LETTER Communicated by Raj Rao

Exact Inferences in a Neural Implementation of a Hidden Markov Model

Jeffrey M. Beck
jbeck@bcs.rochester.edu
Alexandre Pouget
alex@bcs.rochester.edu
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, U.S.A.

From first principles, we derive a quadratic nonlinear, first-order dynamical system capable of performing exact Bayes-Markov inferences for a wide class of biologically plausible stimulus-dependent patterns of activity while simultaneously providing an online estimate of model performance. This is accomplished by constructing a dynamical system that has solutions proportional to the probability distribution over the stimulus space, but with a constant of proportionality adjusted to provide a local estimate of the probability of the recent observations of stimulus-dependent activity-given model parameters. Next, we transform this exact equation to generate nonlinear equations for the exact evolution of log likelihood and log-likelihood ratios and show that when the input has low amplitude, linear rate models for both the likelihood and the log-likelihood functions follow naturally from these equations. We use these four explicit representations of the probability distribution to argue that, in contrast to the arguments of previous work, the dynamical system for the exact evolution of the likelihood (as opposed to the log likelihood or log-likelihood ratios) not only can be mapped onto a biologically plausible network but is also more consistent with physiological observations.

1 Introduction

It is becoming increasingly clear that some of the computations performed by the nervous system are akin to Bayesian inferences, whether in the context of sensory perception, reasoning, or motor control (Carpenter & Williams, 1995; Gold & Shadlen, 2001; Knill & Pouget, 2004). However, the neural mechanisms of these inferences remain unclear. Here, we explore one particular neural implementation of Bayesian inferences for hidden Markov models (HMMs). HMMs are particularly appealing because they can be used to model a wide variety of problems involving dynamic stimuli such as cue integration, decision making, and motor control. The goal is to understand how a neural network can infer $\Pr(\theta(t)|\vec{v}(t), \ldots, \vec{v}(0))$, where

$\theta(t)$ is a time-varying stimulus and $\{v(t), \ldots, v(0)\}$ is a set of observations or measurements related to $\theta(t)$, where $\theta(t)$ could be the position of a moving object inferred from a sequence of images $\{v(t), \ldots, v(0)\}$. Consistent with the work of others, we will limit ourselves to the consideration of recurrent networks for which the firing rate $u_i(t)$ of neuron i is related to the probability that the stimulus takes on some value θ_i through some monotonically increasing function,

$$u_i(t) = f(\Pr(\theta(t) = \theta_i | \vec{v}(s) \forall s < t)) \tag{1.1}$$

where $f(x)$ is typically equal to the identity, a convolution (Anderson, 1994), or a log transform (Rao, 2004; Yu & Dayan, 2005; Denève, 2005).

Exact solutions to this problem have recently been derived for a single neurons with a binary variable θ (Denève, 2005) and for multiple neurons with a static variable θ (Koechlin, Anton, & Burnod, 1999). However, for nontrivial, dynamic $\theta(t)$ and a population of neurons, the only available solution relies on the assumption that the log of a sum can be approximated by the sum of logs (Rao, 2004). In particular, this study showed that when $f(x) = \log(x)$, a linear dynamical system of the form

$$\frac{d\vec{u}}{dt} = -\vec{u} + \mathbf{W}\vec{u} + \mathbf{M}\vec{v}, \tag{1.2}$$

where $\mathbf{W}$ is the recurrent connectivity matrix and $\mathbf{M}$ is the feedforward connectivity matrix, approximates Bayes inferences for such an HMM after the weight matrix $\mathbf{W}$ was adjusted by a numerical fit.

Here we present a simple, rigorous, and systematic derivation of a quadratic, nonlinear first-order dynamical system that has parameters that may be linked directly to Bayes rule and has solutions that give the exact probability that a Markov stimulus currently takes on a particular value given the entire history of the observed pattern of activity. Furthermore, it will be shown that such a system follows solely from the assumptions that the stimulus is Markov and that the stimulus-dependent pattern of activity consists of a collection of independent Poisson processes with stimulus-dependent (and nonzero) rates. A companion equation for the evolution of the log likelihood of the observed stimulus-dependent activity given the model parameters will also be derived. We argue that this equation provides the proper objective function for a Bayes-linked learning rule and indicates that the goal of such a network is, in fact, the mean prediction of the input patterns. Moreover, we will demonstrate that the rate of increase of the probability of the data given model parameters can be incorporated into the amplitude of the output probability distribution in a manner that allows for an online estimate of model performance.

For the purposes of comparison with previous work, we then generate the associated nonlinear evolution equations for the log-likelihood and

log-likelihood ratios. The assumption of low-amplitude input will then be applied to the log-likelihood equations to demonstrate that the recurrent weights of the linearized model may be calculated directly, as opposed to by a numerical fit as in Rao (2004). We then show that the assumption of low-amplitude input may also be applied to the nonlinear equation for the evolution of the likelihood function to obtain an equivalent linear dynamical system with an associated Hebbian learning rule. Finally, we discuss the biological plausibility of the various dynamical systems for Bayesian inference that we have generated and argue the case for a neural implementation, which directly represents the amplitude-adjusted probability density function as opposed to a representation in the log likelihood or log-likelihood ratio domain.

2 Model

To demonstrate the distinction between our approach and that of Rao (2004) and Denève (2005), we begin by replicating their derivation of a discrete dynamical system that performs exact Bayesian inferences. We begin with a formulation of Bayes' rule that calculates the probability that at the time step n, a particular stimulus (θ_i) is being presented given the total history of all previous stimulus-dependent activity ($\vec{v}^n, \vec{v}^{n-1}, \ldots, \vec{v}^0$). Thus, if we denote $\theta^n = \theta_i$ as θ_i^n, we may write the conditional probability that the stimulus takes on some value given the history of the activity pattern as

$$\Pr\left(\theta_i^n|\vec{v}^n, \vec{v}^{n-1}, \ldots, \vec{v}^0\right) = \frac{\Pr\left(\vec{v}^n|\theta_i^n, \vec{v}^{n-1}, \ldots, \vec{v}^0\right)}{\Pr\left(\vec{v}^n|\vec{v}^{n-1}, \ldots, \vec{v}^0\right)} \times \sum_j \Pr\left(\theta_i^n|\theta_j^{n-1}, \vec{v}^{n-1}, \ldots, \vec{v}^0\right) \Pr\left(\theta_j^{n-1}|\vec{v}^{n-1}, \ldots, \vec{v}^0\right). \tag{2.1}$$

The assumptions of a first-order Markov process for the stimulus and conditional independence of $\vec{v}^n$ given θ_i^n (see Figure 1) imply that a discrete dynamical system for the quantity

$$u_i^n = \Pr(\theta_i^n|\vec{v}^n, \vec{v}^{n-1}, \ldots, \vec{v}^0), \tag{2.2}$$

is given by

$$u_i^n = \frac{\Pr\left(\vec{v}^n|\theta_i^n\right)}{\Pr\left(\vec{v}^n|\vec{v}^{n-1}, \ldots, \vec{v}^0\right)} \sum_j \Pr\left(\theta_i^n|\theta_j^{n-1}\right) u_j^{n-1}. \tag{2.3}$$

At this point, Rao (2004) takes the natural log of this equation so that dependence on the noisy input ($\vec{v}^n$) comes in as an additive driving force

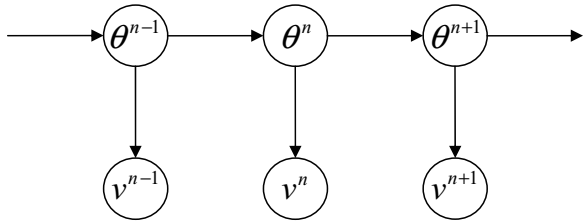

Figure 1: Graphical representation of a hidden Markov model. Arrows indicate conditional independence.

rather than a multiplicative one. Unfortunately, this causes the transition probabilities to come into the equation through a term that is the natural log of a sum. This nonlinearity is eliminated through a numerical parameter fit. It is then assumed that both the additive driving term and the resulting linear approximation to the transition probability term are of order Δt, so that a first-order linear differential equation may be obtained. In contrast, here we start with the goal of obtaining a temporally continuous first-order dynamical system and explicitly delineate the assumptions necessary to come to this conclusion. In particular, we begin by noting that for a discrete system of the form

$$\vec{u}^n = \vec{h}(\vec{u}^{n-1}, \vec{v}^n, \Delta t), \tag{2.4}$$

a first-order dynamical system can be obtained when

$$\vec{h}(\vec{u}^{n-1}, \vec{v}^n, \Delta t) = \vec{u}^{n-1} + \Delta t \vec{h}^1(\vec{u}^{n-1}, \vec{v}^n) + O(\Delta t^2). \tag{2.5}$$

It can be shown that to put equation 2.3 into this form, we need only make the reasonable assumptions that (1) the probability of transitioning from state j to state $i \neq j$ is proportional to the discrete time interval Δt and that (2) the ability of a particular pattern of activity to demonstrate a preference for a given stimulus is also proportional to Δt, the time interval over which the pattern of activity is observed. Written in terms of the relevant probability distributions, assumption 1 is given by

$$\Pr\left(\theta_i^n | \theta_j^{n-1}\right) = \delta_{ij} + \Delta t w_{ij}, \tag{2.6}$$

where δ_{ij} is the Kronecker delta and $\sum_i w_{ij} = 0$, and $w_{ij} > 0$ for all $i \neq j$. Similarly, assumption 2 is given by

$$\frac{\Pr\left(\theta_i^n | \vec{v}^n\right)}{\Pr\left(\theta_i^n\right)} = 1 + \Delta t d_i^n + O(\Delta t^2), \tag{2.7}$$

where $d_i^n = d_i(v^n)$ are subject to the condition $\vec{d}^n \cdot \vec{b} = \vec{0}$ for all n, for $b_i = \Pr(\theta_i^n)$ gives the prior or bias in the stimulus. Note that d_i^n can be thought of as the growth rate of the conditional preference of the probability of θ_i^n given v^n.

Clearly, assumption 1 is just a restatement of the assumption that the stimulus is governed by a continuous-time Markov process and thus has Poisson transition times. Assumption 2 is slightly more restrictive. It is easy to show that any set of discrete-time HMMs that converges to a continuous HMM as Δt goes to zero has the property that the left-hand side of equation 2.7 has a Taylor expansion in Δt. However, it is not generally the case that the first term of this expansion is equal to one as is required to put equation 2.3 into the form of equation 2.5. Fortunately, a wide class of biologically plausible stimulus-dependent patterns of activity has this property. Specifically, a layer of cortex modeled by a set of independent Poisson neurons with stimulus-dependent rates has this property, as does a layer cortex modeled by a covariant gaussian distribution associated with a dense set of tuning curves. (See appendix A for an example of the Poisson case.)

Regardless, equations 2.6 and 2.7 can be used with equation 2.3 after another application of Bayes' rule,

$$\Pr(\vec{v}^n|\theta_i^n) = \frac{\Pr\left(\theta_i^n|\vec{v}^n\right)\Pr(\vec{v}^n)}{\Pr\left(\theta_i^n\right)}, \tag{2.8}$$

to yield

$$\Pr\left(\vec{v}^n|\vec{v}^{n-1}\ldots\vec{v}^0\right)\vec{u}^n = \Pr(\vec{v}^n)\left(\mathbf{I}+\Delta t\mathbf{D}^n\right)\left(\mathbf{I}+\Delta t\mathbf{W}\right)\vec{u}^{n-1} + O(\Delta t^2), \tag{2.9}$$

where $\mathbf{D}^n$ is a diagonal matrix with diagonal entries given by the d_i^n defined by equation 2.7, $\mathbf{W}$ is the input-independent, recurrent weight matrix with elements given by transition probabilities, and $\mathbf{I}$ is the identity matrix.

To obtain an expression for $\Pr\left(\vec{v}^n|\vec{v}^{n-1}\ldots\vec{v}^0\right)$, we simply take the dot product of equation 2.9 with $\vec{\mathbf{1}}$, the vector composed entirely of ones, while recalling that $\vec{\mathbf{1}}\cdot\vec{u}^n = \vec{\mathbf{1}}\cdot\vec{u}^{n-1} = 1$ since both $\vec{u}^n$ and $\vec{u}^{n-1}$ represent probability distribution. This yields

$$\Pr\left(\vec{v}^n|\vec{v}^{n-1}\ldots\vec{v}^0\right) = \Pr(\vec{v}^n)\left[1+\Delta t\vec{\mathbf{1}}\cdot\left(\mathbf{D}^n+\mathbf{W}\right)\vec{u}^{n-1}+O\left(\Delta t^2\right)\right]. \tag{2.10}$$

Then, taking the ratio of equations 2.9 and 2.10, we obtain

$$\vec{u}^n = \frac{\vec{u}^{n-1} + \Delta t\{(\mathbf{D}^n + \mathbf{W})\vec{u}^{n-1}\} + O(\Delta t^2)}{1 + \Delta t\{\vec{\mathbf{1}} \cdot (\mathbf{D}^n + \mathbf{W})\vec{u}^{n-1}\} + O(\Delta t^2)}, \tag{2.11}$$

where the denominator corresponds to a form of divisive normalization (Heeger, 1992; Nelson, 1994), which, in this particular context, is required to ensure that $\vec{u}(t)$ is a probability distribution. Expanding this equation to first order in Δt, we obtain

$$\vec{u}^n = \vec{u}^{n-1} + \Delta t\{(\mathbf{D}^n + \mathbf{W})\vec{u}^{n-1} - \vec{\mathbf{1}} \cdot (\mathbf{D}^n + \mathbf{W})\vec{u}^{n-1}\} + O(\Delta t^2). \tag{2.12}$$

Taking the limit as Δt goes to zero, we find that

$$\frac{d\vec{u}}{dt} = \mathbf{D}(t)\vec{u} + \mathbf{W}\vec{u} - (\vec{d}(t) \cdot \vec{u})\vec{u}, \tag{2.13}$$

where we have used the definition of the matrix $\mathbf{D}(t)$ and the fact that $\vec{\mathbf{1}} \cdot \mathbf{W} = \vec{\mathbf{0}}^T$ to simplify the uniform inhibitory term. Recalling that $\mathbf{D}(t)$ is a function of the time-dependent stimulus $v(t)$, we see that we have obtained a first-order quadratic nonlinear recurrent network for the exact computation of Bayesian inferences. Additionally, we note that the quadratic inhibitory term in equation 2.13 is a result of the divisive normalization in equation 2.11, and has the same effect in both equations. For example, when $\mathbf{D}(t)$ is constant, the solution to equation 2.13 is

$$\vec{u}(t) = \frac{\exp[(\mathbf{D} + \mathbf{W})t]\vec{u}_0}{(1 - \vec{\mathbf{1}} \cdot \vec{u}_0) + \vec{\mathbf{1}} \cdot \exp[(\mathbf{D} + \mathbf{W})t]\vec{u}_0}. \tag{2.14}$$

This expression is similar to the divisive normalization reported in cortical circuits (Heeger, 1992; Nelson, 1994).

3 Estimating Model Performance

Solutions to equation 2.13 have the property that $\vec{\mathbf{1}} \cdot \vec{u}(t) = 1$ for all time. As noted in previous work (Sahani & Dayan, 2003) this property, while useful in the context of the above derivation, is not a necessary feature of the activity of a recurrent network, which directly represents a probability distribution and may be used to represent some other variable. In this context, the natural choice would be to use the amplitude to represent an estimate of model performance.

The performance of the model can be assessed at any given time by computing the likelihood of the data under the current model parameters M. A procedure similar to that used above may be used to generate an

expression for the evolution of this quantity, leading to

$$L^n = \Pr(\vec{v}^n, \vec{v}^{n-1} \ldots \vec{v}^0 | \mathrm{M}) = \Pr(\vec{v}^n | \vec{v}^{n-1} \ldots \vec{v}^0, \mathrm{M}) L^{n-1}. \tag{3.1}$$

Using equation 2.9 to evaluate the conditional dependence of the current activity pattern on the history of the activity pattern, we see that

$$L^n = L^{n-1} \Pr(\vec{v}^n | \mathrm{M})(1 + \Delta t \vec{d}^n \cdot \vec{u}^{n-1}). \tag{3.2}$$

Taking the natural log, expanding to first order in Δt, and eliminating terms that are independent of model parameters yields

$$L^*(t) - L^*(0) = \int_0^t \vec{d}(s) \cdot \vec{u}(s) ds, \tag{3.3}$$

where $L^*(t)$ is a function optimized by the same model parameters that optimize the $L(t)$ when the model parameters accurately represent the average stimulus-independent and time-independent behavior of the input patterns, that is, when terms of the form $\Pr(\vec{v}^n | \mathrm{M})$ may be neglected. (See appendixes A and B for details.) Thus, the expected value of $\vec{d} \cdot \vec{u}$ provides the proper objective function for evaluating the performance of the model while a local average of $\vec{d} \cdot \vec{u}$ provides a local estimate of model performance.

Our goal is to find a differential equation over a new variable $\tilde{u}$, which has a solution proportional to the solution of equation 2.13: $\tilde{u}(t) = \alpha(t)\vec{u}(t)$, and for which $\alpha(t)$ can be related to a local average of $\vec{d} \cdot \vec{u}$. This can be achieved by suitably adjusting the term responsible for divisive normalization. In particular, a dynamical system of the form

$$\frac{d\tilde{u}}{dt} = (\mathrm{D}(t) + \mathrm{W})\,\tilde{u} - f(\alpha)\,\tilde{u} \tag{3.4}$$

has a solution given by $\tilde{u}(t) = \alpha(t)\vec{u}(t)$, where $\alpha(t)$ obeys

$$\frac{d \log(\alpha)}{dt} = \left(\vec{d} \cdot \vec{u} - f(\alpha)\right). \tag{3.5}$$

Thus, if $f(\alpha)$ is a monotonically increasing function, this quantity is attracted to a local average of $\vec{d} \cdot \vec{u}$ as desired. In particular, if $f(\alpha) = \beta \log(\alpha)$, then $\beta \log(\alpha)$ is a local average of $\vec{d} \cdot \vec{u}$. Moreover, since the effective time constant of equation 3.5 is given by the inverse of the parameter β, we may adjust the effective time window over which the "local" average of model performance is taken by simply manipulating this parameter.

Of course, our choice of $f(\alpha)$, the coefficient of this divisive normalization term in equation 3.4, is fairly arbitrary. Indeed, if we wish to restrict

ourselves to dynamical systems with quadratic nonlinearities, then we should choose $f(\alpha)$ to be linear and obtain

$$\frac{d\tilde{u}}{dt} = (\mathbf{D}(t) + \mathbf{W} + \beta\mathbf{I})\tilde{u} - \gamma(\vec{n} \cdot \tilde{u})\tilde{u}, \tag{3.6}$$

which has solutions given by $\tilde{u} = \alpha(t)\vec{u}$ but with $\alpha(t)$ obeying the equation

$$\frac{d\alpha}{dt} = \alpha\big(\vec{d} \cdot \vec{u} + \beta - \gamma\alpha\big), \tag{3.7}$$

where β and γ may be used to adjust the effective time constant of integration for the evolution of the amplitude.

4 Special Case: Log Probability

The equations we have derived so far are general in the sense that they make no assumption about the mapping between the activity of the neurons and the encoded distributions (i.e., the function $f(x)$ in equation 1.1). We now consider special cases that have been investigated in other studies and show that our exact equation may be reduced to yield the results of previous work when identical assumptions are made.

For instance, Denève (2005) has recently obtained a set of equations for exact Bayesian inferences for binary Markov random variables using neurons encoding the log-likelihood ratio: $f(x) = \log(x) - \log(1 - x)$. To determine the multidimensional analog, we first assume that $f(x) = \log(x)$ and transform equation 3.6 to obtain

$$\frac{d\tilde{\phi}_i}{dt} = d_i(t) + \exp\left(-\tilde{\phi}_i\right)\sum_j w_{ij}\exp\left(\tilde{\phi}_j\right) - \sum_j \exp\left(\tilde{\phi}_j\right), \tag{4.1}$$

where $\tilde{\phi}_i = \log(\tilde{u}_i) = \log(\alpha) + \log(u_i)$. This equation can be further simplified by considering the complete set of likelihood ratios defined as $\phi_{ij} = \log(u_i) - \log(u_j)$. This yields the multidimensional equivalent of the binary case derived by Denève (2005):

$$\frac{d\phi_{ij}}{dt} = d_i - d_j + \sum_k w_{ik}\exp\left(\phi_{ki}\right) - w_{jk}\exp\left(\phi_{kj}\right). \tag{4.2}$$

A simple linear transformation allows an equation of this form to be mapped onto a biologically plausible network that functions by linearly summing inputs under the assumption that the action of a dendrite is to exponentiate the rate of the presynaptic neuron.

5 Linearized Equations

A related study (Rao, 2004) considered a similar log probability encoding ($f(x) = \log(x)$). However, since many models of neural activity assume that a linear dynamical system can be used to model neural activity (Dayan & Abbott, 2001), Rao generated a linear differential equation for Bayesian inference by numerically approximating the log of a sum as a sum of logs. Unfortunately, as equation 4.1 indicates, Bayes-Markov inference cannot always be reduced to a linear dynamical system under a log-encoding scheme, and while this approach worked well enough for several cases of interest, that study did not make clear the conditions under which the approximation is likely to hold.

Here, we address this issue by considering the assumptions necessary to linearize equations 2.13, 4.1, or 4.2. As shown in appendix C, an equivalent linearization follows naturally from the assumption that the input has a low amplitude, that is, $|\vec{d}(t)| \ll 1$. For equation 4.1, this implies that $|\tilde{\phi}_i - \tilde{\phi}_j - (\log(b_i) - \log(b_j))| \ll 1$ and, similar to Rao (2004), this leads to a linear dynamical system of the form

$$\frac{d\tilde{\phi}_i^1}{dt} = d_i(t) + \sum_j w_{ij}^* \left(\tilde{\phi}_j^1 - \log(b_j)\right) - \gamma \sum_j b_j \left(\tilde{\phi}_j^1 - \log(b_j)\right), \tag{5.1}$$

but with recurrent weights given by

$$w_{ij}^* = w_{ij} \frac{b_j}{b_i} \tag{5.2}$$

rather than by a numerical fit. Note also that the parameter $\gamma > 0$ does not affect the output probability distribution but is necessary to ensure the linear stability of solutions.

Alternatively, we note that the same assumption necessary to linearize the equation in the log domain can be applied to equation 2.13, leading to a compellingly simple equation for an approximate probability distribution,

$$\frac{du_i^1}{dt} = b_i d_i(t) + \sum_j w_{ij} u_j^1. \tag{5.3}$$

This implies that the approximation used in Rao is valid in the same regime as the simple linear model in the probability domain given by equation 5.3. Furthermore, when the driving function $d_i(t)$ is obtained by linearly summing input rates, this equation takes the form of a typical rate model

for neural activity (Dayan & Abbott, 2001),

$$\frac{d\tilde{u}^1}{dt} = \mathbf{BM}\vec{v}(t) + \mathbf{W}\tilde{u}^1 + \gamma\vec{b}(1 - \vec{\mathbf{1}} \cdot \tilde{u}^1), \tag{5.4}$$

where $\mathbf{B}$ is the diagonal matrix given by the bias vector, so that the matrix $\mathbf{BM}$ has the property that $\vec{\mathbf{1}} \cdot \mathbf{BM} = \vec{\mathbf{0}}^T$. Once again, $\gamma > 0$ is an adjustable parameter that does not affect the output probability distribution, but is necessary to ensure that the solution is asymptotically stable and, in the absence of inputs, equal to the bias vector. Moreover, equation 3.3 implies that when the stimulus is unbiased, maximum likelihood parameter estimation for this equation has an objective function given by

$$\langle \vec{d} \cdot \vec{u} \rangle = \langle \vec{u}^1 \cdot \mathbf{M}\vec{v} \rangle, \tag{5.5}$$

which we recognize as the quantity maximized by a constrained Hebbian learning rule associated with such a simple linear model of neural activity (Dayan & Abbott, 2001). Once again, note that as in the log-likelihood ratio case, linearizing eliminates the possibility of using the amplitude to encode model performance.

6 Discussion

We have presented a systematic technique for the generation of a system of equations that track the evolution of a dynamical neural network for the case of hidden (and hierarchical hidden) Markov models. This technique was shown to benefit from being exact, requiring no parameter fit, and retaining a simple interpretation of recurrent weights—that is, excitatory recurrent weights are transition probabilities. Moreover, we have demonstrated that since the amplitude of a population that directly encodes a probability distribution is arbitrary, we can use it to incorporate an online estimate of model performance. For the purposes of comparison with previous work, we then generated equations for the evolution of the log likelihood and log-likelihood ratios, as well as linear approximations to the dynamical systems that calculate the likelihood and log-likelihood functions.

We now consider the biologically plausibility of Bayes inferences in the likelihood and log-likelihood domains. One advantage of the likelihood implementation we have proposed is that exact Bayesian inference for an HMM requires only divisive normalization and dendritic multiplication, both of which have been observed in neurons (Heeger, 1992; Pena & Konishi, 2001; Gabbiani, Krapp, Koch, & Laurent, 2002). In contrast, much previous work has argued for the implementation of Bayesian inferences in the log probability domain, typically by arguing that such implementations do not require that neurons perform any complicated actions such as

multiplication. Unfortunately, this is not the case. While it is true that the log transformation eliminates the multiplicative terms associated with the feedforward terms, equation 4.1 clearly demonstrates that for tasks more complicated than binary decision making, this transformation can result in a dynamical system that has an even more complicated nonlinear interaction associated with the recurrent terms. Rao (2004), resolved this issue through the generation of a linear dynamical system that approximates the full nonlinear system. However, the derivation of equation 5.1 demonstrates that the assumption necessary to make this type of approximation valid in the log-likelihood domain could just as easily have been applied in the likelihood domain to obtain an equally plausible linear dynamical system. Additionally, an implementation of Bayes inferences for an HMM capable of incorporating prior knowledge on short timescales would require a network capable of dynamically adjusting the recurrent connection strengths. In both the likelihood and log-likelihood domains, the recurrent connection strengths are multiplied by the synaptic input, so that if priors are to be included, then even the linear equations may ultimately require neurons capable of performing a multiplicative operation. Thus, it seems that there is both little distinction between and little utility of linearization in the likelihood and log-likelihood domains.

This may be why the log of the sum approximation has been dropped in a more recent work (Rao, 2005), which identifies the attracting state of the membrane potential of a neuron with the log of the equation for Bayesian inference (see equation 2.3). The firing rate of such a neuron is then obtained by exponentiating this quantity so that the firing rate is approximately proportional to the likelihood function. The log of sums is implemented by a dendritic filtering function that applies a log transform to a linear combination of the input rates. In the work presented here, we are agnostic as to the interpretation and evolution of membrane potential and simply remark that, if neuronal dynamics can implement the model described in Rao (2005) for inputs generated by an HMM (as opposed to the static stimulus which is analyzed in that work), then the resulting rate dynamics should approximate the exact evolution equation for the likelihood function, equation 2.13, derived above. Moreover, we note that the utility of this work lies in the demonstration that the consideration of a dynamic as opposed to static stimulus actually simplifies network dynamics required to implement Bayesian inference.

Finally, we argue that simply representing the likelihood function is inefficient and suggest that the ability of equation 3.6 to represent model performance through mean activity of the network can be used to explain one of the experimental discrepancies of the Bayesian information integration model used by Shadlen and Newsome (2001) to model the behavior of LIP neurons. In that work, a weakly tuned binary stimulus (left or right motion) was shown, and neurons in LIP were observed to initially exhibit a linear increase or decrease in activity depending on their direction selectivity.

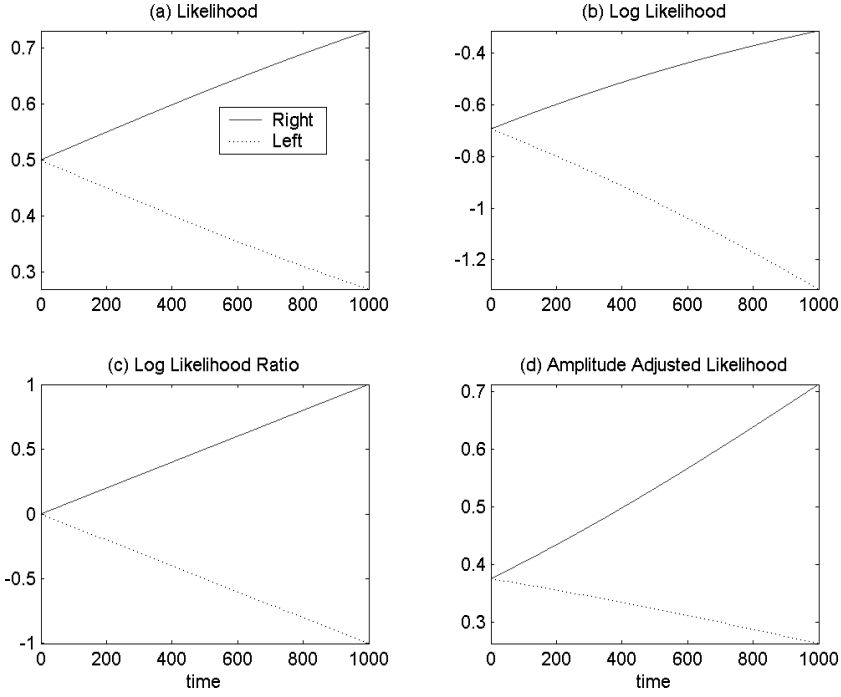

Figure 2: Predictions for the four explicit representations of the probability distribution for a binary decision task. Solid lines represent right selective neurons, while dashed lines represent left selective neurons. In the likelihood and log-likelihood ratio cases, left selective neurons decrease in activity at the same rate as right selective neurons. In the log-likelihood case, the decrease is larger than the increase. Only when the amplitude is used to encode model performance does the increase in activity of the right selective neuron exceed the decrease in the left selective neurons, as is observed in Shadlen and Newsome (2001).

However, the linear decrease was always observed to be of lesser absolute slope than the linear increase, indicating that mean activity increases. As noted in that work, this is incompatible with the interpretation of the activity of left and right selective neurons as representative of log-likelihood ratios. In particular, if these LIP neurons are tracking log-likelihood ratios or potential functions or any linearized quantity that is a function of the probability, then the mean activity should not change (see Figure 2c). This is also the case when these LIP neurons are assumed to represent the likelihood function directly (see Figure 2a). Similarly, when the mean activity represents the log likelihood, the concavity of the log function implies that the mean activity should actually decrease (see Figure 2b). In contrast, if the amplitude of the population (or in this case the average activity of the

pair of neurons) also encodes the likelihood of the recent data given the model parameters, then we would expect an increase in mean activity to accompany the presentation of a familiar stimulus (see Figure 2d). To our knowledge, this is the only model that explicitly represents the probability distribution and accounts for this feature of the relevant experimental data in LIP.

Thus, we conclude that the implementation of exact Bayes-Markov inferences for biologically plausible stimulus-dependent patterns of activity in the likelihood domain may be amplitude modulated to provide an online representation of model performance in a manner consistent with both biological constraints and physiological measurements.

Appendix A: Conditional Preference

The only departure from a standard hidden Markov that was required to yield an exact implementation of Bayesian inference through a first-order differential equation was the requirement that the conditional preference have a Taylor expansion in Δt, given by

$$\frac{\Pr\left(\vec{v}^n|\theta_i^n\right)}{\Pr\left(\vec{v}^n\right)} = \frac{\Pr\left(\theta_i^n|\vec{v}^n\right)}{\Pr\left(\theta_i^n\right)} = 1 + \Delta t d_i^n + O\left(\Delta t^2\right). \tag{A.1}$$

By considering a series of discrete-time hidden Markov models that converge to a continuous-time HMM, it is possible to show that a Taylor expansion in Δt necessarily exists, but that it is not generally the case that the first term of the expansion is one. This additional restriction holds only when a single snapshot of the stimulus-dependent pattern of activity is noninformative, that is, when $\Pr(\theta_i^n|\vec{v}, \Delta t = 0) = \Pr(\theta_i^n)$. Fortunately, when the stimulus-dependent pattern of activity is composed of independent asynchronous events, such an event occurs with zero probability in a time window of zero width. This implies that zero-width time windows are noninformative. As an example, consider a stimulus-dependent pattern of activity that is a collection of independent Poisson processes with stimulus-dependent (and nonzero) rates,

$$\Pr(\vec{v}^n|\theta, \Delta t) = \prod_i \frac{(\Delta t f_i(\theta))^{\Delta t v_i^n} \exp\left(-\Delta t f_i(\theta)\right)}{\left(\Delta t v_i^n\right)!}, \tag{A.2}$$

where v_i^n is now thought of as a discrete estimate of the rate of neuron i obtained by summing Poisson spikes in a time window of length Δt and $f_i(\theta)$ is the tuning curve of that neuron. For such a distribution

equations 2.7 and A.1 take the form

$$\frac{\Pr(\vec{v}^n|\theta, \Delta t)}{\Pr(\vec{v}^n|\Delta t)} = \prod_i \frac{(\Delta t f_i(\theta))^{\Delta t v_i^n} \exp(-\Delta t f_i(\theta))}{(\Delta t \bar{f}_i)^{\Delta t v_i^n} \exp(-\Delta t \bar{f}_i)}. \tag{A.3}$$

The log of this equation yields

$$\log\left(\frac{\Pr(\vec{v}^n|\theta, \Delta t)}{\Pr(\vec{v}^n|\Delta t)}\right) = \Delta t \left(\sum_i v_i^n \log\left(\frac{f_i(\theta)}{\bar{f}_i}\right) - f_i(\theta) + \bar{f}_i\right), \tag{A.4}$$

which is proportional to Δt since v_i^n is an estimate of rate (as opposed to spike count) and may be considered an order one quantity. Exponentiation of this equation then leads to the desired expansion in Δt provided that $f_i(\theta)$ is nonzero for all values of the stimulus.

Appendix B: Estimating Model Performance

To generate an expression for the evolution of the probability of the input pattern of activity given the model parameters M, we recall that

$$L^n = \Pr(\vec{v}^n, \vec{v}^{n-1}, \ldots, \vec{v}^0|\mathrm{M}) = \Pr(\vec{v}^n|\vec{v}^{n-1}, \ldots, \vec{v}^0, \mathrm{M}) L^{n-1}. \tag{B.1}$$

Using equation 2.9 to evaluate the conditional dependence of the current activity pattern on the history of the activity pattern, we see that

$$L^n = L^{n-1} \Pr(\vec{v}^n|\mathrm{M})(1 + \Delta t \vec{d}^n \cdot \vec{u}^{n-1}), \tag{B.2}$$

where we have used the fact that $\mathbf{W}^T \vec{n} = \vec{0}$. Taking the natural log, expanding to order Δt yields

$$\log(L^n) = \log(L^{n-1}) + \log(\Pr(\vec{v}^n|\mathrm{M})) + \Delta t \vec{d}^n \cdot \vec{u}^{n-1} \tag{B.3}$$

or, equivalently,

$$\log(L^N) = \log(L^0) + \sum_{j=1}^{N} \log(\Pr(\vec{v}^j|\mathrm{M})) + \sum_{j=1}^{N} \Delta t \vec{d}^j \cdot \vec{u}^{j-1}. \tag{B.4}$$

Written in this way, we note that the first sum in equation B.4 represents the probability of the stimulus-dependent activity under the assumption that each interval is independent. However, for a hidden Markov model, each stimulus-dependent pattern of activity is independent only when conditioned on the stimulus; therefore, this term represents the log likelihood

of the stimulus-dependent pattern of activity under the assumption that the pattern is independent of the time course of the stimulus. This implies that if the model parameters accurately characterize the average behavior of the stimulus-dependent pattern of activity, but not necessarily the actual stimulus dependence, then this term may be neglected, giving rise to the expression in equation 2.17.

Once again a consideration of the independent Poisson case is helpful. In this case, equation A.3 implies that the first sum in equation B.4 has terms of the form

$$\log\left(\Pr(\vec{v}^n|\mathrm{M},\Delta t)\right) = \Delta t \sum_i v_i^n \log\left(\bar{f}_i\right) - \bar{f}_i + \sum_i \Delta t v_i^n \log\left(\Delta t_i\right) - \sum_{j=1}^{\Delta t v_i^n} \log(j) \tag{B.5}$$

However, only terms that contain $\bar{f}_i$ are affected by model parameters so that a maximum likelihood estimate may ignore the second line of equation B.5. Thus maximizing L is equivalent to maximizing L^* if $L^*(T)$ is given by

$$L^*(T) = L^*(0) + \int_0^T \vec{d}(s)\cdot\vec{u}(s)ds + \int_0^T \left(v_i(s)\log\left(\bar{f}_i\right) - \bar{f}\right)_i ds. \tag{B.6}$$

Finally, if the model parameters do not affect the stimulus-independent mean rates of the input neurons $\left(\bar{f}_i\right)$, then we may conclude that

$$\frac{dL^*}{dt} = \vec{d}\cdot\vec{u}, \tag{B.7}$$

as desired.

Appendix C: Linearizations

To generate a linear dynamical system that approximates a Bayes-Markov network we could linearize any of the exact equations 2.13, 3.6, 4.1, or 4.2. To demonstrate that such linearizations follow naturally from the assumption of weakly tuned input, we now assume that $\mathbf{D}(t) = \varepsilon\mathbf{D}^1(t)$ and $u_i(t) = b_i + \varepsilon u_i^1(t) + O(\varepsilon^2)$. Inserting this ansatz into equation 2.14 yields

$$\frac{du_i^1}{dt} = b_i d_i^1(t) + \sum_j w_{ij}u_j^1 + O(\varepsilon). \tag{C.1}$$

Equation 5.4 is obtained by noting that the recurrent weight matrix is singular with the left and right eigenvectors given by $\vec{\mathbf{1}}$ and $\vec{b}$, respectively. Since the dynamical system is linear, an addition of a recurrent term proportional to $(\vec{\mathbf{1}} \cdot \vec{u})\vec{b}$ and a driving term proportional to $\vec{b}$ will have the effect of adding a vector proportional to $\vec{b}$ to the solution, while enhancing the linear stability of the solution.

For the purposes of comparison with Rao (2004), we also choose to linearize equation 4.1. First, we note the uniform inhibitory term in equation 4.1 can be replaced by an arbitrary function without affecting the resulting probability distribution since

$$u_i(t) = \frac{\exp(\phi_i(t) + \chi(t))}{\sum_j \exp(\phi_i(t) + \chi(t))} = \frac{\exp(\phi_i(t))}{\sum_j \exp(\phi_i(t))} \tag{C.2}$$

for any $\chi(t)$. Thus, a linear dynamical system that approximates a Bayes-Markov network is concerned only with the accuracy of the linear approximation to the nonlinearity, which contributes the excitatory term, that is, the term given by

$$\sum_j w_{ij} \exp\left(\tilde{\phi}_j - \tilde{\phi}_i\right). \tag{C.3}$$

When $\mathbf{D}$ is of order ε,

$$\tilde{\phi}_j - \tilde{\phi}_i = \phi_j - \phi_i \approx \log(b_j) - \log(b_i) + \varepsilon \frac{u_j^1}{b_j} - \varepsilon \frac{u_i^1}{b_i}, \tag{C.4}$$

which implies that

$$\sum_j w_{ij} \exp\left(\tilde{\phi}_j - \tilde{\phi}_i\right) \approx \sum_j w_{ij} \frac{b_j}{b_i} \left(1 + \varepsilon \frac{u_j^1}{b_j} - \varepsilon \frac{u_i^1}{b_i}\right) = \varepsilon \sum_j w_{ij} \frac{u_j^1}{b_i}, \tag{C.5}$$

where we have used the fact that

$$\sum_j w_{ij} b_j = 0. \tag{C.6}$$

Defining $\tilde{\phi}_j^1$ so that

$$\varepsilon \frac{u_j^1}{b_j} = \tilde{\phi}_j^1 - \log(b_j) - \log(\alpha(t)), \tag{C.7}$$

we may combine equation 5.1 with the approximation given by equation C.5 to obtain the linearized equation

$$\frac{d\tilde{\phi}_i^1}{dt} = d_i(t) + \beta + \sum_j w_{ij}^* \tilde{\phi}_j^1 - \sum_j w_{ij}^* \log(b_j), \tag{C.8}$$

with

$$w_{ij}^* = w_{ij} \frac{b_j}{b_i}. \tag{C.9}$$

Recalling that any uniform term may added to an equation that tracks the log likelihood, we obtain equation 5.1 by simply adding an arbitrary uniform term given from a sum of the $\tilde{\phi}_j$'s.

Acknowledgments

J.B. was supported by NIH training grant T32-MH19942, and A.P. was supported by NSF grant 0346785.

References

Anderson, C. (1994). Neurobiological computational systems. In J. M. Zurada, R. J. Marks, & C. J. Robinson (Eds.), *Computational intelligence imitating life*. New York: IEEE Press.

Carpenter, R. H., & Williams, M. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature, 377*(6544), 59–62.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Denève, S. (2005). Bayesian inferences in spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 353–360). Cambridge, MA: MIT Press.

Gabbiani, F., Krapp, H. G., Koch, C., & Laurent, G. (2002). Multiplicative computation in a visual neuron sensitive to looming. *Nature, 420*(6913), 320–324.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*, 10–16.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*, 181–197.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci., 27*(12), 712–719.

Koechlin, E., Anton, J. L., & Burnod, Y. (1999). Bayesian inference in populations of cortical neurons: A model of motion integration and segmentation in area MT. *Biol. Cybern., 80*(1), 25–44.

Nelson, M. E. (1994). A mechanism for neuronal gain control by descending pathways. *Neural Computation, 6*, 242–254.

Pena, J. L., & Konishi, M. (2001). Auditory spatial receptive fields created by multiplication. *Science, 292*(5515), 249–252.

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits. *Neural Comput., 16*(1), 1–38.

Rao, R. P. (2005). Bayesian inference and attention modulation in the visual cortex. *Neuroreport, 16*(16), 1843–1848.

Sahani, M., & Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Comput., 15*(10), 2255–2279.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol., 86*(4), 1916–1936.

Yu, A. J., & Dayan, P. (2005). Inference, attention, and decision in a Bayesian neural architecture. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Received June 10, 2005; accepted August 18, 2006.