7th Annual International IEEE EMBS Conference on Neural Engineering
Montpellier, France, 22 - 24 April, 2015

Inherently Stochastic Spiking Neurons for Probabilistic Neural Computation

Maruan Al-Shedivat[1,*], Rawan Naous[1], Emre Neftci[2], Gert Cauwenberghs[2] and Khaled N. Salama[1]

***Abstract*— Neuromorphic engineering aims to design hardware that efficiently mimics neural circuitry and provides the means for emulating and studying neural systems. In this paper, we propose a new memristor-based neuron circuit that uniquely complements the scope of neuron implementations and follows the stochastic spike response model (SRM), which plays a cornerstone role in spike-based probabilistic algorithms. We demonstrate that the switching of the memristor is akin to the stochastic firing of the SRM. Our analysis and simulations show that the proposed neuron circuit satisfies a neural computability condition that enables probabilistic neural sampling and spike-based Bayesian learning and inference. Our findings constitute an important step towards memristive, scalable and efficient stochastic neuromorphic platforms.**

I. INTRODUCTION

Neuromorphic engineering has been tightly following theoretical developments of neuroscience by designing efficient silicon neurons [1] as analog or digital very large scale integration (VLSI) circuits. The neuron has been successfully decomposed into three functional blocks: the synapse (spike receiver), the soma (spatio-temporal input signal integrator), and the spike generator; application-dependent implementations for each block has been extensively studied [2].

Natural noise in biological neural networks [3] is seen beneficial for information processing [4], and it can explain probabilistic inference in cortical microcircuits [5]. Nevertheless, neuromorphic research has mainly focused on generalized deterministic integrate-and-fire (I&F) neurons and has overlooked the possibility of building natively probabilistic spiking units [2], [6]. At the same time, recent theoretical work demonstrated that the behavior of networks built of probabilistic neurons can be interpreted as Bayesian computation [7]. Such networks can implement probabilistic sampling and inference algorithms [8], and serve as building blocks for biologically plausible implementations of Boltzmann machines and deep belief networks [9]–[11]. The common approach to add stochasticity to a deterministic neuron is to inject uncorrelated background noise into every neuron [12]. Such an approach lacks power efficiency and constrains scalability.

Circuits based on memristors have become one of the recent trends in neuromorphic engineering as extremely low-power and compact devices [14]. However, the focus has been mainly on the memory and learning properties of the memristor [15]. Moreover, recently discovered non-deterministic behavior of the memristive switching [13], [16], [17] based on stochasticity of nano-filament formation in thin metal-oxide films [18] seems promising in the context of neuromorphic applications: Abrupt switching of the resistance in such devices can be used for generating spontaneous events further converted into spikes. This can provide efficient, low-power, and scalable alternative implementation of the neuron for probabilistic applications.

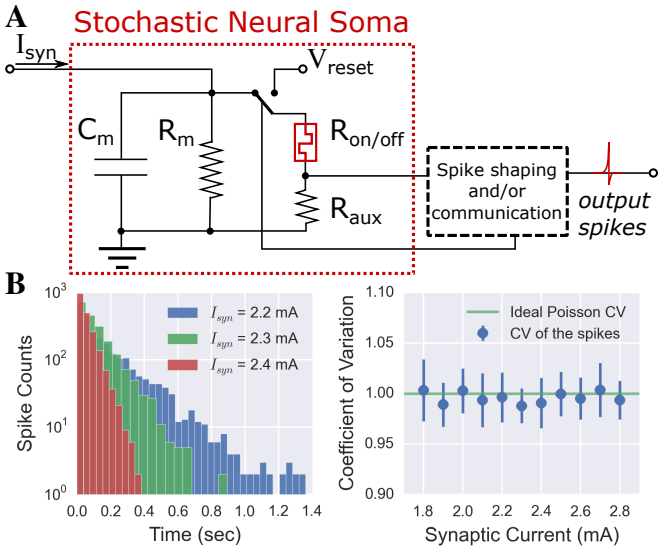

Fig. 1. **(A) The memristor-based stochastic SRM neural soma circuit. (B) Inter-spike interval (ISI) distribution and the coefficient of variation (CV) of ISI for the proposed stochastic neuron. Simulation parameters were close to the data fit [13]:** $V_0 = 156\ mV$, $\tau_0 = 2.85 \cdot 10^5 s$.

In this work, we propose a memristor-based stochastically spiking neural soma circuit that natively realizes the spike response model (SRM) with stochastic firing rate and is compatible with arbitrary synaptic and spike shaping and communication blocks (Fig. 1A). We show that probabilistic switching of the metal-oxide memristor in the sub-threshold regime matches probabilistic firing of the SRM. The inter spike time intervals (ISI) generated by such a model precisely follow a Poisson distribution (Fig. 1B) and satisfy the neural computability condition for Boltzmann distributions [8]. Our simulations based on experimental data suggest that such neurons can be effectively used for building efficient neuromorphic platforms for probabilistic computation.

II. STOCHASTIC NEURON IMPLEMENTATION

The stochastic SRM model is a crucial building block for a number of novel spike-based probabilistic algorithms. Here, we introduce the SRM and show that memristive switching

This work was supported by King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

[1]M.A., R.N., and K.N.S. are with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, KSA.

[2]E.N. and G.C. are with the Institute for Neural Computation, University of California, San Diego, La Jolla, CA 92093, USA.

*Corresponding author: maruan.shedivat@kaust.edu.sa.

statistics, supported by recent experimental results, is akin to the model behavior. The resistance of a memristive device depends on the applied voltage history and is bounded by the minimal R_{on} and maximal R_{off} stable values [19]. When the voltage drop across the memristor is below a certain threshold, the device exhibits a non-deterministic behavior: it experiences spontaneous jumps in resistance from R_{off} to R_{on} or vice versa, depending on the polarity [16]. We propose a simple implementation of the stochastic neural soma block that exploits such memristive switching.

A. SRM Neuron With Stochastic Firing Intensity

The spike response model (SRM) generalizes the classic integrate-and-fire (I&F) neuron [20]. Such a model with stochastic firing threshold was shown to be in a good agreement with experimental cortical data [21]. At any time point, such stochastic neuron has the following instantaneous firing probability (also called stochastic firing intensity):

$$r(V-\theta) = \frac{1}{\tau_s}\exp\left(\frac{V-\theta}{\delta V}\right), \tag{1}$$

where θ is the effective threshold voltage, δV is the width of the spike emission zone, and τ_s is the mean time to spike emission at threshold. In other words, regardless of the membrane potential value, at any time point, this neuron can generate a spike with some probability.

Memristive switching process is intrinsically stochastic in the sub-threshold regime and can be well approximated by the inhomogeneous Poisson process with a time constant that depends on the voltage drop across the memristor [13]:

$$\tau(V) = \tau_0 \exp\left(-\frac{V}{V_0}\right), \tag{2}$$

where τ_0 and V_0 are some constants of the appropriate units. Since the Poisson firing rate is the inverse of the Poisson time constant, it is apparent that (1) and (2) describe identical processes. Based on this correspondence, we can implement stochastic firing using spontaneous memristive switching events for triggering neural spike generation.

Importantly, the firing rate exponentially depends on the membrane voltage and makes such units satisfy the neural computability condition for the Boltzmann distribution [8]:

$$\begin{cases} P[\mathbf{x}] = \dfrac{1}{Z}\exp\left(\mathbf{x}^T\mathbf{W}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right), \\ e^{V_k(t)} \sim P[x_k(t)|\mathbf{x}_{-k}(t)], \end{cases} \tag{3}$$

where k is the neuron index in the network, x_k is the binary (refractory / non-refractory) state of the neuron k, $\mathbf{W}$ and $\mathbf{b}$ are the synaptic and bias weights, respectively. This property allows us to use stochastic SRM, and hence the memristor-triggered spiking as well, for implementing neural networks that can sample from the Boltzmann distribution and perform Bayesian inference (Fig. 2).

B. Neural Soma Circuit

The soma collects pre-synaptic input and generates action potentials. The leaky I&F neuron model captures this functionality by linearly summing synaptic currents and firing

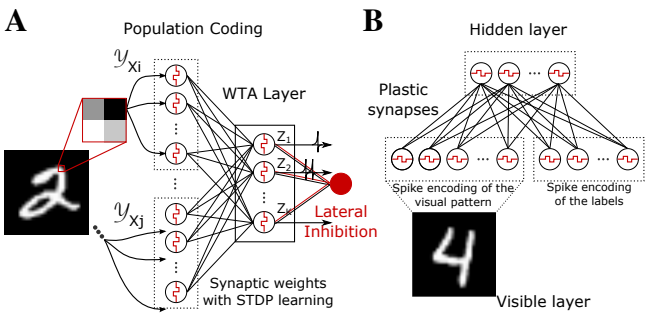

Fig. 2. **(A) Winner-take-all (WTA) network has three layers: the input encoding layer, the WTA output layer, and the inhibitory layer. Neurons of the first two layers are stochastically spiking; the inhibitory layer is triggered whenever it receives a spike. (B) Spiking restricted Boltzmann machine (RBM) has two layers: the visible layer that encodes the data and the labels, and the hidden layer. Each visible neuron is connected to each hidden via a bi-directional plastic synapse.**

an action potential once the membrane voltage V_m crosses a fixed threshold. After each spike, V_m is reset back to the resting potential, V_{reset} [20]. Such behavior can also be described by a more general SRM with a fixed threshold.

To implement stochastic firing, spontaneous memristor switching can be used. In the model we propose, a memristor is added in parallel to the I&F circuit with the membrane resistance R_m (Fig. 1A). The reset mechanism ensures that the initial state of the memristor is R_{off}. Once it switches to R_{on}, the current flowing through it undergoes a strong positive jump, which is detected and converted into a spike.

The voltage across the memristor (the membrane voltage) obeys the leaky integrator dynamics:

$$\tau_m \frac{dV_m}{dt} = -V_m + RI_{\text{syn}}(t), \tag{4}$$

where $\tau_m = RC_m$ is the membrane time constant, and R is the total resistance of R_m connected in parallel to $R_{\text{on/off}}$ and R_{aux}. With a high $R_{\text{on/off}}$-to-R_m ratio, the voltage across the memristor V_m would be effectively independent of the resistance changes of the memristor. Consequently, from (4) it follows that the switching probability (2) is not affected by the memristor dynamics and exclusively depends on the integrated synaptic input, I_{syn}.

The ratio between the currents through the memristor after and before switching, I_{on} and I_{off}, is proportional to the ratio between R_{off} and R_{on}. Depending on fabrication technology, R_{off} to R_{on} ratio of a memristor varies in the range $10^2 \div 10^4$. This enables reliable detection of the switching events.

We simulated the behavior of a single stochastic neuron for some arbitrary noisy synaptic input. Actual memristor parameters τ_0 and V_0 were found by fitting (2) to the experimental data in [13]. The simulations demonstrated that the inter-spike interval (ISI) distribution of the generated spike trains precisely followed the stochastic SRM model (1) for different membrane voltages (Fig. 1B).

III. RESULTS

We consider two applications of the probabilistic neurons. The first is a WTA network that can asynchronously adapt

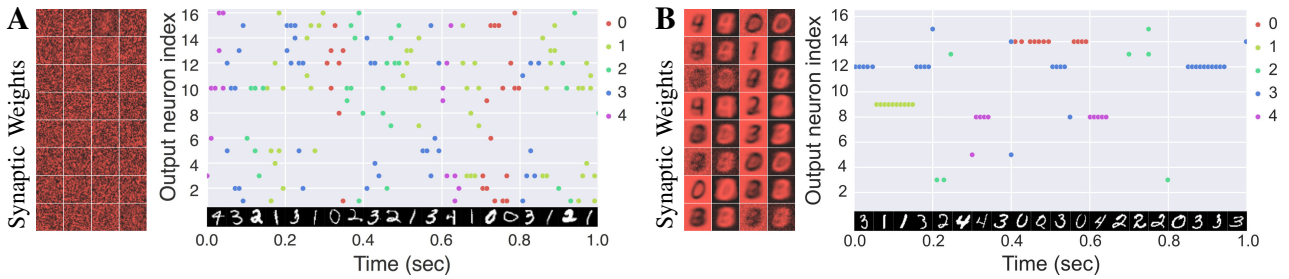

Fig. 3. **Synaptic weights and output spiking. The synapses that connect input and output neurons are represented as images: black pixels denote strong inhibitory connections, while red correspond to strong excitatory ones. (A) Initial random synaptic weights lead to random spiking of the WTA neurons. (B) The network was exposed to the handwritten digits from 0 to 4 (5 classes): Input layer was encoding each input pattern for $40\ ms$ with a silent delay of $10\ ms$ between the patterns. Input images were fed in a randomly shuffled order. After several iteration of learning with STDP rule, each WTA neuron had chosen a pattern and refined its incoming synapses to be sensitive to this pattern the most.**

to the input patterns in unsupervised fashion. The second is the spike-sampling algorithm that can be performed by our neurons: we show how a small network of stochastic neurons can perform sampling from a Boltzmann distribution.

A. Probabilistic Spiking Winner-take-all Network

Probabilistic neurons are necessary for building stochastic spiking WTA networks [7]. The network is composed of three layers: a layer of input encoding neurons is connected in a feedforward fashion to a WTA layer followed by an inhibitory layer; the latter is recursively connected to each WTA neuron (Fig. 2A). The spikes generated by the middle WTA layer represent the network's output. Stochastically firing competing neurons drive the network and asynchronously learn a Naive Bayes probabilistic model by adjusting their incoming synaptic weights using a simple STDP rule.

We implemented the stochastic neural behavior using spontaneous memristive switching model and trained the network on MNIST handwritten digits. Our network had 1568 input neurons (2 encoding neurons per pixel for a 28×28 raster image), 16 competing WTA output neurons, and one inhibitory unit. The spiking output was random and unstructured before learning (Fig. 3A). After a few minutes of learning, each neuron became almost exclusively receptive to one of the patterns from the data, as evidenced by the post-learning spiking behavior of the WTA and by the resulting synaptic matrices (Fig. 3B).

Even though the network is trained in unsupervised manner, each output neuron can be assigned the label of an input pattern it actively spikes for. In this case, the network can be used for classification. We tested WTA with different number of output neurons and achieved the accuracy of about 78% with a network that had 128 output neurons which closely matches the results of the original theoretical work [7].

In order to learn incoming visual patterns, probabilistic WTA performs the so-called spiking expectation-maximization (SEM) algorithm. SEM relies on the assumption that each neuron is stochastic and satisfies the neural computability condition, i.e. its rate should exponentially depend on the membrane voltage (1). We observed that if this condition is violated—for example, by substituting stochastic neurons with deterministic ones—the system no longer converged to the desired Bayesian model.

B. Spike Sampling and Boltzmann Machine

The SRM is an important building block for the neural sampling of Boltzmann machines [8]. This sampling strategy is an alternative to Gibb's sampling commonly used for Boltzmann machines and is ideal for neuromorhic implementations [11]. Furthermore, it enables an on-line spike-driven variant of the commonly used contrastive divergence algorithm for training them. The individual units in the Boltzmann machine are typically endowed with a sigmoid activation function. To emulate this behavior with our neuron model, we include a refractory period after every spike. During this period, the spiking of the neuron is prevented. The connections between units are implemented using linear synapses with a time constant equal to the refractory period.

We trained the restricted Boltzmann machine (RBM) on the MNIST data set using event-driven contrastive divergence [11]. After training, the parameters of the RBM are mapped onto the spiking neural network consisting of 824 visible units and 500 hidden units (Fig. 2B). 784 of these visible units represented 28×28 images of handwritten digits, and the remaining 40 were used for class labels (4 neurons per label). Thanks to its generative properties, the same network is capable of both discrimination (classification) and generation (Fig 4). The classification accuracy of such networks trained with event-driven CD is about 92% (chance is 10%). The results suggest that our memristor-based neuron model can become an ideal hardware building block for neurormorphic Boltzmann machines.

IV. ALTERNATIVE SYSTEMS

An alternative approach to implementing stochastically spiking neurons is based on the I&F model. It was shown that the mean spiking rate of the leaky I&F neuron can approximate the desired stochastic firing intensity when uncorrelated noise is injected into the neuron's membrane [22]. However, this strategy requires a source of uncorrelated noise [11]. For this purpose, one can use a multichannel uncorrelated pseudorandom bit stream generator based on a pair of linear feedback shift registers accompanied with a global clocking mechanism [23]. To implement this, each neuron must be equipped with an XOR element, a low pass filter, and a low gain amplifier to convert the bit stream into

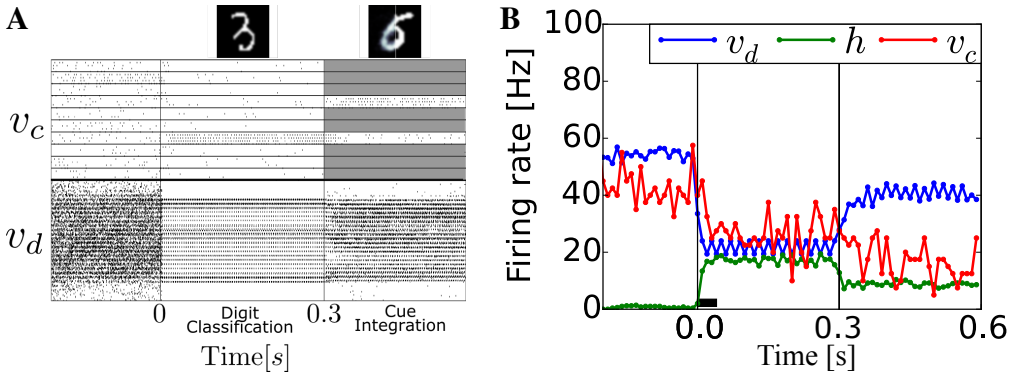

Fig. 4. **The neuromorphic restricted Boltzmann machine (RBM) using our memristive spiking neuron model. (A) Raster plot of the visible layer spiking activity in the neuromorphic RBM: the neurons encoding data, v_d, and the neurons encoding classes, v_c. Until $0s$, no input is provided and the network is free running. At time $0s$, a 28×28 image of a hand-written digit 3 is presented, and the network correctly classifies it as a 3 by activating class neurons corresponding to a 3. At $0.3s$, the right half of a digit 5 is presented, and the class labels are biased such that only 3 and 6 can activate (the others are strongly inhibited). Because only 6 is consistent with the presented data, the network generates the remaining half as a 6 and activates the corresponding class label. (B) Layer-wise population firing rates during the experiment.**

analog Gaussian noise signal [24]. Besides the area overhead created by additional noise generation circuitry, the system becomes less scalable and only approximately matches the SRM. On the contrary, our proposed implementation exploits intrinsic stochasticity of the memristive switching, avoids noise generation expenses, and exactly matches the SRM.

V. CONCLUSION

In this work, we demonstrated a native implementation of the stochastic SRM neuron based on non-determinism of the memristive switching. According to the experimental evidence, we approximated the switching with an inhomogeneous Poisson process and proposed a neural soma circuit that uses the switching for triggering spike events. The probabilistic spiking fully satisfied the neural computability condition necessary for spike-based probabilistic computing. We demonstrated how spiking winner-take-all network and Boltzmann machine can make use of such stochastic neurons. The analysis and simulations confirmed that our neurons natively support probabilistic computation in spiking neural networks.

REFERENCES

[1] C. Mead and M. Ismail, *Analog VLSI implementation of neural systems*. Springer, 1989.

[2] G. Indiveri *et al.*, "Neuromorphic silicon neuron circuits," *Frontiers in neuroscience*, vol. 5, 2011.

[3] A. A. Faisal, L. P. Selen, and D. M. Wolpert, "Noise in the nervous system," *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 292–303, 2008.

[4] M. McDonnell and L. Ward, "The benefits of noise in neural systems: bridging theory and experiment," *Nature Reviews Neuroscience*, vol. 12, no. 7, pp. 415–426, 2011.

[5] W. Maass, "Noise as a resource for computation and learning in networks of spiking neurons," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 860–880, May 2014.

[6] M. Rahimi Azghadi *et al.*, "Spike-based synaptic plasticity in silicon: Design, implementation, application, and challenges," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 717–737, 2014.

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

[8] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons," *PLoS CB*, vol. 7, no. 11, p. e1002211, 2011.

[9] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking deep belief network," *Frontiers in neuroscience*, vol. 7, 2013.

[10] B. U. Pedroni *et al.*, "Neuromorphic adaptations of restricted boltzmann machines and deep belief networks," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013.

[11] E. Neftci *et al.*, "Event-driven contrastive divergence for spiking neuromorphic systems," *Frontiers in neuroscience*, vol. 7, 2014.

[12] H. E. Plesser and W. Gerstner, "Noise in integrate-and-fire neurons: From stochastic input to escape rates," *Neural Computation*, vol. 12, no. 2, pp. 367–384, 2000.

[13] S. H. Jo, K.-H. Kim, and W. Lu, "Programmable resistance switching in nanoscale two-terminal devices," *Nano Letters*, vol. 9, no. 1, pp. 496–500, 2008.

[14] C. Zamarreño-Ramos *et al.*, "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers in neuroscience*, vol. 5, 2011.

[15] S. H. Jo, T. Chang, I. Ebong, B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Letters*, vol. 4, pp. 1297–1301, 2010.

[16] S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," *Nanoscale*, vol. 5, no. 13, p. 5872, 2013.

[17] Q. Li, A. Khiat, I. Salaoru, H. Xu, and T. Prodromakis, "Stochastic switching of TiO2-based memristive devices with identical initial memory states," *Nanoscale research letters*, vol. 9, no. 1, 2014.

[18] Y. Yang *et al.*, "Observation of conducting filament growth in nanoscale resistive memories," *Nature Communications*, vol. 3, p. 732, 2012.

[19] A. Ascoli *et al.*, "Memristor model comparison," *Circuits and Systems Magazine, IEEE*, vol. 13, no. 2, pp. 89–105, 2013.

[20] W. Gerstner and W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[21] 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.

[22] A. Renart, N. Brunel, and X.-J. Wang, "Mean-field theory of irregularly spiking neuronal populations and working memory in recurrent cortical networks," *Computational neuroscience: A comprehensive approach*, pp. 431–490, 2004.

[23] G. Cauwenberghs, "An analog VLSI recurrent neural network learning a continuous-time trajectory," *Neural Networks, IEEE Transactions on*, vol. 7, no. 2, pp. 346–361, 1996.

[24] M. B. Parker and R. Chu, "A VLSI-efficient technique for generating multiple uncorrelated noise sources and its application to stochastic neural networks," *IEEE TCAS*, vol. 38, no. 1, p. 109, 1991.