Cédric Archambeau, Dan Cornford, Manfred Opper and John Shawe - Taylor, Journal of Machine Learning Research: Workshop and Conference, Proceedings, 1:1–16. (2007).

Variational Inference for Diffusion Processes, Cedric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford and John Shawe-Taylor, Advances in Neural Information Processing Systems 20 (2008).

Variational inference for Markov jump processes, Manfred Opper and Guido Sanguinetti, Advances in Neural Information Processing Systems 20, 1105–1112 (2008).

A comparison of variational and Markov Chain Monte Carlo methods for inference in partially observed stochastic dynamic systems, Yuan Shen, Cédric Archambeau, Dan Cornford, Manfred Opper, John Shawe-Taylor and Remi Barillec. Journal of Signal Processing Systems (2009).

Switching Regulatory Models of Cellular Stress Response, Guido Sanguinetti, Andreas Ruttor, Manfred Opper and Cédric Archambeau, Bioinformatics, doi: 10.1093/bioinformatics/btp138 (2009).