EXPECTATION PROPAGATION FOR INFERENCE IN NON-LINEAR DYNAMICAL MODELS WITH POISSON OBSERVATIONS

Byron M. Yu[1], *Krishna V. Shenoy*[1,2], *Maneesh Sahani*[3]

[1]Dept. of Electrical Engineering, [2]Neurosciences Program, Stanford University, Stanford, CA, USA
[3]Gatsby Computational Neuroscience Unit, UCL, London, UK

ABSTRACT

Neural activity unfolding over time can be modeled using non-linear dynamical systems [1]. As neurons communicate via discrete action potentials, their activity can be characterized by the numbers of events occurring within short predefined time-bins (*spike counts*). Because the observed data are high-dimensional vectors of non-negative integers, non-linear state estimation from spike counts presents a unique set of challenges. In this paper, we describe why the expectation propagation (EP) framework is particularly well-suited to this problem. We then demonstrate ways to improve the robustness and accuracy of Gaussian quadrature-based EP. Compared to the unscented Kalman smoother, we find that EP-based state estimators provide more accurate state estimates.

1. INTRODUCTION

Consider the following dynamical system for modeling neural spike counts:

$$\mathbf{x}_t \mid \mathbf{x}_{t-1} \sim \mathcal{N}\left(\mathbf{f}\left(\mathbf{x}_{t-1}\right),\, Q\right) \tag{1a}$$

$$y_t^i \mid \mathbf{x}_t \sim \text{Poisson}\left(\lambda_i\left(\mathbf{x}_t\right) \cdot \Delta\right), \tag{1b}$$

where $\mathbf{x}_t \in \mathbb{R}^{p \times 1}$ is the state vector at time $t = 1, \ldots, T$, and $y_t^i \in \{0, 1, 2, \ldots\}$ is the corresponding observed spike count for neuron $i = 1, \ldots, q$ taken in a time bin of width Δ. The functions $\mathbf{f} : \mathbb{R}^{p \times 1} \to \mathbb{R}^{p \times 1}$ and $\lambda_i : \mathbb{R}^{p \times 1} \to \mathbb{R}_+$ are, in general, non-linear. The initial state $\mathbf{x}_1$ is Gaussian-distributed. For notational compactness, the spike counts for all q simultaneously-recorded neurons are assembled into a $q \times 1$ vector $\mathbf{y}_t$, whose ith element is y_t^i. Note that the observations are discrete-valued and that, typically, $q \gg p$.

Given sequences of observed spike counts from a group of simultaneously-recorded neurons, we would like to estimate both the state $\mathbf{x}_t$ at each timepoint, and the model parameters in (1). This goal can be naturally approached using the Expectation-Maximization algorithm, as in [1]. Here, we focus on the first of the two objectives, namely state estimation. In particular, we seek *smoothed* state estimates, conditioned on all past, present, and future observations (denoted $\{\mathbf{y}\}_1^T$).

The extended Kalman smoother is a common tool for non-linear state estimation; unfortunately, it cannot be directly applied to our problem because the observation noise in (1b) is not additive Gaussian. A possible alternative is the unscented Kalman smoother (UKS) [2, 3], which employs quadrature techniques to approximate multi-dimensional Gaussian integrals that are analytically intractable. For smoothing, the UKS requires that the state dynamics be run backwards in time, either exactly, or approximately using, e.g., a neural network. However, inverting non-linear state dynamics is generally difficult and may not be possible without altering the behavior of the system. Furthermore, the UKS makes Gaussian approximations in the observation space. For discrete-valued observations as in (1b), this approximation may not be appropriate.

Another technique for non-linear state estimation was recently developed [4, 5, 6] using the expectation propagation (EP) framework [7]. By contrast to the UKS, the EP-based approach i) does not require inverting the state dynamics, ii) makes Gaussian approximations only in the state space and not in the observation space, and iii) allows state estimates to be refined iteratively using multiple forward-backward passes. We generally observe tens to hundreds of neurons simultaneously, and the number of spikes emitted by a neuron in a single time bin is most often 0 or 1. Thus, the observations are high-dimensional and distinctly non-Gaussian. In such settings, property ii) above is critical.

The EP framework requires estimating the moments of the joint state posterior distribution $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$. This can be done using either Gaussian quadrature (GQ-EP) [5] or a modal Gaussian approximation (Laplace-EP) [6]. Whereas Laplace-EP estimates moments based on a local region of the distribution, GQ-EP takes into account more global properties of the distribution. While promising, GQ-EP is known to be sensitive to outlying observations [5] and can only be used with quadrature rules that satisfy certain properties.

In the following, we first summarize the EP framework for non-linear state estimation. We then show how the sensitivity to outliers in GQ-EP can be overcome. Next, we demonstrate how quadrature rules that are more accurate than existing ones for GQ-EP can be derived. Finally, we compare the state estimation accuracy of the UKS, GQ-EP, and Laplace-EP techniques for the model neural dynamical system (1).

This work was supported by NIH-NINDS-CRCNS-R01, NSF, NDSEGF, Gatsby, CRF, BWF, ONR, Sloan, and Whitaker.

2. EXPECTATION PROPAGATION

The application of EP [7] to general dynamical models is summarized in this section; for more details, see [4]. We seek to compute the marginal $P\left(\mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$ and pairwise joint $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$ state posteriors. These distributions can be expressed in terms of forward α_t and backward β_t messages as follows

$$P\left(\mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right) = \frac{1}{P\left(\{\mathbf{y}\}_1^T\right)} \alpha_t\left(\mathbf{x}_t\right) \beta_t\left(\mathbf{x}_t\right) \tag{2}$$

$$P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right) = \frac{\alpha_{t-1}\left(\mathbf{x}_{t-1}\right) P\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) P\left(\mathbf{y}_t \mid \mathbf{x}_t\right) \beta_t\left(\mathbf{x}_t\right)}{P\left(\{\mathbf{y}\}_1^T\right)}, \tag{3}$$

where

$$\alpha_t\left(\mathbf{x}_t\right) = P\left(\mathbf{x}_t, \{\mathbf{y}\}_1^t\right) \quad \text{and} \quad \beta_t\left(\mathbf{x}_t\right) = P\left(\{\mathbf{y}\}_{t+1}^T \mid \mathbf{x}_t\right).$$

The messages α_t and β_t are typically approximated by an exponential family density, in our case an unnormalized Gaussian. These approximate messages are then iteratively updated by matching the moments of the marginal posterior (2) with the corresponding moments of the pairwise joint posterior (3). The updates are usually performed sequentially via multiple forward-backward passes. During the forward pass, the α_t are updated while the β_t remain fixed. During the backward pass, the β_t are updated while the α_t remain fixed.

Two different techniques have been proposed to estimate the moments of (3). First, the moments can be expressed as

$$\iint g\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right) P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right) d\mathbf{x}_{t-1} d\mathbf{x}_t \tag{4}$$

for appropriate choices of the function g. For example, if $g\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right) = \mathbf{x}_t$, the mean of $\mathbf{x}_t$ based on the pairwise joint posterior is obtained. By introducing a proposal distribution $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$[1], (4) can be expressed as an integral over a known "weighting" function

$$\iint g\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right) \frac{P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)}{Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)} Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right) d\mathbf{x}_{t-1} d\mathbf{x}_t.$$

Gaussian quadrature [5, 8] approximates this integral by

$$\sum_{j=0}^{n-1} w_j \cdot g\left(\boldsymbol{\chi}_{t-1}^j, \boldsymbol{\chi}_t^j\right) \frac{P\left(\boldsymbol{\chi}_{t-1}^j, \boldsymbol{\chi}_t^j \mid \{\mathbf{y}\}_1^T\right)}{Q\left(\boldsymbol{\chi}_{t-1}^j, \boldsymbol{\chi}_t^j\right)}, \tag{5}$$

where w_j and $[(\boldsymbol{\chi}_{t-1}^j)' \ (\boldsymbol{\chi}_t^j)']'$ are, respectively, the jth quadrature point and weight based on $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$. An example of a quadrature rule (i.e., a set of quadrature points and weights) based on a Gaussian proposal will be given in Section 4. The EP-based state estimator that employs Gaussian quadrature to compute moments is referred to as GQ-EP.

A second way to estimate the moments is to fit a Gaussian to a mode of $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$, as in the Laplace approximation of an integral [6, 9]. The moments of this fitted Gaussian are taken to be the approximate moments of (3). The EP-based state estimator that computes moments in this way is referred to as Laplace-EP.

3. PROPOSAL DISTRIBUTIONS FOR GQ-EP

Using the proposal distribution

$$Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right) \propto \alpha_{t-1}\left(\mathbf{x}_{t-1}\right) \beta_{t-1}\left(\mathbf{x}_{t-1}\right) \alpha_t\left(\mathbf{x}_t\right) \beta_t\left(\mathbf{x}_t\right),$$

with Gaussian messages α_t and β_t, Zoeter and colleagues [5] reported that GQ-EP was sensitive to outlying observations. In particular, the quadrature points may lie in regions where $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$ has negligible density relative to $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$. As a result, covariance matrices estimated from (5) may be ill-conditioned, and GQ-EP becomes largely unusable. Outlying observations are common in the early stages of learning the model parameters, when the parameters are not a good match with the observed data. Even without outlying observations per se, quadrature point locations can be poorly chosen during the first forward pass if $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$ is determined without knowledge of the current observation $\mathbf{y}_t$.

To overcome this problem, we choose $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$ to be a Gaussian matched to the location and curvature of a mode of $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$, as in the Laplace approximation of an integral [9]. Note that this is the same Gaussian used to estimate the moments of $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$ in Laplace-EP, but it is used here as a proposal distribution for GQ-EP. With this choice of proposal distribution, the quadrature points are centered on a mode of $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$, making GQ-EP more robust to outlying observations.

4. QUADRATURE RULES WITH NON-NEGATIVE WEIGHTS

Covariance matrices are formed in (5) by a sum of outer products. If one or more of the quadrature weights w_j is negative, the resulting covariance matrix may have negative eigenvalues. It is important to emphasize that this appearance of negative eigenvalues is not due to numerical instabilities; in particular, if a square-root filter [2] is used, negative quadrature weights may lead to invalid Cholesky updates. Thus quadrature rules with non-negative w_j are necessary to stabilize quadrature-based EP.

Furthermore, evaluating $P\left(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \{\mathbf{y}\}_1^T\right)$ at the quadrature points in (5) requires computing the data likelihood

$$P\left(\{\mathbf{y}\}_1^T\right) = \iint \alpha_{t-1}\left(\mathbf{x}_{t-1}\right) P\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \cdot P\left(\mathbf{y}_t \mid \mathbf{x}_t\right) \beta_t\left(\mathbf{x}_t\right) d\mathbf{x}_{t-1} d\mathbf{x}_t. \tag{6}$$

[1] $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$ determines the distribution of quadrature points and so is referred to as a proposal distribution by analogy to importance sampling.

This integral is generally analytically intractable, and must also be approximated by Gaussian quadrature (frequently using the same proposal distribution $Q\left(\mathbf{x}_{t-1}, \mathbf{x}_t\right)$). Once again, negative quadrature weights may lead to instability, here in the form of an impossible negative likelihood estimate.

Here, we consider two quadrature rules with non-negative weights. For notational clarity, a Gaussian integral is approximated by Gaussian quadrature as follows

$$\int h(\mathbf{z})\, \mathcal{N}\left(\mathbf{z};\, \boldsymbol{\mu},\, \Sigma\right) d\mathbf{x} \approx \sum_{j=0}^{n-1} w_j h\left(\mathbf{z}_j\right), \tag{7}$$

where $\mathbf{z} \in \mathbb{R}^{r\times 1}$, $\boldsymbol{\mu} \in \mathbb{R}^{r\times 1}$, $\Sigma \in \mathbb{R}^{r\times r}$, h is a deterministic nonlinear function, $\mathbf{z}_0, \ldots, \mathbf{z}_{n-1}$ are the quadrature points, and $w_0, \ldots, w_{n-1}$ are the quadrature weights. The first quadrature rule is the classical precision 3 rule [2, 8, 10], which prescribes the following points and weights

$$\begin{aligned} \mathbf{z}_0 &= \boldsymbol{\mu} & w_0 &= 1 - \frac{r}{\gamma^2} \\ \mathbf{z}_i &= \boldsymbol{\mu} + \gamma\left(\sqrt{\Sigma}\right)_i & w_i &= \frac{1}{2\gamma^2} \\ \mathbf{z}_{r+i} &= \boldsymbol{\mu} - \gamma\left(\sqrt{\Sigma}\right)_i & w_{r+i} &= \frac{1}{2\gamma^2}, \end{aligned} \tag{8}$$

where $i = 1, \ldots, r$ and $\gamma \in \mathbb{R}$ is a free parameter. $\left(\sqrt{\Sigma}\right)_i$ is the ith column of $R \in \mathbb{R}^{r\times r}$, where $RR' = \Sigma$. The number of quadrature points is $n = 2r + 1$. This quadrature rule is exact if $h(\mathbf{z})$ in (7) is a monomial of degree 3 or less. Furthermore, as long as γ is chosen such that $\gamma^2 \geq r$, the quadrature weights w_j in (8) are non-negative.

The second quadrature rule is a custom "precision 3" rule derived using Gaussian processes (GPs) under the constraint of non-negative weights. Whereas the classical rule achieves zero error for monomials of degree 3 or less and offers no guarantees for monomials of higher degree, the custom rule minimizes the average error across an entire family of functions. In the GP approach, the task of selecting quadrature points and weights is formulated as an optimization problem. The details of how to derive quadrature rules in this way can be found in [11]; here, we describe how this technique was applied to derive the custom "precision 3" rule. We first transformed the unconstrained optimization problem into a constrained optimization problem by introducing a non-negativity constraint on the quadrature weights. Assuming the same constellation of quadrature points as in (8) up to the scaling factor γ, the optimization problem was then solved to obtain γ and a new set of quadrature weights $w_0, \ldots, w_{2r}$. Note that these optimized weights will not necessarily be the same as the classical weights of (8). A GP requires the specification of a covariance function. We chose the commonly-used radial basis function

$$K(\mathbf{z}_j, \mathbf{z}_k) = e^{-\frac{b}{2}\|\mathbf{z}_j - \mathbf{z}_k\|^2}, \tag{9}$$

where the free parameter b sets the relative importance of monomials of varying degree. As $b \to 0$, monomials of lower degree have priority. This GP approach is general and can be used to derive other quadrature rules with non-negative weights.

In the classical precision 3 rule (8), only the central quadrature weight w_0 can be negative. Julier and colleagues [12] recognized that, if a covariance estimate is expanded about a point away from the estimated mean, positive semidefiniteness can be guaranteed even though $w_0 < 0$. To illustrate this, let $\mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\mu},\, \Sigma\right)$ and $h(\mathbf{z})$ be a column vector. The estimated covariance of $h(\mathbf{z})$ using Gaussian quadrature is

$$\widehat{C} = \sum_{j=0}^{n-1} w_j \left[h\left(\mathbf{z}_j\right) - \widehat{\mathbf{m}}\right]\left[h\left(\mathbf{z}_j\right) - \widehat{\mathbf{m}}\right]', \tag{10}$$

where $\widehat{\mathbf{m}}$ is the estimated mean of $h(\mathbf{z})$ from (7). Julier and colleagues expanded $\widehat{C}$ about $h(\mathbf{z}_0)$ rather than $\widehat{\mathbf{m}}$. As a result, the $j = 0$ term disappears and all remaining terms have positive quadrature weights. The UKS tested in Section 5 uses this expansion. While effective for the precision 3 rule, this expansion doesn't generalize to the precision 5 rule [8, 10], where multiple quadrature weights can be negative. Furthermore, this technique cannot be used to estimate data likelihoods.

Another way to ensure positive semidefiniteness is to use a one-dimensional quadrature rule along each dimension of $\mathbf{z}$, rather than a multi-dimensional rule such as (8). However, the number of quadrature points required would grow exponentially, rather than linearly, with r. In addition, Lerner [8] showed how to project a covariance matrix with predominantly known components to the positive semidefinite cone. However, applying this technique to problems discussed in this paper would require extension to covariance matrices whose elements are entirely unknown.

5. RESULTS

We compare here the state estimation accuracy of the UKS, GQ-EP, and Laplace-EP for state dimensionalities $p = 3, 10$ and observation dimensionality $q = 100$ (Table 1). We generated 50 state trajectories, each with 50 time points, and corresponding spike counts from the models (1a) and (1b), where

$$\mathbf{f}(\mathbf{x}) = (1 - k)\,\mathbf{x} + k \cdot W \cdot \mathrm{erf}(\mathbf{x}) \tag{11}$$

$$\lambda_i(\mathbf{x}) = \log\left(1 + e^{\mathbf{c}_i'\mathbf{x} + d_i}\right). \tag{12}$$

with the error function (erf) acting element-by-element on its argument. The model parameters $W \in \mathbb{R}^{p\times p}$, $\mathbf{c}_i \in \mathbb{R}^{p\times 1}$, and $d_i \in \mathbb{R}$ were randomly chosen within a range that provided biologically realistic spike counts (typically, 0 or 1 spike in each bin). This procedure was repeated three times for each state dimensionality. The time constant $k \in \mathbb{R}$ was set to 0.1.

	$p = 3$	$p = 10$
UKS	1.94±0.02	4.10±0.03
GQ-EP, classical	0.93±0.01	2.62±0.02
GQ-EP, custom	0.93±0.01	2.35±0.02
Laplace-EP	0.94±0.01	2.22±0.01

Table 1. Root-mean-square error (mean±sem) between the actual and estimated state trajectories.

For the UKS, we applied the classical precision 3 rule with $\gamma = \sqrt{3}$, which yields quadrature points that match some fourth order moments of a Gaussian distribution [12]. The UKS requires computing and inverting a predicted observation covariance $P_{yy} \in \mathbb{R}^{q \times q}$ [2, 3, 10]. Because the observations here are high-dimensional, with a large number of elements equal to 0, P_{yy} was usually ill-conditioned. Thus, to make the inversion possible, we added a constant (0.5, which was determined by a systematic sweep) to the diagonal elements of P_{yy}, by analogy to ridge regression. For the UKS backward pass, we defined an artificial state prior and approximated the backward-time dynamics with a linear-Gaussian system [3].

For the EP-based estimators, the results are based on a single forward-backward pass. GQ-EP was tested using the modal Gaussian proposal distribution from Section 3 in tandem with each of the two quadrature rules from Section 4. For the classical rule (8), we set $\gamma^2 = r$ to ensure non-negative quadrature weights. Larger values of γ led to higher estimation errors.

The UKS yielded higher estimation errors than the EP-based estimators because i) it makes Gaussian approximations in the observation space where the data are distinctly non-Gaussian, and ii) it approximates the backward-time dynamics of the non-linear system (1a) using a linear-Gaussian system. For $p = 3$, the three EP-based estimators provide comparable performance. However, for $p = 10$, Laplace-EP is preferred and the custom quadrature rule that we derived using Gaussian processes outperforms the classical rule.

Higher precision quadrature rules have been proposed (e.g., precision 5 rules [8, 10]), but the techniques used to guarantee positive semidefinite covariances and non-negative data likelihoods for the classical precision 3 rule don't apply. In particular, there is no free parameter that can be chosen to keep weights non-negative. Furthermore, because more than one weight can be negative, it is not possible to guarantee valid covariances by expanding about a different point. We are currently developing quadrature rules that further improve the estimation accuracy of GQ-EP, especially at higher state dimensionalities.

6. REFERENCES

[1] B.M. Yu, A. Afshar, G. Santhanam, S.I. Ryu, K.V. Shenoy, and M. Sahani, "Extracting dynamical structure embedded in neural activity," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, Cambridge, MA, 2006.

[2] E. A. Wan and R. van der Merwe, "The unscented Kalman filter," in *Kalman Filtering and Neural Networks*, S. Haykin, Ed., chapter 7. Wiley Publishing, 2001.

[3] M. Briers, A. Doucet, and S. Maskell, "Smoothing algorithms for state-space models," Tech. Rep. CUED/F-INFENG/TR.498, Cambridge University Engineering Department, Aug. 2004.

[4] T. Heskes and O. Zoeter, "Expectation propagation for approximate inference in dynamic Bayesian networks," in *Proceedings UAI-2002*, A. Darwiche and N. Friedman, Eds., 2002, pp. 216–223.

[5] O. Zoeter, A. Ypma, and T. Heskes, "Improved unscented Kalman smoothing for stock volatility estimation," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, A. Barros, J. Principe, J. Larsen, T. Adali, and S. Douglas, Eds., 2004.

[6] A. Ypma and T. Heskes, "Novel approximations for inference in nonlinear dynamical systems using expectation propagation," *Neurocomputing*, vol. 69, pp. 85–99, 2005.

[7] T.P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[8] U.N. Lerner, *Hybrid Bayesian Networks for Reasoning about Complex Systems*, Ph.D. thesis, Stanford University, 2002.

[9] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[10] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[11] T.P. Minka, "Deriving quadrature rules from Gaussian processes," Tech. Rep., 2000, `http://research.microsoft.com/~minka/papers/quadrature.html`.

[12] S. Julier, J. Uhlmann, and H.F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 477–482, Mar. 2000.