Novel approximations for inference and learning in nonlinear dynamical systems

Alexander Ypma Tom Heskes

SNN, Geert Grooteplein 21, 6525 EZ, University of Nijmegen

Abstract. We formulate the problem of inference in nonlinear dynamical systems (NLDS) in the Expectation-Propagation framework, and propose two novel inference algorithms based on Laplace approximation and the unscented transform (UT). The algorithms are compared empirically and employed as an improved E-step in a conjugate gradient learning algorithm. We illustrate its use for data mining with two high-dimensional time series from marketing research. This contribution is based on work that appeared in [1] and [2]. More information can be obtained via `www.snn.kun.nl/~ypma/papers/list_of_papers.html` or `ypma@snn.kun.nl`. The work is supported by STW, project NNN.5321 "Graphical models for data mining". Data was provided by BrandmarC.

Model. We consider dynamical systems with nonlinearities in the state- and observation equations and additive Gaussian noise,

$$x_t = f(x_{t-1}) + v_t,\ v_t \sim \mathcal{N}(0, Q); \qquad y_t = g(x_t) + w_t,\ w_t \sim \mathcal{N}(0, R) \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ are nonlinear functions. In the well-known Kalman filter and smoother all functions are assumed linear and posterior beliefs on the hidden states can be computed exactly. In the nonlinear model, forward and backward messages cannot be computed exactly any more, so one has to resort to approximations.

Inference with unscented and Laplace approximation. One can express inference in a graphical model as a sequence of multiplications and a summation (or integral) of local factors and messages. In the NLDS model, $p(x_{1:T}, y_{1:T})$ factorizes: $p(x_{1:T}, y_{1:T}) = \prod_t \Psi_t(x_{t-1}, x_t) = \prod_t p(x_t|x_{t-1})p(y_t|x_t)$. Beliefs $p(x_t|y_{1:T})$ are computed by $p(x_t|y_{1:T}) = \hat{\alpha}_t(x_t)\hat{\beta}_t(x_t)$, where $\hat{\alpha}_t(x_t)$ and $\hat{\beta}_t(x_t)$ are the forward and backward messages at x_t. We express a two-slice belief as a product of a two-slice potential and 'incoming messages', $\hat{p}_t(x_{t-1}, x_t) \propto \hat{\alpha}_{t-1}(x_{t-1})\Psi_t(x_{t-1}, x_t)\hat{\beta}_t(x_t)$. Belief $q_t(x_t)$ is obtained as $q_t(x_t) = \text{collapse}\, \hat{p}_t(x_{t-1}, x_t)dx_{t-1}$, where "collapse" involves projection to a Gaussian and marginalization (in this case over x_{t-1}). In our *first approach* we collapse the nongaussian marginal onto a Gaussian by applying Laplace approximation. In our *second approach*, we use the unscented transform to collapse the nongaussian two-slice joint $p_t(x_{t-1}, x_t)$ to a Gaussian, in three steps:

1. *prediction*: approximate $\hat{\alpha}_{t-1}(x_{t-1})\Psi^a_t(x_{t-1}, x_t)$ $p^*_t(x_{t-1}, x_t)$ with UT;
2. *correction*: compute $p^*_t(x_t)$ by marginalization; approximate $p^*_t(x_t)\Psi^b_t(x_t, y_t)$ with a Gaussian $p^*_t(x_t, y_t)$ using UT; incorporate evidence into $p^*_t(y_t|x_t) = p^*_t(x_t, y_t)/p^*_t(x_t)$, resulting in $p^{**}_t(y_t|x_t)$;

3. *combination*: compute $q_t(x_{t-1}, x_t) = p_t^*(x_{t-1}, x_t) p_t^{**}(y_t|x_t)\hat{\beta}_t(x_t)$, and obtain $q_t(x_{t-1})$ and $q_t(x_t)$ by marginalization.

We use UT for computing moments of the joints $p_t^*(x_{t-1}, x_t)$ and $p_t^*(x_t, y_t)$. For example, in the prediction step we need to compute

$$\int\int \hat{\alpha}_{t-1}(x_{t-1})\Psi_t^a(x_{t-1}, x_t)h(x_{t-1}, x_t)dx_{t-1}dx_t \approx \sum_i w_i \mathcal{F}_h(\chi_i) \tag{2}$$

Here we denote with $h(x_{t-1}, x_t)$ a generalization of $G_\tau^i(x_{t-1}, x_t)$ that also includes the cross-moment of x_{t-1} and x_t. This is needed since we need to compute *all* first- and second-order moments of the two-slice posterior $p_t^*(x_{t-1}, x_t)$ using UT. With $\mathcal{F}_h$ we express that the 'effective nonlinearity' through which the samples have to be propagated is now determined by $h(x_{t-1}, x_t)$ as well.

Conjugate gradient learning. We parameterize the nonlinearities in (1) with radial basis functions ρ_f^i (dynamics) and ρ_g^i (observer), and include weighted inputs u_t. E.g. for the dynamics, $x_{t+1} = \sum_{i=1}^{I_f} h_f^i \rho_f^i(x_t) + A_f x_t + B_f u_t + b_f + v_t \equiv \theta_f \Phi_t^f + v_t$ where $v_t \sim \mathcal{N}(0, Q)$ and $\rho_f^i(x_t)$ are Gaussians in x_t space. We then compute the gradient of the loglikelihood $\mathcal{L}$ with respect to Q and θ_f as

$$\begin{aligned} \nabla_Q(\mathcal{L}) &= \frac{1}{2}Q^{-1}SQ^{-1} - \frac{J}{2}Q^{-1} \\ \nabla_{\theta_f}(\mathcal{L}) &= Q^{-1}\left(\sum_t \langle x_{t+1}\Phi_t^{f,T}\rangle_t - \theta_f \sum_t \langle \Phi_t^f \Phi_t^{f,T}\rangle_t\right) \end{aligned} \tag{3}$$

Results. In experiments with a one-dimensional nonlinear dynamical system, our unscented algorithm proved to be robust and consistently better than extended and unscented Kalman filtering. Our Laplace algorithm allowed for the best estimates of the hidden state means, but also proved less robust to high observation noise. We then applied our combined inference-learning algorithm to the task of data mining of marketing time series, where the underlying assumption is that a marketing steering variable has both an immediate influence on the output (via the observer) and a delayed influence via the dynamics (e.g. when 'the general opinion' about a brand gradually changes as a result of PR activities).

References

[1] A. Ypma and T. Heskes. Novel approximations for inference and learning in nonlinear dynamical systems. In *Proceedings of 12th European Symposium on Artificial Neural Networks ESANN'2004*, pages 361 – 366, 2004.

[2] A. Ypma and T. Heskes. Novel approximations for inference in nonlinear dynamical systems using expectation propagation. *Neurocomputing - Special issue on Neural Networks for Signal Processing*, 2004.