Understanding Local Linearisation in Variational Gaussian Process State Space Models

Talay M Cheema (tmc49@cam.ac.uk)

Motivation

- Many data can be well-described by dynamical systems models, e.g. mechanistic differential equations (Figure 1).
- In application domains such as model based reinforcement learning or synthetic biology, systems must be learnt from noisy data.
- Where the task requires uncertainty estimates (e.g. active learning), this invites a Bayesian approach.
- If the prior on the function is a Gaussian process, we have a Gaussian process state space model (GPSSM). Previous variational inference approaches either
  - assume independence of dynamics f and states x, which we show badly biases the process noise hyperparameter Q, leading to poor predictions, and/or
  - require parameters scaling with the number of latent states, which is particularly limiting in continuous time, but unecessary.

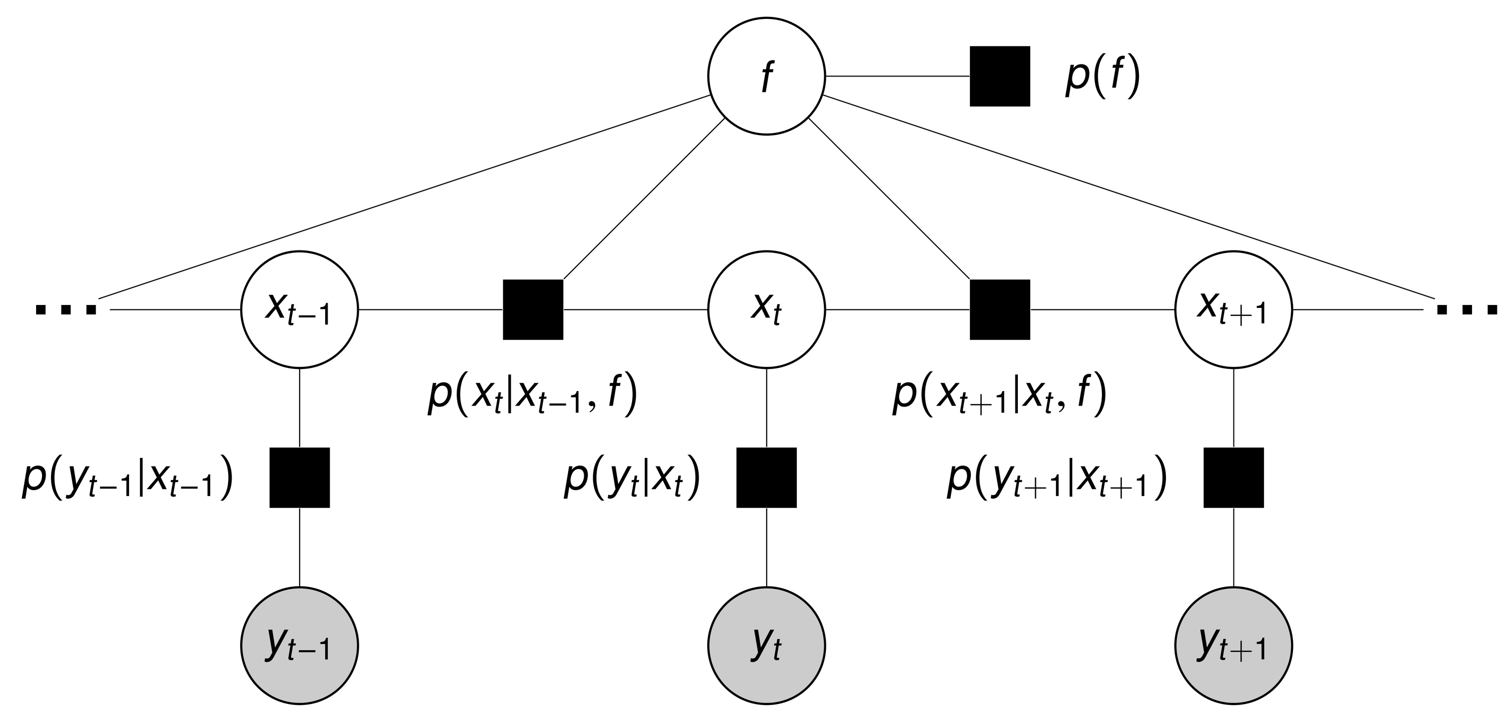

Figure 1: Dynamical system factor graph. The states x are a Markovian intermediary from the observations to f. Here $p(x_t|x_{t-1}, f) = \mathcal{N}(x_t|f(x_{t-1}), Q), p(y_t|x_t) = \mathcal{N}(y_t|Cx_t, R)$.

Inferring states (x) given dynamics (f)

If the transitions are **linear Gaussian**

$$q(x_{t+1}|x_t, f) = \mathcal{N}(x_{t+1}|h_t(x_t), \tilde{Q}_t), \quad h_t(x_t) = A_t x_t + b_t \tag{1}$$

then **Kalman filtering/smoothing gives the exact posterior marginals** $p(x_t|y_t)$ efficiently.

For nonlinear f, **any jointly Gaussian approximation** $q(x|f)$ implicitly **has linear-Gaussian transitions** (Table 1), so is a local linearisation.

Table 1: Approximations for Kalman-like filter/smoothers. $E_t[\cdot] = \int \cdot q(x_t = x)dx$.

Method	A_t	b_t	$\tilde{Q}_t$
Extended Kalman filter (EKF)	$\frac{\partial f}{\partial x}\Big\vert_{\mu_t}$	$f(\mu_t) - A_t\mu_t$	Q
Statistical linearised filter (SLF)	$\mathbb{E}_t\left[\frac{\partial f}{\partial x}\right]$	$\mathbb{E}_t[f(x)] - A_t\mu_t$	Q
Assumed density filter (ADF)	0	$\mathbb{E}_t[f(x)]$	$\mathrm{Cov}_t[f(x)]$

If $q(x)$ is not jointly Gaussian, we would only be able to access samples from the marginals. For example, we could use the particle filter.

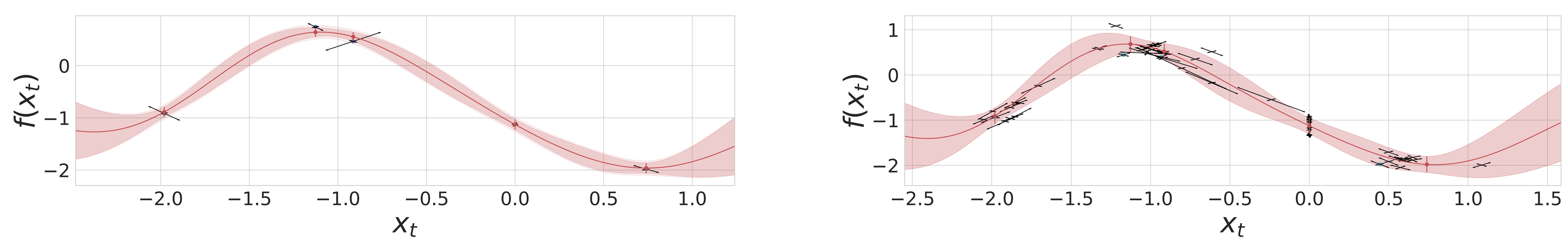

Figure 2: *(Left)* $q(f)$ from $q(x)$ and *(right)* x filtered from $p(f|v)$ for sampled v. Inducing points, predictive mean and confidence region in red, x confidence intervals in black.

Inferring dynamics (f) given state (x) distributions

If x were known, we have input-output pairs for f; inference is exact GP regression, which has an analytic solution.

If we have x only in distribution, we have noisy input-output pairs. The standard approach is to introduce M deterministic inputs z and corresponding outputs $v \sim q(v)$, then (z, v) stand in for input-output pairs for regression. Then,

$$p(f|y) \approx q(f) = q(f, v) = q(v)p(f|v) \tag{2}$$

We can make $q(v)$ Gaussian, and optimise it and z as variational parameters.

Existing methods are biased or memory intensive

The objective is

$$\mathcal{F} = \int \log p(y|x)dq(x) - D_{KL}(q(v)\|p(v)) - \int\int D_{KL}(q(x|f)\|p(x|f))\, p(f|v, z)df\, q(v)dv \leq p(y). \tag{3}$$

Then the maximising process noise variance Q for T latent time points is

$$Q_{\text{opt}} = \frac{1}{T}\sum_{t=1}^{T}(\tilde{Q}_{t-1} + \mathbb{E}[(h_{t-1}(x) - f(x))(h_{t-1}(x) - f(x))^\top]) \tag{4}$$

which is biased larger wherever

- $h \perp f$ (the mean-field approximation)
- or h cannot approximate the shape of f well (e.g. h is linear and we have very noisy observations).

Previous approaches mostly parameterise $A_t, b_t, \tilde{Q}_t$ directly. But **the optimal Gaussian filter is the SLF**. An optimal iterative Gaussian smoother can also be described.

These have **no additional parameters** (reduction from $O(TD^2)$ to $O(M^2D^2)$ where the number of inducing points M could be $O(\log T)$).

Proposal: locally linearised GPSSM

The proposal is to construct $q(x|f) = q(x|v)$ by sampling inducing points, then carrying out Kalman-like smoothing using $p(f|v)$.

Initial results are promising. An interesting extension would be to break the linearity constraint by using a particle filter instead.

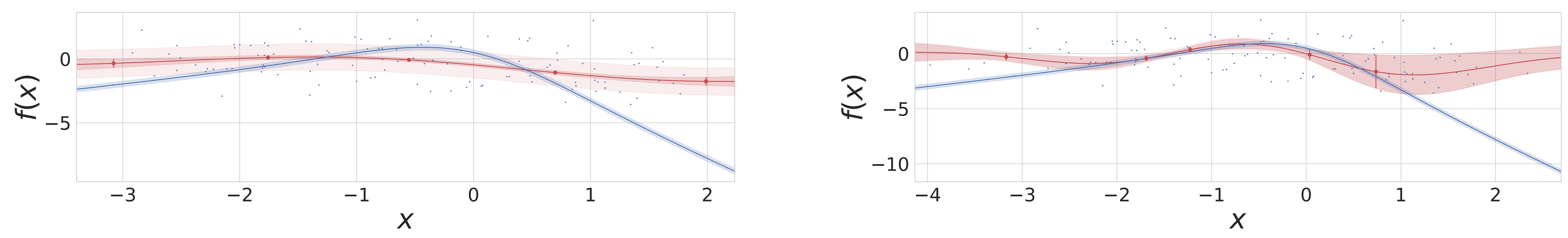

Figure 3: *(Left)* The fit for the mean field case. *(Right)* The fit for the proposed method, with $q(x|f)$ constructed using statistically linearised smoothing. Ground truth in blue, model fit in red. The faint blue dots show the observed pairs (y_t, y_{t+1}). The correlated approach has much better calibrated posterior uncertainty.