Unification of MAP Estimation and Marginal Inference in Recurrent Neural Networks

Zhaofei Yu, Feng Chen, *Member, IEEE,* and Fei Deng

***Abstract*— Numerous experimental data show that human brain can represent probability distributions and perform Bayesian inference. However, it remains unclear how the brain implements probabilistic inference in the form of neural circuits. Several models have been proposed that aim at explaining how the network of neurons carry out maximum a posterior inference (MAP) estimation and marginal inference, but they are all task specific in that they treat MAP estimation and marginal inference separately. In this brief, we propose that human brain could implement MAP estimation and marginal inference in the same network of neurons. We illustrate our result in hidden Markov models and prove that a recurrent neural network (RNN) implementation of belief propagation can be tuned to perform approximate Bayesian inference (to provide posterior or conditional distribution over the latent causes of observations) or identify the MAP or peak of the joint distribution. The key tuning parameter is a temperature parameter that controls the precision of probability distributions that are optimized. Theoretical analyses and experimental results demonstrate that RNNs can carry out near-optimal MAP estimation and marginal inference.**

***Index Terms*— Maximum a posterior inference (MAP) estimation, marginal inference, networks, recurrent neural, hidden Markov models.**

I. Introduction

NUMEROUS experimental data suggest that the brain is able to represent probability distribution and perform Bayesian inference [1]–[3]. This holds especially for sensory processing [4], [5], motor control [6], [7], and cognitive reasoning [8]–[10]. Although from a macro point of view, probabilistic graphical models, such as Bayesian networks and Markov random fields, have been found to be successful in modeling the inference of human brain [1], [3], [11], how probabilistic inference might be carried out by the network of neurons in the brain is still an open question.

Computational neuroscientists and computer scientists have started to shed light on the relationship between inference equations and dynamics of neural circuits. There have been many proposals for the process theory of Bayesian inference in the brain [12]. A process theory is a theory of the neuronal processes that could implement the belief propagation (BP) (for discrete state space models) [13], [14] or Bayesian filtering for continuous state space models) [15]–[17]. We will focus on discrete latent states, in the form of a hidden Markov model (HMM) or Markov decision process. In this context, variational formulations of BP speak to the biological plausibility of recurrent neural networks (RNNs) as for real neuronal processing [18]. These formulations suggest that neuronal depolarization can be associated with the log probability over latent causes of sensory data, while neuronal firing encodes the expected states of the world generating those sensory samples. From the perspective of inference problem, all these work can be divided into two classes: maximum a posterior (MAP) estimation and marginal inference. Here we review these work as follows.

A. MAP Estimation

The goal of MAP estimation is to find the values of all unobserved variables that maximize a posterior distribution given observation variables. There has been considerable results on implementing BP with neural circuits. Inspired by the experimental evidence that neurons can encode logarithm of probability [19], Ming and Hu [20] proposed to implement BP of Markov random fields with likelihood neurons, belief neurons, and message neurons. Taking another approach, Litvak and Ullman [21] designed the neural circuits for summation and maximization operations, which were the basic computation units in BP for MAP estimation. Thus, they could implement MAP estimation for arbitrary probabilistic graphical models by expanding neural circuits.

B. Marginal Inference

The goal of marginal inference is to compute the posterior or conditional probability distribution of some variables from a joint distribution through the summation over all other variables. Rao [22] proved that the marginal inference equation of BP for HMMs is equivalent to the dynamic equation of RNNs, which means that RNNs can implement marginal inference for HMMs directly. Similarly, Ott and Stoop [23] built the relationship between the inference equation of Markov random fields and the dynamic equation of Hopfield networks, in which all the neurons should have specially initialized messages. Yu *et al.* [24] went further to overcome this unreasonable assumption and came up with a more precise equivalence relation.

As we know, the previously proposed circuits are task specific as they treat marginal inference and MAP estimation separately, which is contrary to the general-purpose computing principle of the human brain. As suggested by Ma and Rahmati [25], the human brain can perform general-purpose inference for a large palette of related tasks with only minor adjustments. It is still unclear whether the human brain could implement marginal inference and MAP estimation in the same neural circuit. To the best of our knowledge, we are the first to try to answer this question. We illustrate our results in HMMs and prove that, with Nesterov's smoothing strategy [26], the inference equation of MAP estimation can be converted to a variant of marginal inference equation. The difference lies in that a temperature parameter is used to shape the distribution, in the sense that marginal inference is a flattened version of MAP estimation with high temperature. Thus, we can unify MAP estimation and marginal inference into the

Manuscript received January 25, 2017; revised November 30, 2017 and February 5, 2018; accepted February 10, 2018. Date of publication March 9, 2018; date of current version October 16, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671266 and Grant 61327902, in part by the Research Project of Tsinghua University under Grant 20161080084, and in part by the National High-tech Research and Development Plan under Grant 2015AA042306. *(Corresponding author: Feng Chen.)*

Z. Yu is with the Department of Automation, Center for Brain-Inspired Computing Research, Tsinghua University, Beijing 100084, China, and also with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: yzf714@126.com).

F. Chen and F. Deng 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 100085, China (e-mail: chenfeng@mail.tsinghua.edu.cn; dengf15@mails.tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2805813

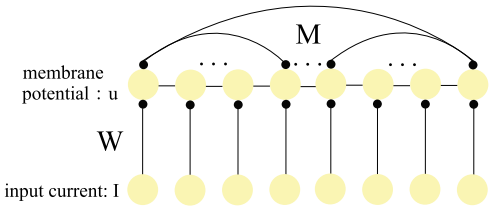

Fig. 1. Recurrent neural network.

same dynamic equation, which is in accordance with that of RNNs. We specify how RNNs can perform unified MAP estimation and marginal inference and illustrate the results with the experiments of synthetic data and visual orientation estimation.

The rest of this brief is organized as follows. Section II briefly reviews RNNs and HMMs. In Section III, we unify MAP estimation equation and marginal inference equation into the dynamic equation of RNNs. We show the experimental results in Section IV and conclude in Section V

II. Preliminaries

In this section, we briefly review RNNs and HMMs.

A. Recurrent Neural Networks

A RNN is a ubiquitous motif of cortical microcircuits, which can perform the same cognitive tasks as animals [27]. Furthermore, it also has been successfully used in natural language processing and nonlinear optimization [28]–[30]. As shown in Fig. 1, the dynamic equation of a RNN can be described as [31]

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{I} + M\mathbf{u} \tag{1}$$

where $\mathbf{I}$ denotes the input currents of the network and $\mathbf{u}$ denotes the membrane potentials of the recurrently connected neurons in the network. W is the feedforward weight matrix and M is the recurrent weight matrix. τ denotes the time constant of the network.

B. Hidden Markov Models

A HMM is a simple dynamic Bayesian network [32], [33], which captures the relationship between hidden variables and observation variables over time t (see Fig. 2). The latent state sequence $Y = \{y_1, y_2, \ldots, y_t\}$ is a first-order Markov chain in which the conditional distribution of the current state $p(y_t|y_1, y_2, \ldots, y_{t-1})$ only depends on the previous latent state y_{t-1}, that is, $p(y_t|y_1, y_2, \ldots, y_{t-1}) = p(y_t|y_{t-1})$. The observation sequence $X = \{x_1, x_2, \ldots, x_t\}$ is governed by the latent state sequence and each observation x_i ($i = 1, 2, \ldots, t$) only depends on the corresponding latent variable y_i. The latent states are discrete states and that the observation sequence can be discrete or continuous (as we will see later). Thus, the joint distribution of HMM can be written in the form

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

The inference problems of HMMs include MAP estimation and marginal inference. By MAP estimation, we refer to the point estimation of the maximum of a posterior. Conversely, marginal inference refers to inferring the posterior or conditional distribution over the latent causes of observations, either exactly or in the sense of approximate Bayesian inference. Specifically, MAP estimation is to compute $\arg\max_{y_1, y_2, \ldots, y_t} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$, while marginal inference is to compute $\sum_{y_1, y_2, \ldots, y_{t-1}} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$.

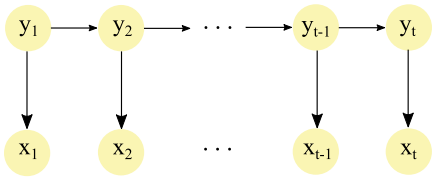

Fig. 2. Hidden Markov model.

III. Neural Implementation of Both MAP Estimation and Marginal Inference

In this section, we will present how RNNs can implement both MAP estimation and marginal inference for HMMs. We first formulate MAP estimation and marginal inference equations into two dynamic equations, respectively, and then, we unify them into the same equation by introducing a temperature parameter T. Finally, we prove that this equation is equivalent to the dynamic equation of RNNs.

A. MAP Estimation With Dynamic Programming

Here we will rewrite MAP estimation equation into a dynamic equation. As $p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t) = (1/Z_t)p(y_1, y_2, \ldots, y_t, x_1, x_2, \ldots, x_t)$ holds for arbitrary observation sequence $x_1, x_2, \ldots, x_t$, where Z_t is a normalized constant, the MAP estimation problem can be rewritten as $\exp(\max_{y_1, y_2, \ldots, y_t} \log(1/Z_t)p(y_1, y_2, \ldots, y_t, x_1, x_2, \ldots, x_t))$. Then, we can solve it with dynamic programming [34]. To be specific, we shall write out the max operator in term of its components by conditional independence

$$\begin{aligned}
&\max_{y_1, y_2, \ldots, y_t} \log \frac{1}{Z_t} p(y_1, y_2, \ldots, y_t, x_1, x_2, \ldots, x_t) \\
&= -\log Z_t \\
&\quad + \max_{y_t}\big(\log p(x_t|y_t) \\
&\qquad + \max_{y_{t-1}}\big(\log p(y_t|y_{t-1}) + \log p(x_{t-1}|y_{t-1}) + \ldots \\
&\qquad + \max_{y_2}\big(\log p(y_3|y_2) + \log p(x_2|y_2) \\
&\qquad + \max_{y_1}(\log p(y_1) + \log p(y_2|y_1) + \log p(x_1|y_1))\big)\big)\big).
\end{aligned} \tag{3}$$

Then, we obtain the following dynamic programming algorithm, which is also called max-sum algorithm [35]:

$$\log f(y_i) = \begin{cases} \log p(y_1), & i = 1 \\ \max\limits_{y_{i-1}} \{\log f(y_{i-1}) + \log p(y_i|y_{i-1}) \\ \quad + \log p(x_{i-1}|y_{i-1})\}, & i = 2, 3, \ldots, t \end{cases} \tag{4}$$

$$\max_{y_1, y_2, \ldots, y_t} \log \frac{1}{Z_t} p(y_1, y_2, \ldots, y_t, x_1, x_2, \ldots, x_t) = \max_{y_t}\{\log f(y_t) + \log p(x_t|y_t) - \log Z_t\}. \tag{5}$$

The definition of MAP estimation of HMM is an exponential function of $\max_{y_1, y_2, \ldots, y_t} \log(1/Z_t)p(y_1, y_2, \ldots, y_t, x_1, x_2, \ldots, x_t)$, which also can be rewritten into a dynamic equation. We have the following theorem:

Theorem 1: Supposing that $F_1(y_t) = f(y_t)p(x_t|y_t)$ holds for $t = 1, 2, \ldots$, then $\max_{y_t}((F_1(y_t))/Z_t)$ equals to the MAP estimation $\max_{y_1, y_2, \ldots, y_t} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$ and satisfies the following dynamic equation:

$$F_1(y_t)^{\frac{1}{T}} = p(x_t|y_t)^{\frac{1}{T}} \sum_{y_{t-1}} F_1(y_{t-1})^{\frac{1}{T}} \; p(y_t|y_{t-1})^{\frac{1}{T}} \tag{6}$$

with T denoting the temperature parameter and tending to 0.

Proof: With the definition $F_1(y_t) = f(y_t)p(x_t|y_t)$ and (5), it is evident that $\max_{y_t}((F_1(y_t))/Z_t) = \max_{y_1,y_2,\ldots,y_t} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$. In order to prove (6), we first rewrite (4) into dynamic equation. Based on Nesterov's smoothing strategy [26] $\max\{a_1, a_2, \ldots, a_n\} \leq T\log\{e^{(a_1/T)} + \cdots + e^{a_n/T}\} \leq \max\{a_1, a_2, \ldots, a_n\} + T\log n$, the max operation can be approximated by the sum operation

$$\begin{aligned}\log f(y_t) &= \max_{y_{t-1}}(\log f(y_{t-1}) + \log p(y_t|y_{t-1}) + \log p(x_{t-1}|y_{t-1})) \\ &\approx T\log\left(\sum_{y_{t-1}} e^{\frac{\log f(y_{t-1})+\log p(y_t|y_{t-1})+\log p(x_{t-1}|y_{t-1})}{T}}\right) \\ &= T\log\left(\sum_{y_{t-1}}(f(y_{t-1})p(y_t|y_{t-1})p(x_{t-1}|y_{t-1}))^{\frac{1}{T}}\right) \quad (7)\end{aligned}$$

which is equivalent to

$$f(y_t)^{\frac{1}{T}} = \sum_{y_{t-1}}(f(y_{t-1})p(y_t|y_{t-1})p(x_{t-1}|y_{t-1}))^{\frac{1}{T}}. \quad (8)$$

Another explanation for (8) is using infinity norm to represent maximum norm. If we multiply $p(x_t|y_t)^{\frac{1}{T}}$ for both sides of (8), we obtain

$$\begin{aligned}&f(y_t)^{\frac{1}{T}}p(x_t|y_t)^{\frac{1}{T}} \\ &\quad = (p(x_t|y_t))^{\frac{1}{T}}\sum_{y_{t-1}}(f(y_{t-1})p(y_t|y_{t-1})p(x_{t-1}|y_{t-1}))^{\frac{1}{T}}. \quad (9)\end{aligned}$$

Hence, from the definition $F_1(y_t) = f(y_t)p(x_t|y_t)$, we conclude that

$$F_1(y_t)^{\frac{1}{T}} = p(x_t|y_t)^{\frac{1}{T}}\sum_{y_{t-1}} F_1(y_{t-1})^{\frac{1}{T}}\ p(y_t|y_{t-1})^{\frac{1}{T}}. \quad (10)$$

B. Implement Marginal Inference With Dynamic Equation

In this section, we formulate the marginal inference equation into a dynamic equation. The marginal probability $p(y_t|x_1, x_2, \ldots, x_t) = \sum_{y_1,y_2,\ldots,y_{t-1}} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$ at time t can be computed with BP

$$\begin{aligned}p(y_t|x_1, x_2, \ldots, x_t) = &\frac{p(x_t|y_t)}{p(x_t|x_1, x_2, \ldots, x_{t-1})} \\ &\cdot\sum_{y_{t-1}} p(y_{t-1}|x_1, x_2, \ldots, x_{t-1})\ p(y_t|y_{t-1}).\end{aligned} \quad (11)$$

Supposing that $F_2(y_t) = p(y_t|x_1, x_2, \ldots, x_t)$, we obtain

$$F_2(y_t) = \frac{p(x_t|y_t)}{p(x_t|x_1, x_2, \ldots, x_{t-1})}\sum_{y_{t-1}} F_2(y_{t-1})\ p(y_t|y_{t-1}). \quad (12)$$

Equation (12) differs from (6) in that a constant $p(x_t|x_1, x_2, \ldots, x_{t-1})$ is used at each time step to normalize the probability. Rao [22] interprets that a form of global recurrent inhibition may implement the normalization. In contrast to this step-by-step normalization, we now show that normalization is only required at the last step. The following theorem demonstrates why this is the case:

Theorem 2: Supposing that $F_3(y_1) = F_2(y_1)$ and $F_3(y_t) = p(x_t|y_t)\sum_{y_{t-1}} F_3(y_{t-1})\ p(y_t|y_{t-1})$ for $t > 1$, then $F_3(y_t) \propto F_2(y_t)$ holds for $t \geq 1$.

Proof: Here we prove it with mathematical induction. Obviously, the statement is true for $t = 1$. Now we show that if $F_3(y_i) \propto F_2(y_i)$ (for some unspecified value of i), then also $F_3(y_{i+1}) \propto F_2(y_{i+1})$ holds. This can be done as follows.

Using the induction hypothesis that $F_3(y_i) = a_i F_2(y_i)$ ($a_i \neq 0$), where a_i is a constant with respect to step i, we obtain the following equation with (12):

$$\begin{aligned}F_3(y_{i+1}) &= p(x_{i+1}|y_{i+1})\sum_{y_i} F_3(y_i)\ p(y_{i+1}|y_i) \\ &= p(x_{i+1}|y_{i+1})\sum_{y_i} a_i F_2(y_i)\ p(y_{i+1}|y_i) \\ &= a_i\, p(x_{i+1}|x_1, x_2, \ldots, x_i)F_2(y_{i+1}) \\ &\propto F_2(y_{i+1}). \quad (13)\end{aligned}$$

Note that $p(x_{i+1}|x_1, x_2, \ldots, x_i)$ is a constant as $x_1, x_2, \ldots, x_{i+1}$ are known observation variables. By mathematical induction, $F_3(y_t) \propto F_2(y_t)$ holds for $t \geq 1$. One can also prove that $F_3(y_t) = p(y_t, x_1, x_2, \ldots, x_t)$. As $Z_t = \sum_{y_t} p(y_t, x_1, x_2, \ldots, x_t) = \sum_{y_t} F_3(y_t)$, the normalized constant Z_t can be computed easily if a neural network can compute $F_3(y_t)$.

Theorem 2 shows that $F_3(y_t)$ can also be used to implement marginal inference of HMMs and $p(y_t|x_1, x_2, \ldots, x_t)$ is the normalization of $F_3(y_t)$. Compare the dynamic equation of $F_1(y_t)$ and $F_3(y_t)$, we can unify them into one equation

$$F(y_t)^{\frac{1}{T}} = p(x_t|y_t)^{\frac{1}{T}}\sum_{y_{t-1}} F(y_{t-1})^{\frac{1}{T}}p(y_t|y_{t-1})^{\frac{1}{T}}. \quad (14)$$

Equation (14) carries out marginal inference when $T = 1$ and MAP estimation when T tends to 0.

C. Unified MAP Estimation and Marginal Inference in Recurrent Neural Networks

Here we build the relationship between inference equation of HMMs and dynamic equation of RNNs. The dynamic equation (1) can be rewritten into a discrete form

$$\mathbf{u}_t = \left(1 - \frac{1}{\tau}\right)\mathbf{u}_{t-1} + \frac{1}{\tau}W\mathbf{I}_{t-1} + \frac{1}{\tau}M\mathbf{u}_{t-1} \quad (15)$$

or equivalently

$$u_t^i = \left(1 - \frac{1}{\tau}\right)u_{t-1}^i + \frac{1}{\tau}\sum_j w_{ij} I_{t-1}^j + \frac{1}{\tau}\sum_j m_{ij} u_{t-1}^j \quad (16)$$

where u_t^i denotes the membrane potential of the ith neuron at time t, I_{t-1}^j denotes input current to the jth neuron at time $t-1$. w_{ij} and m_{ij} are the elements in the ith row and jth column of matrix W and M, respectively.

Then, we rewrite (14) in the log domain

$$\log F(y_t)^{\frac{1}{T}} = \log p(x_t|y_t)^{\frac{1}{T}} + \log\sum_{y_{t-1}} p(y_t|y_{t-1})^{\frac{1}{T}}F(y_{t-1})^{\frac{1}{T}}, \quad (17)$$

that is,

$$\begin{aligned}&\log F(y_t = y^i)^{\frac{1}{T}} \\ &\quad = \log p(x_t|y_t = y^i)^{\frac{1}{T}} \\ &\qquad + \log\sum_j p(y_t = y^i|y_{t-1} = y^j)^{\frac{1}{T}}F(y_{t-1} = y^j)^{\frac{1}{T}} \quad (18)\end{aligned}$$

where $y_t = y^i$ means random variable y_t is in state y^i.

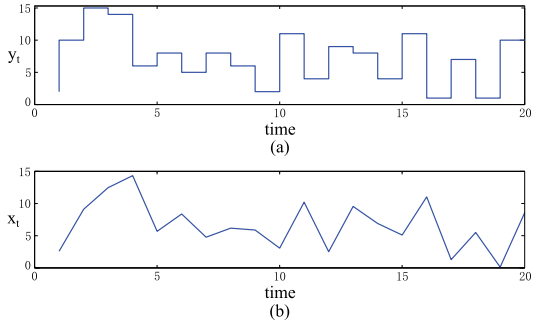

Fig. 3. (a) Example state sequence generated by the HMM. (b) Example observation sequence generated by the HMM.

Obviously, (18) is equivalent to (16) if

$$\tau = 1 \tag{19}$$

$$u_t^i = \log F(y_t = y^i)^{\frac{1}{T}}, \tag{20}$$

$$\sum_j w_{ij} I_{t-1}^j = \log p(x_t | y_t = y^i)^{\frac{1}{T}}, \tag{21}$$

$$\sum_j m_{ij} u_{t-1}^j = \log \sum_j p(y_t = y^i | y_{t-1} = y^j)^{\frac{1}{T}} F(y_{t-1} = y^j)^{\frac{1}{T}}. \tag{22}$$

Equation (20) shows that the neural membrane potential in RNNs encodes the logarithm of the probability, which is possible and has been proved by physiological experiments [19]. The temperature parameter T represents a hyperparameter of RNNs that controls the precision of probability distributions or beliefs. Mathematically, it plays the same role as the precision or sensitivity parameters of classical softmax response rules and related formulations for discrete state spaces. If the feedforward weight matrix W is an identity matrix, (21) means input current $\mathbf{I}_{t-1}$ encode the logarithm of conditional probability $p(x_t|y_t)$, that is, $I_{t-1}^i = (1/T) \log p(x_t|y_t = y^i)$. Equation (22) uses a sum-of-logs to approximate a log-sum, which holds true with the appropriate recurrent weight matrix M. Similar to [22], we compute M with the standard pseudoinverse method. The difference lies in that M is computed only with respect to $F(y_{t-1})$ and $F(y_{t-2})$.

IV. Experiments

A. Synthetic Data

In order to test the accuracy of our method, we use the 20 data points generated from a fifteen state HMM (shown in Fig. 3). The initial distribution of $p(y_1)$ is created by randomly generating 15 numbers from a uniform distribution on $[0, 1]$ and then normalizing them. The transition matrix A is created by randomly generating a square matrix of order 15 from a uniform distribution on $[0, 1]$ and then normalizing each row. With initial distribution $p(y_1)$ and transition matrix A, it is convenient to generate the 20 hidden variables $y_1, y_2, \ldots, y_{20}$ [shown in Fig. 3(a)]. The observation data $x_i (i = 1, 2, \ldots, 20)$ are chosen from Gaussian distribution with mean value y_i and variance 1, which are indicated in Fig. 3(b).

Fig. 4(a) shows the performance of marginal inference with the corresponding RNN. Here we plot the Kullback–Leibler (KL) divergence between the normalization of $F_3(y_t)$ and the actual distribution $p(y_t|x_1, x_2, \ldots, x_t)$ with respect to time. It is observed that the KL divergence is less than 3×10^{-16}, which implies the marginal inference of the RNN is near-optimal. Note that since the values of KL divergence are rather small, which induce numerical artifacts with some rather small negative values, the KL values plotted here are truncated to nonnegative. To see how temperature parameter

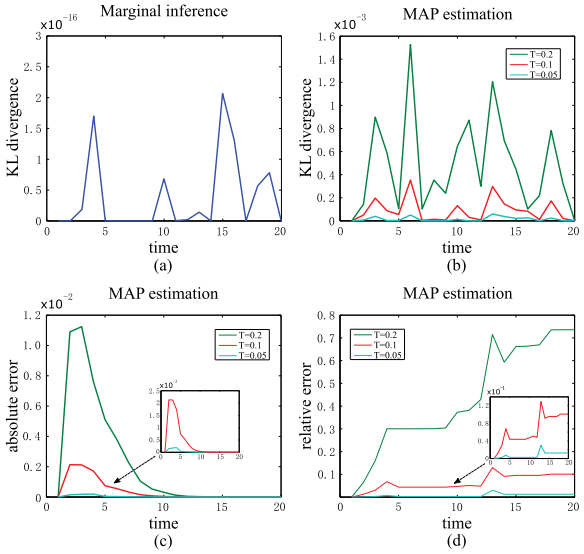

Fig. 4. (a) Performance of marginal inference. (b) Performance of MAP estimation. (c) and (d) Absolute error and relative error of MAP value vary with time and temperature.

affects the performance of MAP estimation of the RNN, we plot the KL divergence between the normalization of $F_1(y_t)$ and the normalization of $\max_{y_1, y_2, \ldots, y_{t-1}} p(y_1, y_2, \ldots, y_t | x_1, x_2, \ldots, x_t)$ with respect to different temperature values [see Fig. 4(b)]. Note that the use of a KL divergence to compare a point probability mass from the MAP estimator with a probability distribution $(F_1(y_t))$ is a bit redundant. However, we use this KL measure for symmetry with the evaluation of posterior inference. From Fig. 4(b), one can find that the accuracy of MAP estimation improves as the temperature decreases, the KL divergence is less than 1×10^{-4} when $T = 0.05$. In addition, we compute the absolute error and relative error of MAP estimation $\max_{y_t} F_1(y_t)$ comparing with the truth $\max_{y_1, y_2, \ldots, y_t} p(y_1, y_2, \ldots, y_t | x_1, x_2, \ldots, x_t)$. As illustrated in Fig. 4(c) and (d), both the absolute error and relative error decrease as the temperature T decreases. The absolute error is less than 2×10^{-4} and the relative error is less than 3% when $T = 0.05$. Consequently, we conclude that the RNN can carry out near-optimal and unified MAP estimation and marginal inference.

B. Visual Orientation Estimation

Next we investigated whether the unified inference framework can be scale up to the biologically more realistic task. To do so, we applied the framework outlined above to a visual discrimination task, which is to estimate stimulus orientation from a sequence of noisy images containing an oriented stimulus [shown in Fig. 5(a)]. This process can be modeled by a HMM and the orientation estimation problem is equivalent to computing the posterior probability or MAP estimation of the HMM given observation variables (noisy images). To be specific, the hidden variable y_t contains 18 states, each of which represents one orientation (0 to 180 in 10 degree step). As the stimulus orientation is fixed during each trial, the transition probability matrix $p(y_{t+1}|y_t)$ is an identity matrix. Similar to [22], the likelihoods equal the convolution of the image with a set of oriented gabor filters $G(y^i)$ [shown in Fig. 5(b)]

$$p(x_t | y_t^i) = a \sum_{m,n} (G(y^i) * x_t)_{m,n} \tag{23}$$

where $(.)_{m,n}$ represents the element in mth row and nth column, a is a small constant that keeps $p(x_t|y_t^i)$ smaller than 1.

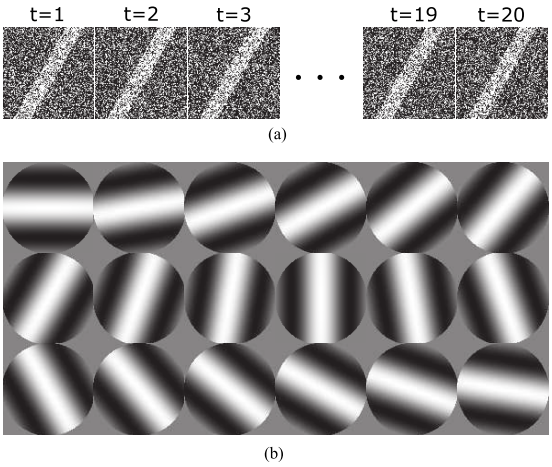

Fig. 5. (a) Sequence of input noisy images. (b) 18 oriented gabor filters (0 to 180 in 10 degree step).

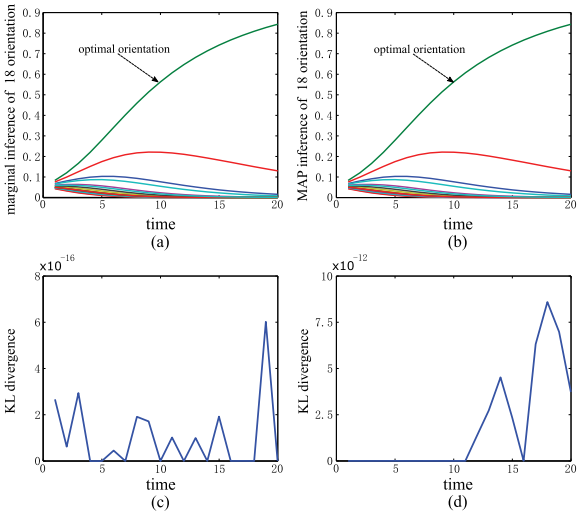

Fig. 6. (a) Distribution of marginal inference change with respect to time. (b) Distribution of MAP estimation change with respect to time. (c) Performance of marginal inference. (d) Performance of MAP estimation.

Fig. 6(a) and (b) shows how the distribution of marginal inference [the normalization of $F_3(y_t)$] and MAP estimation ($F_1(y_t)$) change with respect to time, respectively. Note that Fig. 6(a) and (b) is nearly the same. This phenomenon could be explained as follows. Since the transition probability matrix $p(y_{t+1}|y_t)$ is an identity matrix, the marginal inference $p(y_t|x_1, x_2, \ldots, x_t)$ equals to MAP $\max_{y_1,y_2,\ldots,y_{t-1}} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$. Then, $F_3(y_t) \propto F_1(y_t)$ holds for $t \geq 1$. From Fig. 6(a) and (b), we see that the probability of optimal orientation increases through recurrent accumulation evidence over time. Besides, the probability of optimal orientation is always larger than those of other directions, which implies the accuracy of visual orientation estimation with RNNs. In addition, we compute the KL divergence between $F_3(y_t)$ and the actual distribution $p(y_t|x_1, x_2, \ldots, x_t)$ [see Fig. 6(c)] , and the KL divergence between $F_1(y_t)$ and $\max_{y_1,y_2,\ldots,y_{t-1}} p(y_1, y_2, \ldots, y_t|x_1, x_2, \ldots, x_t)$ [see Fig. 6(d)], which are less than 8×10^{-16} and 1×10^{-11}, respectively. All these results show that a RNN can implement accurate visual orientation estimation.

V. Conclusion

In this brief, we unify MAP estimation and marginal inference of HMMs into the same dynamic equation and then build the relationship between this dynamic equation and that of RNNs. Theoretical analysis and experimental results show that both MAP estimation and marginal inference of HMMs can be carried out accurately by the same RNN. It remains future work to generalize these results to other important probabilistic graphical models. Besides, the question of how to relate this model to biological data requires considerable future work.

Acknowledgment

The authors would like to thank J. K. Liu and J. Dong for their helpful discussion. They would also like to thank the Editors and Reviewers for their valuable comments and helpful guidance.

References

[1] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham, "Probabilistic brains: Knowns and unknowns," *Nature Neurosci.*, vol. 16, no. 9, pp. 1170–1178, 2013.
[2] W. J. Ma and M. Jazayeri, "Neural coding of uncertainty and probability," *Annu. Rev. Neurosci.*, vol. 37, no. 37, pp. 205–220, 2014.
[3] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and certainty: Distinct probabilistic quantities for different goals," *Nature Neurosci.*, vol. 19, no. 3, pp. 366–374, 2016.
[4] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference," *Annu. Rev. Psychol.*, vol. 55, no. 1, pp. 271–304, 2004.
[5] Z. Shi, R. M. Church, and W. H. Meck, "Bayesian optimization of time perception," *Trends Cognit. Sci.*, vol. 17, no. 11, pp. 556–564, 2013.
[6] P. M. Bays and D. M. Wolpert, "Computational principles of sensorimotor control that minimize uncertainty and variability," *J. Physiol.*, vol. 578, no. 2, pp. 387–396, 2007.
[7] K. P. Körding and D. M. Wolpert, "Bayesian integration in sensorimotor learning," *Nature*, vol. 427, no. 427, pp. 244–247, 2004.
[8] N. Chater, J. B. Tenenbaum, and A. Yuille, "Probabilistic models of cognition: Where next?" *Trends Cognit. Sci.*, vol. 10, no. 7, pp. 292–293, 2006.
[9] A. Yuille and D. Kersten, "Vision as Bayesian inference: Analysis by synthesis?" *Trends Cognit. Sciences*, vol. 10, no. 7, pp. 301–308, 2006.
[10] V. Jampani, S. Nowozin, M. Loper, and P. V. Gehler, "The informed sampler: A discriminative approach to Bayesian inference in generative computer vision models," *Comput. Vis. Image Understand.*, vol. 136, pp. 32–44, Jul. 2015.
[11] A. Perfors, J. B. Tenenbaum, T. L. Griffiths, and F. Xu, "A tutorial introduction to Bayesian models of cognitive development," *Cognition*, vol. 120, no. 3, pp. 302–321, 2011.
[12] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active inference: A process theory," *Neural Comput.*, vol. 29, no. 1, pp. 1–49, 2017.
[13] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.
[14] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Norwell, MA, USA: Now Publishers, 2008.
[15] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, Jul. 2000.
[16] K. Friston, K. Stephan, B. Li, and J. Daunizeau, "Generalised filtering," *Math. Problems Eng.*, vol. 2010, Mar. 2010, Art. no. 621670.
[17] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, no. 4, pp. 695–711, Nov. 2012.
[18] K. J. Friston, T. Parr, and B. de Vries, "The graphical brain: Belief propagation and active inference," *Netw. Neurosci.*, vol. 1, no. 4, pp. 381–414, 2017.
[19] R. H. S. Carpenter and M. L. L. Williams, "Neural computation of log likelihood in control of saccadic eye movements," *Nature*, vol. 377, no. 6544, pp. 59–62, 1995.
[20] Y. Ming and Z. Hu, "Modeling stereopsis via Markov random field," *Neural Comput.*, vol. 22, no. 8, pp. 2161–2191, 2010.
[21] S. Litvak and S. Ullman, "Cortical circuitry implementing graphical models," *Neural Comput.*, vol. 21, no. 11, pp. 3010–3056, 2009.
[22] R. Rao, "Bayesian computation in recurrent neural circuits," *Neural Comput.*, vol. 16, no. 1, pp. 1–38, Jan. 2004.

[23] T. Ott and R. Stoop, "The neurodynamics of belief propagation on binary Markov random fields," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1057–1064.
[24] Z. Yu, F. Chen, and J. Dong, "Neural network implementation of inference on binary Markov random fields with probability coding," *Appl. Math. Comput.*, vol. 301, pp. 193–200, May 2017.
[25] W. J. Ma and M. Rahmati, "Towards a neural implementation of causal inference in cue combination," *Multisensory Res.*, vol. 26, nos. 1–2, pp. 159–176, 2013.
[26] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.
[27] H. F. Song, G. R. Yang, and X.-J. Wang, "Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework," *PLoS Comput. Biol.*, vol. 12, no. 2, p. e1004792, 2016.
[28] L. C. Chang, P. A. Chen, and F. J. Chang, "Reinforced two-step-ahead weight adjustment technique for online training of recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1269–1278, Aug. 2012.
[29] Y. Xia and J. Wang, "A recurrent neural network for nonlinear convex optimization subject to nonlinear inequality constraints," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 7, pp. 1385–1394, Jul. 2004.
[30] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[31] P. Dayan and L. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA, USA: MIT Press, 2001.
[32] M. Robinson, M. R. Azimi-Sadjadi, and J. Salazar, "Multi-aspect target discrimination using hidden Markov models and neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 447–459, Mar. 2005.
[33] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.
[34] T. H. Cormen, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2009.
[35] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.