Emergent Inference of Hidden Markov Models in Spiking Neural Networks through Winner-Take-All

Zhaofei Yu, Shangqi Guo, Fei Deng, Qi Yan, KeKe Huang, Jian K. Liu, Feng Chen, *Member, IEEE*

***Abstract*—Hidden Markov Models (HMMs) underpin the solution to many problems in computational neuroscience. However, it is still unclear how to implement inference of HMMs with a network of neurons in the brain. The existing methods suffer from the problem of being non-spiking and inaccurate. Here we build a precise equivalence between the inference equation of HMMs with time-invariant hidden variables and the dynamics of spiking winner-take-all (WTA) neural networks. We show that the membrane potential of each spiking neuron in the WTA circuit encodes the logarithm of the posterior probability of the hidden variable in each state, and the firing rate of each neuron is proportional to the posterior probability of the HMMs. We prove that the time course of neural firing rate can implement posterior inference of HMMs. Theoretical analysis and experimental results show that the proposed WTA circuit can get accurate inference results of HMMs.**

***Index Terms*—Hidden Markov models, Winner-take-all circuits, spiking neural network, posterior inference, neural implementation**

I. Introduction

HIDDEN Markov models (HMMs) are a kind of dynamic probabilistic graphical model [1], which have been widely used in computational neuroscience [2], [3], [4], [5], [6], computational biology [7], [8], statistical physics [9], [10] and machine learning [11], [12], [13], [14]. In computational neuroscience, HMMs are used to detect hidden regularities with sequential sensory inputs. In particular, they have been proved extremely useful in modeling inference and decision making in the cognitive process of the human brain when the state of the hidden variable is time–invariant [15]. Despite its accurate and powerful computing performance with experimental data, it remains an open question how a network of spiking neurons can implement probabilistic inference of HMMs at the neural circuit level. This problem is of great importance to brain science and artificial intelligence. On the one hand, it can build the bridge between the process of inference and decision making of the human brain at high level and the dynamics of spiking neural networks at low level. One the other hand, one can build the machine that is able to perform inference and make decision like the human brain with these mechanisms.

Various schemes of neural networks have been proposed over the last 15 years to tackle the problem above. Rao [3] built the relationship between the dynamic equation of recurrent neural networks and the inference equation of HMMs, and suggested that the dynamic process of a recurrent neural network is a process of probabilistic inference. However, since the two equations are not exactly equivalent, a sum-of-logs is used to approximate a log-sum, which leads to inaccurate inference results. Beck and Pouget [16] took a further step and built a precise relationship between the inference equation of HMMs and the dynamic equation of a first-order quadratic nonlinear recurrent network. Both methods only focus on non-spiking neural networks while the spike in neuron is the key for computation [17], [18], [19]. Recently some researchers have considered biophysically plausible spike-based networks. For example, Deneve [20] demonstrated that each leaky integrate-and-fire neuron can compute the probability of one hidden variable of HMMs, but it was limited to a binary variable. In summary, most of the previous studies suffer from the problem of being non-spiking, and the existing spiking neural network cannot obtain accurate solution.

In this paper, by focusing on HMMs with time-invariant hidden variables, we found that there is a precise equivalence between their inference equations and the dynamical equations of spiking neurons when the underlying circuit is organized by a winner-take-all (WTA) fashion. Typically, there are two coupled operations during each update of the inference process of HMMs, namely evidence accumulation and normalization, in which the result of normalization in each step serves as the past evidence for the next step. However, we found a WTA spiking neural network with self-connections can naturally decouple these two operations while keeping the precise inference of HMMs. We proved that the inference result of the HMM remains unchanged if the normalization of posterior probabilities is carried out only at the last step instead of each step during evidence accumulation. Based

This work is supported in part by National Postdoctoral Program for Innovative Talents under Grant BX20180005, and in part by China Postdoctoral Science Foundation under Grant and 2018M630036 and 2017M620525, and in part by the National Natural Science Foundation of China under Grant 61671266, 61703439, 61327902, and in part by the Human Brain Project of the European Union #604102 and #720270.

Z. Yu is with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, China, and also with the Department of Automation, Center for Brain-Inspired Computing Research, Tsinghua University, Beijing 100084, China. e-mail: (yzf714@126.com).

S. Guo, F. Deng, Q. Yan and F. Chen are with the Department of Automation, Center for Brain-Inspired Computing Research, Tsinghua University, Beijing 100084, China, and also with LSBDPA Beijing Key Laboratory, Beijing, China. e-mail: (gsq15@mails.tsinghua.edu.cn, dengf15@mails.tsinghua.edu.cn, qyan15@mails.tsinghua.edu.cn, chenfeng@tsinghua.edu.cn).

K. Huang is the School of Information Science and Engineering, Central South University, Changsha 410083, China (e-mail: huangkeke@csu.edu.cn).

J. K. Liu is with Center of Systems Neuroscience, Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, UK, and Institute for Theoretical Computer Science, Graz University of Technology, Graz, Austria. e-mail: (jian.liu@leicester.ac.uk).

Zhaofei Yu and Shangqi Guo contributed equally to this work. Corresponding authors: Jian K. Liu; Feng Chen.

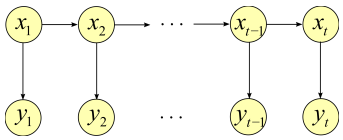


Fig. 1. The scheme of a general hidden Markov model. x_i are hidden variables and y_i are observation variables ($i = 1, 2, ..., t$). Here we consider a special case of HMM with time-invariant hidden variables, i.e., $x_1 = x_2 = ... = x_t$.

on this theory, we can decompose the corresponding neural circuits into two parts: one for updating the posterior with new evidence, and the other for computing the normalization of the distribution.

Furthermore, we showed that the membrane potential of each spiking neuron in the WTA circuit encodes the logarithm of posterior probability of the hidden variable in each state, and the neural firing rate is proportional to the posterior probability of HMM. In addition, we proved that the time course of neural firing rate can implement posterior inference of HMMs. Experimental results with simulation demonstrate that the proposed WTA network can get accurate inference results of HMMs.

The rest of the paper is organized as follows. Section II derives the inference equation of HMMs. In section III we introduce WTA neural network with self-connections and show how it can implement inference of HMMs. We show the experimental results in section IV and conclude in section V.

II. Inference of Hidden Markov models

HMMs are a kind of directed graphical model [21], [22] composed of a hidden variable sequence $X = \{x_1, x_2, ..., x_t\}$ and an observation variable sequence $Y = \{y_1, y_2, ..., y_t\}$ [21] (shown in Fig. 1). The hidden variable sequence X is a first-order Markov chain, and each observation variable y_t is only governed by the corresponding hidden variable x_t through conditional probability $p(y_t|x_t)$. Thus, the joint distribution of a HMM in Fig. 1 can be written as a product of conditional distributions:

$$\begin{aligned} &p(x_1, x_2, ..., x_t, y_1, y_2, ..., y_t) \\ &= p(x_1) \left[\prod_{n=2}^{t} p(x_n|x_{n-1})\right] \prod_{n=1}^{t} p(y_n|x_n). \end{aligned} \tag{1}$$

In this paper, we consider the HMMs with time-invariant hidden variables, that is, $x_1 = x_2 = ... = x_t$. This means the values for the hidden variables will be the same no matter what time they are observed. This model is important to many inference and decision making problems [23], [24], [25] since in many cases we have the prior knowledge where the state of the environment doesn't change or changes very slowly with respect to time [15].

The inference problem is to infer the most probable state of the hidden variable at time t with the observations from 1 to t, that is,

$$\begin{aligned} &\arg\max_{x_t} p(x_t \mid y_1, y_2, ..., y_t) \\ &= \arg\max_{x_t} \sum_{x_1, x_2, ..., x_{t-1}} p(x_1, x_2, ..., x_t \mid y_1, y_2, ..., y_t). \end{aligned} \tag{2}$$

Equation (2) can be calculated by computing the posterior distribution $p(x_t \mid y_1, y_2, ..., y_t)$ and then choosing the state of x_t with maximum probability. In order to implement inference of HMMs with spiking neural networks, a direct idea is to rewrite equation (2) into a dynamic equation and then build the relationship between this equation and the dynamic equation of spiking neural networks. In fact, we can use a difference equation to implement inference of equation (2), and we have the following theorem:

Theorem 1: Supposing that $F(x_0 = x^i) = \ln p(x_1 = x^i)$[1], and

$$F(x_t = x^i) = \ln\ p(y_t|x_t = x^i)\ +\ F(x_{t-1} = x^i) - \ln\ Z_t \tag{3}$$

holds for $t \geq 1$, with $F(x_t)$ denoting a function of x_t, $x_t = x^i$ denoting that random variable x_t is in state x^i, and Z_t being the normalizing constant of $\exp(F(x_t))$ to keep $\sum_i \exp(F(x_t = x^i)) = 1$, that is, $Z_t = \sum_i p(y_t|x_t = x^i) \exp(F(x_{t-1} = x^i))$, then we conclude that for $t \geq 1$:

$$F(x_t) = \ln p(x_t|y_1, y_2, ..., y_t), \tag{4}$$

and

$$\arg\max_{x_t} e^{F(x_t)} = \arg\max_{x_t} p(x_t \mid y_1, y_2, ..., y_t). \tag{5}$$

The proof of Theorem 1 (with all other theorems) is provided in Supplemental Materials. Theorem 1 shows that we can use a difference equation to compute the posterior distribution $\ln p(x_t|y_1, y_2, ..., y_t)$. To be specific, supposing that $F(x_0 = x^i) = \ln p(x_1 = x^i)$, it follows from equation (3) that $\ln p(x_1|y_1)$ can be computed as $F(x_1)$. Likewise, then we can compute $\ln p(x_2|y_1, y_2)$ with equation (3), and so on. Note that there exist two operations in equation (3): evidence accumulation as

$$F(x_t = x^i) = \ln\ p(y_t|x_t = x^i) + F(x_{t-1} = x^i), \tag{6}$$

and normalization as

$$F(x_t = x^i) = F(x_t = x^i) - \ln Z_t. \tag{7}$$

The result of normalization in each step serves as the past evidence for the next step. This means the computations of evidence accumulation and normalization are coupled to each other. This coupling is a challenge for graphical models as well as computational neuroscience, since it is difficult to design a spiking neural circuit to implement accurate inference of HMMs with the coupled equations.

Here we show in the following theorem that the operations of evidence accumulation and normalization can be decoupled. The distribution $e^{F(x_t)}$, namely, $p(x_t|y_1, y_2, ..., y_t)$ is left unchanged if the operation of normalization is carried out at the last step instead of each step.

Theorem 2. Supposing that $G(x_0 = x^i) = F(x_0 = x^i) = \ln p(x_1 = x^i)$ and

$$G(x_t) = \ln p(y_t|x_t) + G(x_{t-1}) \tag{8}$$

holds for $t \geq 1$, then we conclude that the normalization of $e^{G(x_t)}$ equals the distribution $p(x_t|y_1, y_2, ..., y_t)$, that is,

$$e^{G(x_t)} \propto p(x_t|y_1, y_2, ..., y_t), (t \geq 1). \tag{9}$$

[1]Note that the hidden variable sequence of the HMM is $X = \{x_1, x_2, ..., x_t\}$ while the variable sequence of the function $F(.)$ is $x_0, x_1, x_2, ..., x_t$. Here x_0 can be seen as an auxiliary variable.

Combing equation (9) and equation (4), one can find that the normalization of $e^{G(x_t)}$ is the same as the normalization of $e^{F(x_t)}$. Now we can conclude that even if the operation of normalization is carried out at the last step instead of each step, the distribution of $e^{F(x_t)}$ is unchanged. Thus, we can use the difference equation (8) and the initial condition $G(x_0 = x^i) = \ln p(x_1 = x^i)$ to implement posterior inference, and we have $\frac{1}{Z_t} e^{G(x_t)} = p(x_t \mid y_1, y_2, ..., y_t)$, where the normalization constant becomes $Z_t = \sum_{x_t} e^{G(x_t)}$. The benefit of this theorem is that we can decouple the operation of evidence accumulation and normalization, specifically, the result of normalization in current step doesn't need to be the input of the next step. Thus, when we design the corresponding spiking neural network to implement inference of HMMs, we can separate the neural network into two parts: one for updating the posterior with new evidence, that is, $G(x_t) = \ln p(y_t|x_t) + G(x_{t-1})$, and the other for computing the normalization of the posterior distribution $e^{G(x_t)}$. The problem now is whether there exists a plausible spiking neural circuit that can implement these computations of HMMs.

III. Emergent inference in spiking neural network through Winner-take-all

In this section, we show that a spiking neural network of WTA circuit with self-connections can naturally implement inference of HMMs. The membrane potential of spiking neurons in WTA circuits with self-connections can accumulate evidence, namely, update the posterior with new evidence. The competitive mechanism of WTA circuits can normalize the firing rate of each neuron.

We first introduce spiking neural networks and WTA circuits, and then derive the dynamic equations of a spiking neural network of WTA circuit with self-connections. At last we demonstrate that inference of HMMs can be easily implemented by this spiking neural network.

A. Spiking neural network

Spiking neural networks are thought as the third generation of artificial neural network models, which is closer to biological neurons in the brain [26]. In a spiking neural network, each neuron can receive current from other neurons and the membrane potential of which will change. When the membrane potential exceeds a threshold value, an output signal, which is called a spike (or an action potential), will be generated and be delivered to other neurons. Together with neuronal and synaptic state, spike timing is also considered in spiking neural networks model.

Here we consider a network of K spiking neurons $z_1, \ldots, z_K$ and denote the output spike train of neuron z_k by $z_k(t)$ defined as a sum of Dirac delta pulses positioned at the spike times $t_k^{(1)}, t_k^{(2)}, \ldots$, i.e., $z_k(t) = \sum_f \delta(t - t_k^{(f)})$, where $f = 1, 2, ...$ It means $z_k(t) = 1$ if neuron z_k spikes at time $t = t_k^{(f)}$ and $z_k(t) = 0$ otherwise. Neurons $z_1, \ldots, z_K$ are modeled by a standard stochastic variant of the spike response model [27], which is a generalization of leaky integrate-and-fire neuron. In this model, the membrane potential of a neuron z_k at time t is given by:

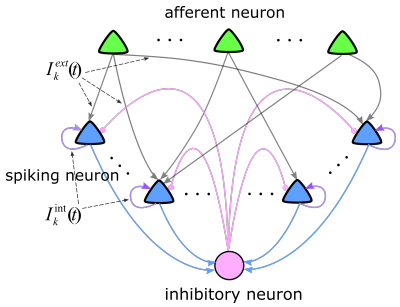


Fig. 2. A scheme of a WTA spiking neuronal circuit with self-connections. There are K output excitatory neurons (blue) and one inhibitory neuron (pick). The temporal sequences of observation variables of HMM are fed by efferent neurons (green). In the end, the hidden state variable of HMM will be represented by one of output neurons due to the competition mechanism of WTA when evidence is cumulated over time.

$$u_k(t) = \sum_f \eta(t - t_k^{(f)}) + \int_0^\infty \kappa(s) I_k(t-s) ds + u_k^{rest}, \quad (10)$$

where $I_k(t)$ denotes the time-dependent current of neuron k at time t, and u_k^{rest} is the rest potential of neuron z_k. $\eta(t - t_k^{(f)})$ is the kernel that describes the reset of the membrane potential of neuron z_k after the spike at $t_k^{(f)}$. $\kappa(s)$ represents the voltage response to a short current pulse. In this paper, we use standard exponential kernels $\eta(t - t_k^{(f)})$ and $\kappa(s)$:

$$\eta(t - t_k^{(f)}) = -\eta_0 \exp\left(-\frac{t - t_k^{(f)}}{\tau}\right), \quad (11)$$

$$\kappa(s) = \varepsilon_0 \exp\left(-\frac{s}{\tau}\right), \quad (12)$$

with the reset potential $\eta_0 = 5$ mV, the membrane time constant $\tau = 20$ ms, and the voltage response amplitude as $\varepsilon_0 = 5$ mV. The parameters set here are similar to that of [28]. Here we consider the escape noise model of spiking neurons, which replaces the strict firing threshold by a noisy threshold [27]. This means that a neuron can fire stochastically. To be specific, the instantaneous firing rate (firing intensity) of neuron z_k is supposed to be stochastic, which is often modeled by an exponential function [29]:

$$\rho_k(t) = \rho \exp(u_k(t) - \theta), \quad (13)$$

with θ representing the firing threshold and ρ scales the firing rate of the neuron. It has been shown by experiments that this model is in good agreement with real neurons [30]. One can find that the instantaneous firing rate (firing intensity) increases as the distance between membrane potential and firing threshold decreases.

B. Winner-take-all circuit

WTA circuit has been suggested as a ubiquitous motif of cortical microcircuits [31], which is widely used to implement normalization [32], visual attention [33] and classification [34]. We consider a WTA circuit of K output spiking neurons

(blue triangles) $z_1, \ldots, z_K$ and an inhibitory neuron (pink circle) as in Fig. 2. The output spiking neurons $z_1, \ldots, z_K$ mutually inhibit each other through the inhibitory neuron. Thus, all the neurons in the output layer are in competition against each other so that they cannot fire simultaneously.

In this paper, we consider the WTA model used in [2], [35], where all neurons are allowed to fire with non-zero probability. Considering all the neurons in WTA circuit are subject to the same lateral inhibition [2] [2], the instantaneous firing rate (firing intensity) of neuron z_k in WTA circuit at time t is determined by [2]:

$$\rho_k(t) = \frac{\rho}{Q(t)} \exp(u_k(t) - \theta), \tag{14}$$

where ρ scales the firing rate of neurons. $Q(t)$ represents the lateral inhibition between the neurons in the WTA circuit, which is defined as:

$$Q(t) = \sum_k \exp(u_k(t) - \theta). \tag{15}$$

Substituting equation (15) into (14) obtains:

$$\begin{aligned} \rho_k(t) &= \frac{\rho}{\sum_k \exp(u_k(t) - \theta)} \exp(u_k(t) - \theta) \\ &= \rho \, \frac{\exp(u_k(t))}{\sum_k \exp(u_k(t))}. \end{aligned} \tag{16}$$

This WTA circuit works like a soft-max function. At each time, all neurons can fire with non-zero probability, but the neuron with the highest membrane potential has the highest firing probability.

C. Implement inference with spiking neural network

In this subsection we demonstrate that the dynamics of spiking neural network of WTA circuits with self-connections can naturally implement inference of HMMs. We show how the spiking neurons in WTA circuits with self-connections can update the posterior probabilities with new evidence in Theorem 3 and how the competitive mechanism of WTA circuit can normalize the posterior distribution in Theorem 4.

In the WTA circuit with self-connections in Fig. 2, the time-dependent current to the network includes two parts: external afferent current $I_k^{ext}(t)$ to the network and internal current $I_k^{int}(t) = \sum_f \delta(t - t_k^{(f)})$ from itself by the self-connections. Thus, equation (10) can be rewritten as:

$$\begin{aligned} u_k(t) &= u_k^{rest} + \sum_f \eta(t - t_k^{(f)}) \\ &\quad + \int_0^\infty \kappa(s) \left(I_k^{ext}(t-s) + I_k^{int}(t-s) \right) ds. \\ &= u_k^{rest} + \sum_f \eta(t - t_k^{(f)}) \\ &\quad + \int_0^\infty \kappa(s) \left(I_k^{ext}(t-s) + \sum_f \delta(t - s - t_k^{(f)}) \right) ds. \end{aligned} \tag{17}$$

[2]lateral inhibition is the capacity of an excited neuron to reduce the activity of its neighborhood neurons.

Then, equation (17) can be reduced to:

$$\begin{aligned} u_k(t) &= u_k^{rest} + \sum_f \eta(t - t_k^{(f)}) \\ &\quad + \int_0^\infty \kappa(s) \left(I_k^{ext}(t-s) + \sum_f \delta(t - s - t_k^{(f)}) \right) ds \\ &= u_k^{rest} + \int_0^\infty \eta(s) \sum_f \delta(t - s - t_k^{(f)}) ds \\ &\quad + \int_0^\infty \kappa(s) \left(I_k^{ext}(t-s) + \sum_f \delta(t - s - t_k^{(f)}) \right) ds \\ &= \int_0^\infty \kappa(s) I_k^{ext}(t-s) ds + u_k^{rest} \\ &= \int_{-\infty}^t \kappa(t-s) I_k^{ext}(s) ds + u_k^{rest} \end{aligned} \tag{18}$$

The second equality holds as $\sum_f \eta(t - t_k^f) = \int_0^\infty \eta(s) \sum_f \delta(t - s - t_k^{(f)}) ds$. The third equality holds due to the definition $\eta(s) = -\kappa(s)$ in Section III-A. Note that an ideal model of spiking neurons is assumed here where the internal currents from self-connections do not lag behind the spike response. We show in the following theorem that the membrane potential of spiking neurons in WTA circuit with self-connections can accumulate afferent current. In other words, if the afferent current encode some variable, then the membrane potential of spiking neurons can compute the sums of a sequence.

Theorem 3. Considering the spiking neural network shown in Fig. 2, the rest potential $u_k^{rest} \leq 0$, and the external current $I_k^{ext}(t) = \sum_j \frac{I_k^j}{\varepsilon_0 \tau} \Theta(t - T_j)$ $(j = 1, 2, 3...)$, where I_k^j represents the jth constant current to neuron z_k and $I_k^j \leq 0$, T_j represents the arriving time of jth current I_k^j, $\Theta(\cdot)$ denotes the Heaviside step function, i.e. $\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise. The voltage response amplitude ε_0 and membrane time constant τ are defined similar to equation (12). Then for arbitrary $m \geq 1$, if $T_{m+1} - T_m \geq 3\tau$ holds, we conclude that:

$$\left| \frac{u_k(T_{m+1}) - \left(u_k^{rest} + \sum_{j=1}^m I_k^j \right)}{u_k^{rest} + \sum_{j=1}^m I_k^j} \right| < 0.05 \tag{19}$$

holds for all k, and:

$$\lim_{T_{m+1} - T_m \to +\infty} u_k(T_{m+1}) = u_k^{rest} + \sum_{j=1}^m I_k^j. \tag{20}$$

Theorem 3 shows that if the time interval $T_{m+1} - T_m$ is large enough, $u_k(T_{m+1})$ can approximate the sums of a sequence, that is, $u_k(T_{m+1}) = u_k^{rest} + \sum_{j=1}^m I_k^j$. In fact, if $T_{m+1} - T_m \geq 3\tau$ holds for $m = 0, 1, 2, ...$, this result can also be rewritten as:

$$\begin{aligned} u_k(T_0) &= u_k^{rest} \\ u_k(T_{m+1}) &= u_k(T_m) + I_k^m \quad (m = 0, 1, 2, ...). \end{aligned} \tag{21}$$

One can find that equation (21) is similar to inference equation (8). Thus if WTA circuit receive appropriate external current, the spiking neurons in WTA circuit with self-connections can accumulate evidence. The problem now is to determine the appropriate input current and demonstrate that this spiking neural network can also implement normalization of the distribution. We have the following theorem.

Theorem 4. Consider the spiking neural network shown in Fig. 2, the rest potential $u_k^{rest} \leq 0$, and the external current $I_k^{ext}(t) = \sum_j \frac{I_k^j}{\varepsilon_0 \tau} \Theta(t - T_j)$ $(j = 1, 2, 3...)$, where $I_k^j = \ln p(y_j | x_j = x^k)$, T_j represents the arriving time of jth current I_k^j, $\Theta(\cdot)$ denotes the Heaviside step function, i.e. $\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise. Then for arbitrary $t \geq 1$, if $u_k^{rest} = \ln p(x_1 = x^k)$, $T_{t+1} - T_t \geq 3\tau$ holds, we can conclude that:

$$u_k(T_{t+1}) = b_t \ln p(x_t = x^k |\ y_1, y_2, ..., y_t), \tag{22}$$

with b_t denoting a constant ($b_t \neq 0$), and

$$\rho_k(T_{t+1}) = \rho\ p(x_t = x^k | y_1, y_2, ..., y_t). \tag{23}$$

Corollary 1. Consider the spiking neural network shown in Fig. 2, the rest potential $u_k^{rest} \leq 0$, and the external current $I_k^{ext}(t) = \sum_j \frac{I_k^j}{\varepsilon_0 \tau} \Theta(t - T_j)$ $(j = 1, 2, 3...)$, where $I_k^j = \ln p(y_j | x_j = x^k)$, T_j represents the arriving time of jth current I_k^j, $\Theta(\cdot)$ denotes the Heaviside step function, i.e. $\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise. T is defined as the minimum time interval, namely $T = \min_t \{T_{t+1} - T_t\}$. Then for all $t \geq 1$, if $u_k^{rest} = \ln p(x_1 = x^k)$, $T \geq 3\tau$ holds, we can conclude that:

$$u_k(T_{t+1}) = b_t \ln p(x_t = x^k |\ y_1, y_2, ..., y_t), \tag{24}$$

with b_t denoting a constant ($b_t \neq 0$), and

$$\rho_k(T_{t+1}) = \rho\ p(x_t = x^k | y_1, y_2, ..., y_t). \tag{25}$$

It is easy to prove with theorem 4.

Theorem 4 and Corollary 1 build the relationship between the dynamics of WTA circuit and the inference equations of HMMs. When the new observation y_t of the HMM comes, an external current of $I_k = \frac{1}{\varepsilon_0 \tau} \ln p(y_t | x_t = x^k)$ is added to the input current of neuron z_k in WTA circuit at time T_t. The membrane potential of each spiking neuron in WTA circuit encodes the logarithm of posterior probability of the hidden variable being in each state (see equation (22)), and the firing rate of each neuron is proportional to the posterior probability of hidden variable in each state (see equation (23)). Moreover, the time course of neural firing rate can implement posterior inference of HMMs. One can read out the inference result by counting spikes from each neuron within a behaviorally relevant time window of a few hundred milliseconds, which is similar to the experimental results of monkey cortex [36], [37].

It is worthwhile to note that for arbitrary t, equation (22) and (23) hold only on the condition that $T_{t+1} - T_t$ is large enough, which has nothing to do with $T_1, T_2, .., T_t$. Thus, if we want to conduct inference of HMM at time t, i.e. to calculate $p(x_t | y_1, y_2, ..., y_t)$, we only need to start from T_t, then wait some time to make $T_{t+1} - T_t \geq 3\tau$ and read out the inference result.

IV. Experiments

In this section, we conduct a series of experiments to validate the proposed spiking neural network. Firstly, we generate data to compare the performance of our spiking neural network with that of belief propagation (optimal solution). Then we move on to demonstrate the convergence of our method by extending the time interval between two evidences. The robustness and applicability of the method to different settings of parameters are provided in Supplemental Materials. At last we scale up the spiking neural network to solve a biologically more realistic task.

A. Testing on the accuracy of our method

We use the data generated from a five-state HMM. The initial distribution $p(x_1)$ is created by randomly generating five numbers from a uniform distribution on $[0, 1]$ and then normalizing them. As the hidden state doesn't change with respect to time, the transition matrix A is set to the identity matrix. With initial distribution $p(x_1)$ and transition matrix A, it is convenient to generate the hidden variables $x_1, x_2, ..., x_m$. The observation data $y_i (i = 1, 2, ..., m)$ is chosen from a Gaussian distribution with mean value being x_i and variance being 1. We set $m = 8$ in the following experiments.

A spiking neural network of WTA circuit with self-connections is used to implement inference of the HMM that is generated with the method above. Here we test the accuracy of our method. Fig. 3a shows how the external current changes with respect to time. The input current for all neurons remains zero before the first evidence y_1 of the HMM comes. Every 150 ms an external current of $I_k = \frac{1}{\varepsilon_0 \tau} \ln p(y_i | x_i = x^k)$ $(i = 1, 2, ..., 8,\ k = 1, 2, ..., 5)$ is added to the input current of neuron k to indicate the new evidence y_i. Note that here we use different colors to denote the currents to different neurons.

The input current will change the membrane potential and the firing activities of the five neurons (shown at the top of Fig. 3b). At the beginning all the neurons can fire. After 700 ms, only neurons 2 and 3 can fire, which indicates that through recurrent accumulation of evidence over time, the state of the hidden variable is most likely to be 2 or 3. At the bottom of Fig. 3b we show the total firing rates of each neuron in the windows of 100 ms (shaded area of Fig. 3b) before each new input current. Note that here we only show four examples. One can find that the firing rate of neuron 3 increases with respect to time and is always larger than that of the other four neurons, which also implies that the state of the hidden variable is most likely to be 3.

In Figs. 3c–3g, we compare the normalized firing rates of spiking neurons (averaged over 500 trials) during inference with the posterior probabilities computed by belief propagation (BP) [38], a commonly used algorithm in machine learning that can get accurate inference results for HMMs [1]. One can find that the spiking neural network can get comparable results as that of BP, which indicates the accuracy of our methods.

B. Testing on the convergence of our method

In Theorem 3 and 4, we proved that our method can converge to the accurate solution as the interval time between

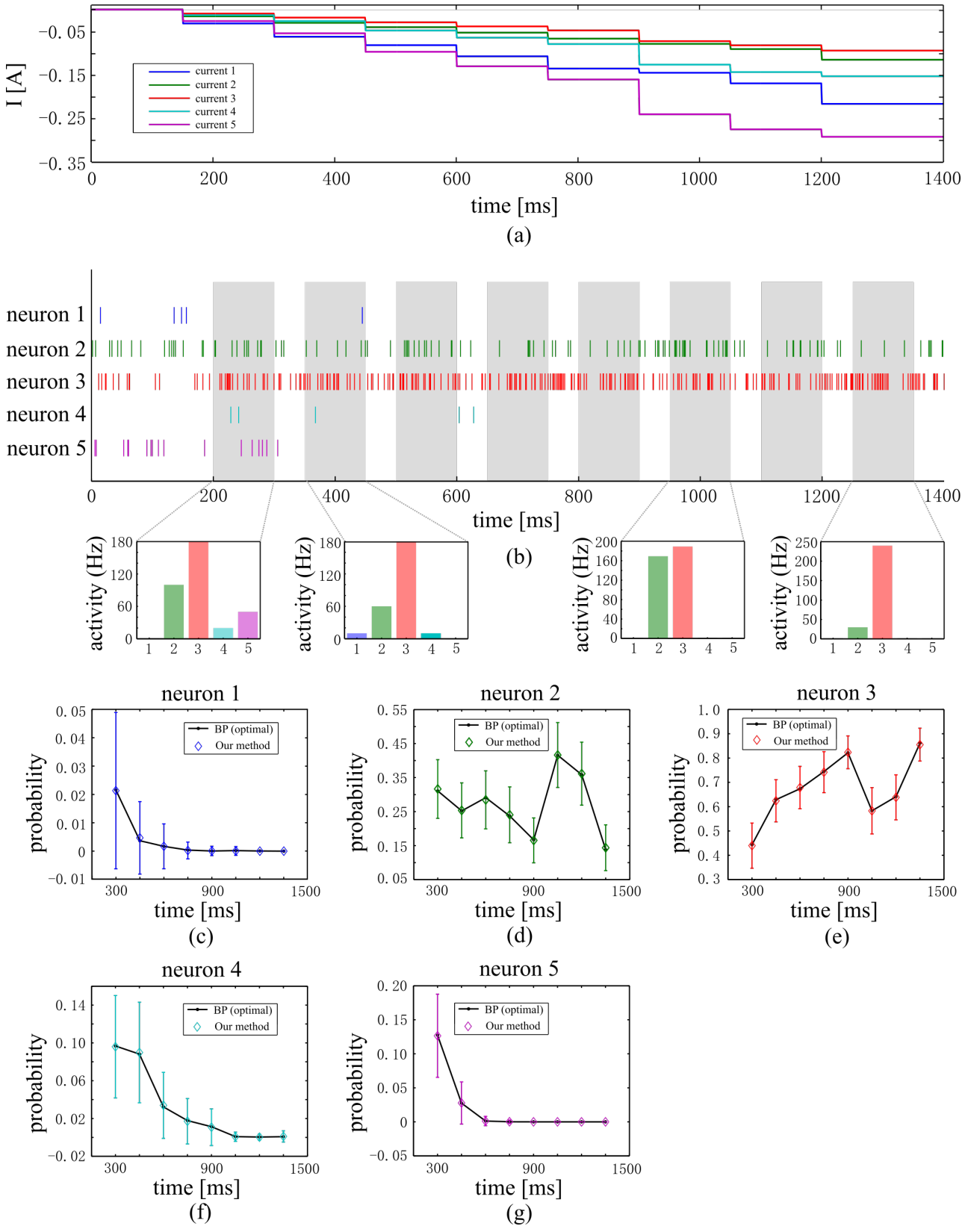


Fig. 3. (a) The external input currents of the five neurons in WTA circuit change with respect to time. Every 150 ms, a new current is added to the input current of each neuron due to the coming of the new evidence of HMM. (b) Top: the firing activities of 5 neurons during inference. Bottom: The total firing rates of different neurons in the windows of 100 ms (shaded area). (c)–(g) Comparison of the normalized firing rate of each spiking neuron (averaged over 500 trials) and accurate inference of HMM with BP.

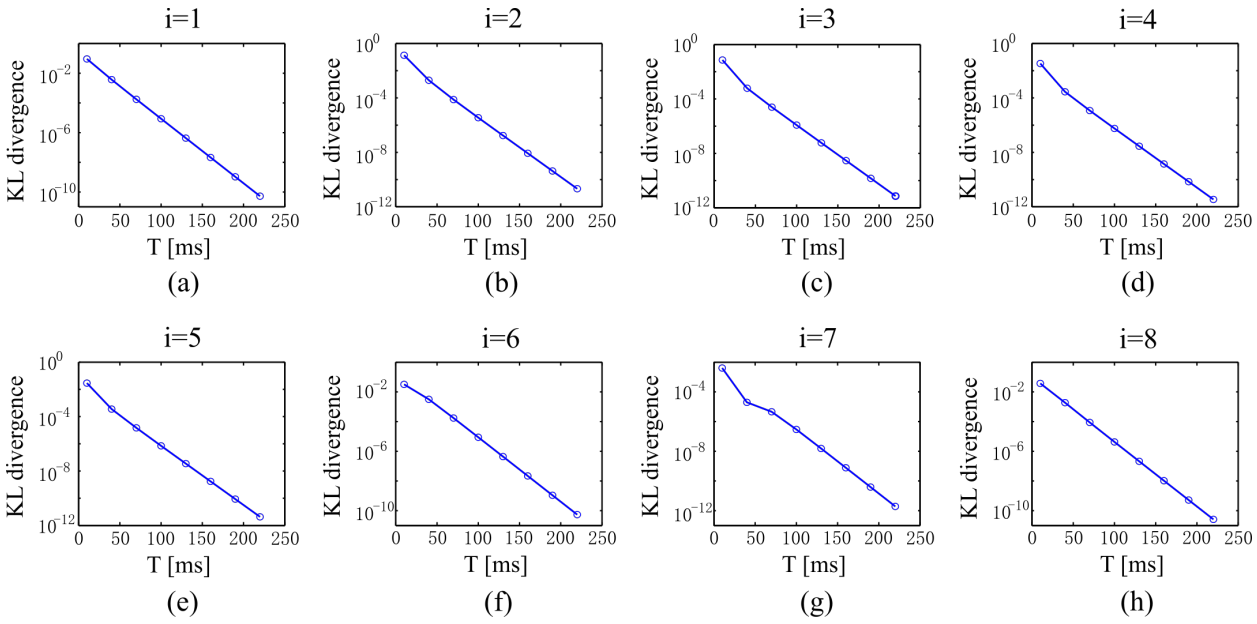


Fig. 4. Kullback-Leibler divergence between the distribution of firing probabilities of all neurons in WTA circuit and the distribution of accurate posterior probability with respect to the interval time. (a) $i = 1$ represents the inference problem of MM with only one evidence, that is, $p(x_1 \mid y_1)$. (b)–(h), the same as (a), but with more evidences.

two evidences increases to infinity. In this experiment we verify this conclusion. The interval time between every two evidences is set as a constant T, ranging from 10 ms to 220 ms. At the beginning of each new external current, we computed the Kullback-Leibler (KL) divergence between the distribution of firing probabilities of all neurons in WTA circuit and the distribution of accurate posterior probability computed by BP, that is, $KL\left(\frac{\exp(u_k(T_{i+1}))}{\sum_k \exp(u_k(T_{i+1}))}, p(x_i \mid y_1, y_2, ..., y_i)\right)$ $(i = 1, 2, ..., 8)$. The results are shown in Fig. 4. We found that the KL divergence decreases as the interval time T increases and converges exponentially to zero when the interval time tends to infinity. These results demonstrate that if the interval time is large enough, our method can implement optimal inference. Also, we can see that the KL divergences are less than 10^{-10} if the interval time is larger than 200 ms. This means $T = 200$ ms is long enough for accurate inference.

C. Cue combination in spiking neural network

Here we investigate whether the spiking neural network can scale up to biologically more realistic task. To do this, we applied the spiking neural network to the task of cue combination. Cue combination is fundamental to our perception [24], which integrates the cues received from multiple sensory modalities in an optimal way. It has been shown by numerous experiments that the process of cue combination is the process of Bayesian inference [39], [40], [41]. We explore whether such process of cue combination can be obtained through our spiking neural network. We consider the simple spiking neural network of WTA circuit with self-connections and currents received from different cues, where the task is to integrate the cues from different sensory modalities.

We first considered the two-cue integration problem, which could be a combination of visual cue and haptic cue. The problem can be modeled by a Bayesian network shown in Fig 5a, in which S represents the location of the stimulus, S_V and S_H denote visual cue and haptic cue respectively. This Bayesian network can also be seen as a HMM with time-invariant hidden variable S, to which the evidences S_V and S_H are given in sequence. Similar to [15], [23], the prior distribution $p(S)$ is supposed to be a uniform distribution, $p(S_V|S)$ and $p(S_H|S)$ are Gaussian distributions with means being S and variances being $\sigma^2_{S_V}$ and $\sigma^2_{S_H}$ respectively. The network receives visual cue and haptic cue in sequence and the problem is to infer the posterior distribution $p(S|S_H, S_V)$. The spiking network to solve this problem is similar to that in Fig. 2. Here the input currents are set to $I_k^{ext}(t) = \frac{I_k^1}{\varepsilon_0 \tau}\Theta(t-T_1) + \frac{I_k^2}{\varepsilon_0 \tau}\Theta(t-T_2)$ with $I_k^1 = \ln p(S_V|S)$ and $I_k^2 = \ln p(S_H|S)$. The parameters are set as follows: $S_V = 55$, $S_H = 65$, $\sigma^2_{S_V} = 16$ and $\sigma^2_{S_H} = 4$. The interval time between every two evidence is supposed to be a constant, which is represented by T. Note that variable S is discretized from 40 to 80 by step 0.5. Thus we need 81 neurons to represent hidden variable S. The results are shown in Fig. 5b–d. The red curve represents the posterior distribution $p(S|S_H, S_V)$, which is the combination of visual cue (green curve) and haptic cue (fuchsia). The blue plus signs are the experimental results of the spiking neural network. One can find that as the interval time T becomes larger, the result of spiking neural network tends to be closer to the accurate curve (red curve). When $T = 100$, the experimental curve is almost the same as the accurate curve, which demonstrates the accuracy of our method.

Next we go a further step and discuss the multi-cue integration. The problem now is to integrate four cues. Similar to two-cue integration problem, we can use a Bayesian network to model it (shown in Fig 5e). Here we use S_1, S_2, S_3 and S_4 to represent four different cues. Supposing that the prior distri-

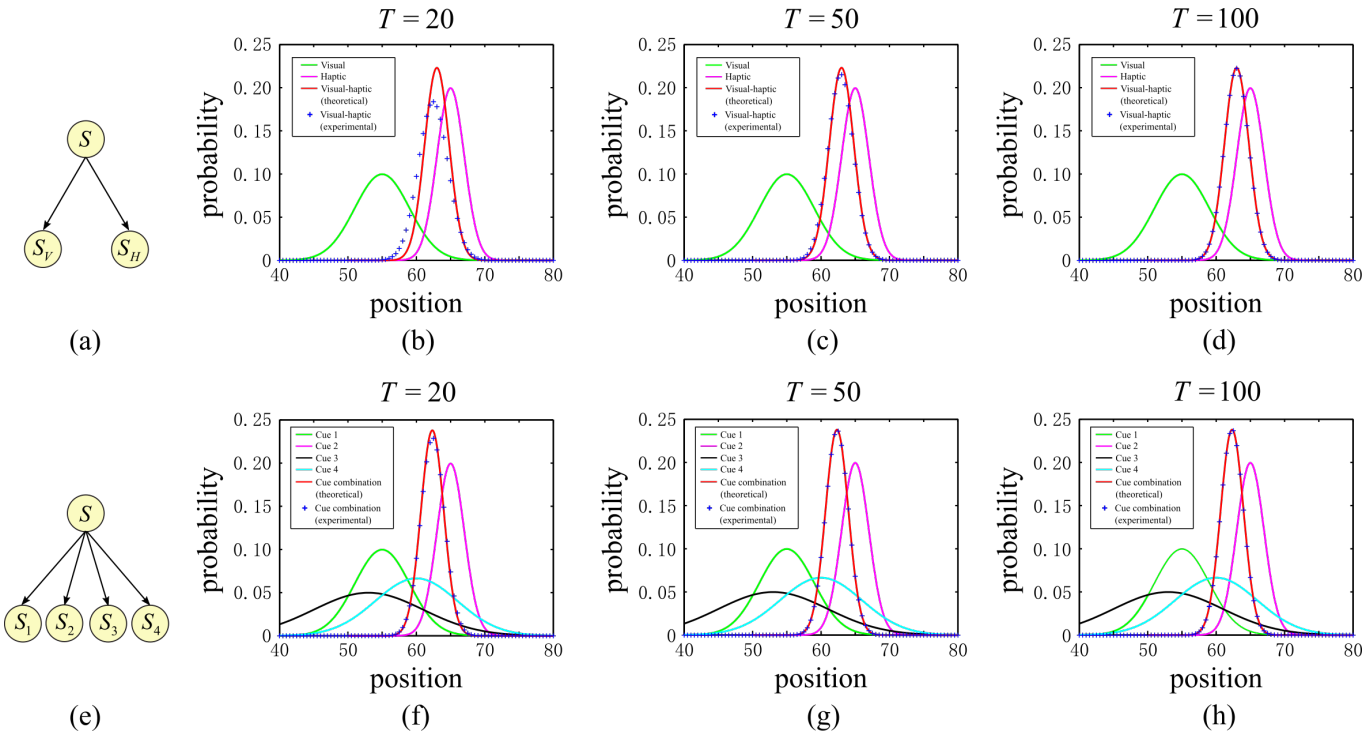


Fig. 5. (a) Bayesian model of two-cue integration. (b)–(d) The performance of spiking neural network with respect to the interval time T. (e) Bayesian model of four-cue integration. (f)–(h) The performance of spiking neural network with respect to the interval time T.

bution $p(S)$ is supposed to be a uniform distribution, $p(S_1|S)$, $p(S_2|S)$, $p(S_3|S)$ and $p(S_4|S)$ are Gaussian distributions with means being S and variances being $\sigma^2_{S_1}$, $\sigma^2_{S_2}$, $\sigma^2_{S_3}$ and $\sigma^2_{S_4}$ respectively. Then we can use spiking neural network to infer the posterior distribution of S given S_1, S_2, S_3 and S_4, that is, $p(S|S_1, S_2, S_3, S_4)$. The results are shown in Fig. 5f–h, here $S_1 = 55$, $S_2 = 65$, $S_3 = 53$, $S_4 = 60$, $\sigma^2_{S_1} = 16$, $\sigma^2_{S_2} = 4$, $\sigma^2_{S_3} = 64$ and $\sigma^2_{S_4} = 36$. The red curve is the theoretical result and the blue plus signs are the experimental results by spiking neural network. Again, we found that as the interval time T becomes larger, the blue plus signs tend to be closer to the accurate curve (red curve). When $T = 100$, the experimental curve is almost the same as the accurate curve, which show the accuracy of our method.

V. Conclusion

In this paper, we show that the dynamics of WTA circuit with self-connections can implement inference of HMM with time-invariant hidden variables. We prove that the membrane potential of each spiking neuron in WTA circuit encodes the logarithm of posterior probability, and the firing intensity of each spiking neuron encodes posterior probability. Theoretical analysis and experimental results demonstrate that our method can get accurate inference result of HMM.

Future work is needed to extend our approach to a more general case of HMM. A possible way is to implement Viterbi algorithm with spiking neural networks. Note that the WTA model used in this paper is a soft WTA model [2], [35], it's interesting to see if other WTA models also work [42], [43]. In addition, one also need to find a biologically plausible way to learn the parameters of HMM. Besides, our present results have suggested to take WTA circuit as the basic unit of computation, which is consistent with previous studies that propose to represent probability distribution with a population of neurons [44], [45], [46]. However, how to implement large-scale Bayesian inference by composition of the basic units of neural circuit is another important yet elusive problem.

Acknowledgment

We would like to thank Jianwu Dong for helpful discussion.

References

[1] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
[2] D. Kappel, B. Nessler, and W. Maass, "STDP installs in winner-take-all circuits an online approximation to hidden Markov model learning," *PLoS Computational Biology*, vol. 10, no. 3, p. e1003511, 2014.
[3] R. P. Rao, "Bayesian computation in recurrent neural circuits," *Neural Computation*, vol. 16, no. 1, pp. 1–38, 2004.
[4] O. Bobrowski, R. Meir, and Y. C. Eldar, "Bayesian filtering in spiking neural networks: Noise, adaptation, and multisensory integration," *Neural Computation*, vol. 21, no. 5, pp. 1277–1320, 2009.
[5] S. Deneve, "Bayesian spiking neurons i: inference," *Neural Computation*, vol. 20, no. 1, pp. 91–117, 2008.
[6] M. Boerlin and S. Deneve, "Spike-based population coding and working memory," *PLoS Computational Biology*, vol. 7, no. 2, p. e1001080, 2011.
[7] R. Corbett-Detig and R. Nielsen, "A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy," *PLoS Genetics*, vol. 13, no. 1, p. e1006529, 2017.
[8] E. Marco, W. Meuleman, J. Huang, K. Glass, L. Pinello, J. Wang, M. Kellis, and G.-C. Yuan, "Multi-scale chromatin state annotation using a hierarchical hidden Markov model," *Nature Communications*, vol. 8, 2017.

[9] S. Gammelmark, K. Mølmer, W. Alt, T. Kampschulte, and D. Meschede, "Hidden Markov model of atomic quantum jump dynamics in an optically probed cavity," *Physical Review A*, vol. 89, no. 4, p. 043839, 2014.

[10] E. Urdapilleta, "Onset of negative interspike interval correlations in adapting neurons," *Physical Review E*, vol. 84, no. 4, p. 041904, 2011.

[11] L. M. Lee and F. R. Jean, "Adaptation of hidden Markov models for recognizing speech of reduced frame rate." *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 2114–2121, 2013.

[12] I. Valera, F. J. Ruiz, and F. Perez-Cruz, "Infinite factorial unbounded-state hidden Markov model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1816–1828, 2016.

[13] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Adaptive hidde Markov model with anomaly states for price manipulation detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 318–330, 2015.

[14] C. Lu, J. M. Schwier, R. M. Craven, L. Yu, R. R. Brooks, and C. Griffin, "A normalized statistical metric space for hidden Markov models," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 806–819, 2013.

[15] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nature Neuroscience*, vol. 9, no. 11, pp. 1432–1438, 2006.

[16] J. M. Beck and A. Pouget, "Exact inferences in a neural implementation of a hidden Markov model," *Neural Computation*, vol. 19, pp. 1344–1361, 2007.

[17] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[18] R. Moreno-Bote and J. Drugowitsch, "Causal inference and explaining away in a spiking network," *Scientific Reports*, vol. 5, 2015.

[19] Q. Yu, R. Yan, H. Tang, K. C. Tan, and H. Li, "A spiking neural network system for robust sequence recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 621–635, 2016.

[20] S. Deneve, "Bayesian inference in spiking neurons," *Advances in Neural Information Processing Systems*, vol. 17, pp. 353–360, 2005, ł.

[21] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[22] M. Robinson, M. R. Azimi-Sadjadi, and J. Salazar, "Multi-aspect target discrimination using hidden Markov models and neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 447–459, 2005.

[23] Z. Yu, F. Chen, J. Dong, and Q. Dai, "Sampling-based causal inference in cue combination and its neural implementation," *Neurocomputing*, vol. 175, pp. 155–165, 2016.

[24] R. L. Seilheimer, A. Rosenberg, and D. E. Angelaki, "Models and processes of multisensory cue combination," *Current Opinion in Neurobiology*, vol. 25, pp. 38–46, 2014.

[25] M. Ursino, C. Cuppini, and E. Magosso, "Neurocomputational approaches to modelling multisensory integration in the brain: a review," *Neural Networks*, vol. 60, pp. 141–165, 2014.

[26] Maas and Wofgang, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[27] G. W, K. WM, N. R, and P. L, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[28] N. Frémaux, H. Sprekeler, and W. Gerstner, "Functional requirements for reward-modulated spike-timing-dependent plasticity," *Journal of Neuroscience*, vol. 30, no. 40, pp. 13 326–13 337, 2010.

[29] J.-P. Pfister, T. Toyoizumi, D. Barber, and W. Gerstner, "Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning," *Neural Computation*, vol. 18, no. 6, pp. 1318–1348, 2006.

[30] R. Jolivet, A. Rauch, H.-R. Lüscher, and W. Gerstner, "Predicting spike timing of neocortical pyramidal neurons by simple threshold models," *Journal of Computational Neuroscience*, vol. 21, no. 1, pp. 35–49, 2006.

[31] R. J. Douglas and K. A. Martin, "Neuronal circuits of the neocortex," *Annual Review of Neuroscience*, vol. 27, pp. 419–451, 2004.

[32] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.

[33] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[34] S. Roy and A. Basu, "An online unsupervised structural plasticity algorithm for spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 4, pp. 900–910, 2017.

[35] B. Nessler, M. Pfeiffer, L. Buesing, and W. Maass, "Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity," *PLoS Computational Biology*, vol. 9, no. 4, p. e1003037, 2013.

[36] T. Yang and M. N. Shadlen, "Probabilistic reasoning by neurons," *Nature*, vol. 447, no. 7148, pp. 1075–1080, 2007.

[37] J. I. Gold and M. N. Shadlen, "The neural basis of decision making," *Annual Review of Neuroscience*, vol. 30, pp. 535–574, 2007.

[38] J. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[39] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[40] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *Trends in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[41] C. Chandrasekaran, "Computational principles and models of multisensory integration," *Current Opinion in Neurobiology*, vol. 43, pp. 25–34, 2017.

[42] P. Tymoshchuk and E. Kaszkurewicz, "A winner-take-all circuit using neural networks as building blocks," *Neurocomputing*, vol. 64, pp. 375–396, 2005.

[43] M. Oster, R. Douglas, and S. Liu, "Computation with spikes in a winner-take-all network." *Neural Computation*, vol. 21, no. 9, p. 2437, 2009.

[44] M. Sahani and P. Dayan, "Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity," *Neural Computation*, vol. 15, no. 10, pp. 2255–2279, 2003.

[45] W. J. Ma and M. Jazayeri, "Neural coding of uncertainty and probability," *Annual Review of Neuroscience*, vol. 37, pp. 205–220, 2014.

[46] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and certainty: distinct probabilistic quantities for different goals," *Nature Neuroscience*, vol. 19, no. 3, pp. 366–374, 2016.