Extending the Neural Engineering Framework for Nonideal Silicon Synapses

Aaron R. Voelker*, Ben V. Benjamin†, Terrence C. Stewart*, Kwabena Boahen† and Chris Eliasmith*
{arvoelke, tcstewar, celiasmith}@uwaterloo.ca {benvb, boahen}@stanford.edu
*Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, Canada.
†Bioengineering and Electrical Engineering, Stanford University, Stanford, CA, U.S.A.

Abstract—The Neural Engineering Framework (NEF) is a theory for mapping computations onto biologically plausible networks of spiking neurons. This theory has been applied to a number of neuromorphic chips. However, within both silicon and real biological systems, synapses exhibit higher-order dynamics and heterogeneity. To date, the NEF has not explicitly addressed how to account for either feature. Here, we analytically extend the NEF to directly harness the dynamics provided by heterogeneous mixed-analog-digital synapses. This theory is successfully validated by simulating two fundamental dynamical systems in Nengo using circuit models validated in SPICE. Thus, our work reveals the potential to engineer robust neuromorphic systems.

I. THE NEURAL ENGINEERING FRAMEWORK

The field of neuromorphic engineering is concerned with building specialized hardware to emulate the functioning of the nervous system [1]. The Neural Engineering Framework (NEF; [2]) compliments this goal with a theory for “compiling” dynamical systems onto spiking neural networks, and has been used to develop the largest functioning model of the human brain, capable of performing various perceptual, cognitive, and motor tasks [3]. This theory allows one to map an algorithm, expressed in software [4], onto some neural substrate realized in silicon [5]. The NEF has been applied to neuromorphic chips including Neurogrid [5], [6] and a VLSI prototype from ETH Zurich [7].

However, the NEF assumes that the postsynaptic current (PSC) induced by a presynaptic spike is modelled by a first-order lowpass filter (LPF). That is, by convolving an impulse representing the incoming spike with an exponentially decaying impulse-response. Furthermore, the exponential time-constant is assumed to be the same for all synapses within the same population. In silicon, synapses are neither first-order nor homogeneous and spikes are not represented by impulses. Synapse circuits have parasitic elements that result in higher-order dynamics, transistor mismatch introduces variability from circuit to circuit, and spikes are represented by pulses with finite width and height. Previously, these features restricted the overall accuracy of the NEF within neuromorphic hardware (e.g., in [5], [6], [7]).

The silicon synapses that we study here are mixed-analog-digital designs that implement a pulse-extender [9] and a first-order LPF [10], modelled as a second-order LPF to account for parasitic capacitances. We also account for the variability (i.e., heterogeneity) introduced by transistor mismatch in the extended pulse’s width and height, and in the LPF’s two time-constants. In §II, we demonstrate how to extend the NEF to directly harness these features for system-level computation. This extension is tested by software simulation in §IV using the circuit models described in §III.

II. EXTENDING THE NEURAL ENGINEERING FRAMEWORK

The NEF consists of three principles for describing neural computation: representation, transformation, and dynamics [2]. This framework enables the mapping of dynamical systems onto recurrently connected networks of spiking neurons. We begin by providing a self-contained overview of these three principles using an ideal first-order LPF. We then extend these principles to the heterogeneous pulse-extended second-order LPF, and show how this maps onto a target neuromorphic architecture.

A. Principle 1 – Representation

The first NEF principle states that a vector $\textbf{x}(t) \in \mathbb{R}^k$ may be encoded into the spike-trains $\delta_i$ of $n$ neurons with rates:

$$r_i(t) = G_i[\alpha_i \textbf{e}_i \cdot \textbf{x}(t) + \beta_i], \quad i = 1 \ldots n, \quad (1)$$

where $G_i$ is a neuron model whose input current to the soma is the linear encoding $\alpha_i \textbf{e}_i \cdot \textbf{x}(t) + \beta_i$, with gain $\alpha_i > 0$, unit-length encoding vector $\textbf{e}_i$ (row-vectors of $\textbf{E} \in \mathbb{R}^{n \times k}$), and bias current $\beta_i$. The state $\textbf{x}(t)$ is typically decoded from spike-trains (see Principle 2) by convolving them with a first-order LPF that models the PSC triggered by spikes arriving at the synaptic cleft. We denote this filter as $h(t)$ in the time-domain and as $H(s)$ in the Laplace domain:

$$h(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} \iff H(s) = \frac{1}{\tau s + 1}. \quad (2)$$

Traditionally the same time-constant is used for all synapses projecting to a given population.\(^2\)

B. Principle 2 – Transformation

The second principle is concerned with decoding some desired vector function $\textbf{f} : S \to \mathbb{R}^k$ of the represented vector. Here, $S$ is the domain of the vector $\textbf{x}(t)$ represented via Principle 1—typically the unit $k$-cube or the unit $k$-ball. Let $r_i(x)$ denote the expected firing-rate of the $i$th neuron in

\(^1\)These statements also hold for real biological systems [8].

response to a constant input \( x \) encoded via (1). To account for noise from spiking and extrinsic sources of variability, we introduce the noise term \( \eta_1 \sim N(0, \sigma^2) \). Then the matrix \( Df(x) \in \mathbb{R}^{n \times k} \) that optimally decodes \( f(x) \) from the spike-trains \( \delta \) encoding \( x \) is obtained by solving the following problem (via regularized least-squares):

\[
Df(x) = \arg \min_{D \in \mathbb{R}^{n \times k}} \int_S \left\| f(x) - \sum_{i=1}^{n} \left( r_i(x) + \eta_1 \right) d_i \right\|^2 d^k x
\]

\[\Leftrightarrow \sum_{i=1}^{n} \left( \delta_i * h_1 \right)(t) d_i f(x) \approx \left( f(x) * h_1 \right)(t).\]  

The quantity in (4) may then be encoded via Principle 1 to complete the connection between two populations of neurons.\(^3\)

\(\text{C. Principle 3 – Dynamics}\)

The third principle addresses the problem of implementing the following nonlinear dynamical system:

\[\dot{x} = f(x) + u, \quad u(t) \in \mathbb{R}^k\]

Since we take the synapse (2) to be the dominant source of dynamics for the represented vector [2, p. 327], we must essentially “convert” (5) into an equivalent system where the integrator is replaced by a first-order LPF. This transformation is accomplished by driving the filter \( h(t) \) with:

\[w := \tau \dot{x} + x = (\tau f(x) + x) + (\tau u)\]

\[\Leftrightarrow (w * h_1)(t) = x(t),\]  

so that convolution with \( h(t) \) achieves the desired integration. Therefore, the problem reduces to representing \( x(t) \) in a population of neurons using Principle 1, while recurrently decoding \( w(t) \) using the methods of Principle 2 (Fig. 1).

\(\text{D. Extensions to Silicon Synapses}\)

Consider an array of \( m \) heterogeneous pulse-extended second-order LPFs (in the Laplace domain):

\[H_j(s) = \frac{\gamma_j (1 - e^{-\epsilon_j s}) s^{-1}}{\tau_j 1 s + 1} \frac{1}{\tau_j 2 s + 1}, \quad j = 1 \ldots m,\]  

where \( \epsilon_j \) is the width of the extended pulse, \( \gamma_j \) is the height of the extended pulse, and \( \tau_{j,1}, \tau_{j,2} \) are the two time-constants of the LPF. \( H_j(s) \), whose circuit is described in

\(^3\)The effective weight-matrix in this case is \( W = E \left( Df(x) \right)^T \).

\(^4\)This form for (6) is \( \Phi = \left[ x \ x \right] \) and \( \Gamma_j = [1, \ \tau]^T \). Further generalizations are explored in [12].
where \( V \) (Fig. 3) that we use to validate our extensions to the NEF. In general, the coefficients \( 1, \epsilon / 2, \epsilon^2 / 12 \), and so on, correspond to the \([0/2] \) Padé approximants of \( \sum_{n=0}^q \frac{x^n}{(n+1)!} \).

Fig. 3. Silicon synapse and soma. The synapse consists of a pulse-extender (\( M_{R2} \)) and a LPF (\( M_{S1-s} \)). The pulse-extender converts subnanosecond digital-pulses—representing input spikes—into submillisecond current-pulses (\( I_{\text{spk}} \)). Then the LPF filters these current-pulses to produce the synapse’s output (\( I_{\text{syn}} \)). The soma integrates this output current on a capacitor (\( C_s \)) and generates a subnanosecond digital-pulse—representing an output spike—through positive-feedback (\( M_{S3-s} \)). This pulse is followed by a reset pulse, which discharges the capacitor. This schematic has been simplified for clarity.

III. CIRCUIT DESCRIPTION

We now describe the silicon synapse and soma circuits (Fig. 3) that we use to validate our extensions to the NEF. An incoming pulse discharges \( C_r \), which \( I_s \) subsequently charges [9]. As a result, \( M_{R2} \) turns on momentarily, producing an output current-pulse \( I_r \) with width and height:

\[
\epsilon = C_r V_f / I_r, \quad \gamma = I_r,
\]

where \( V_f \) is the gate-voltage at which \( M_{R2} \) can no longer pass \( I_r \). This current-pulse is filtered to obtain the synapse’s output, \( I_{\text{syn}} \), whose dynamics obey:

\[
\tau_1 \frac{dI_{\text{syn}}}{dt} + I_{\text{syn}} = A I_{\text{PE}},
\]

where \( \tau_1 = C_{L1} U_T / I_{r1}, \ A = I_{r2} / I_{r1}, \) and \( U_T \) is the thermal voltage. The above assumes all transistors operate in the subthreshold region and ignores all parasitic capacitances [10]. If we include the parasitic capacitance \( C_{L2} \), a small-signal analysis reveals second-order dynamics:

\[
\tau_1 \tau_2 \frac{d^2I_{\text{syn}}}{dt^2} + (\tau_1 + \tau_2) \frac{dI_{\text{syn}}}{dt} + I_{\text{syn}} = A I_{\text{PE}},
\]

where \( \tau_2 = C_{L2} U_T / \kappa I_{r2} \), and \( \kappa \) is the subthreshold-slope coefficient. This second-order LPF and the aforementioned pulse-extender are modelled together in the Laplace domain by (8) after scaling by \( A \).

The dynamics of the soma are described by:

\[
\frac{C_s U_T}{\kappa I_m} \frac{dI_m}{dt} = I_m + I_{\text{syn}},
\]

where \( I_m \) is the current in the positive-feedback loop, assuming all transistors operate in the subthreshold region [13], [14].

IV. VALIDATION WITH CIRCUIT MODELS

We proceed by validating the extended NEF neuromorphic architecture (see Fig. 2) implemented using the circuit models (see §III) on two fundamental dynamical systems: an integrator and a controlled oscillator. We use Nengo 2.3.1 [4] to simulate this neuromorphic system with a time-step of 50 ms. Test data is sampled independently of the training data used to optimize (3) via regularized least-squares. For comparison with (5)—simulated via Euler’s method—spike-trains are filtered using (2) with \( \tau = 10 \) ms. For each trial, the somatic parameters and synaptic parameters (Fig. 4) are randomly sampled from distributions generated using a model of transistor mismatch (validated in SPICE). These parameters determine each linear transformation \( \Gamma_j \), as defined in (11).

A. Integrator

Consider the one-dimensional integrator, \( \dot{x} = u, \ y = x \). We represent the scalar \( x(t) \) in a population of 512 modelled silicon neurons. To compute the recurrent function, we use (11), and assume \( \ddot{x} = \dot{u} \) is given. To be specific, we optimize (3) for \( D^x \), and then add \( u \) and \( \dot{u} \) to the second and third coordinates of the representation, respectively, to decode:

\[
\Phi = [x, u, \dot{u}]^T.
\]

To evaluate the impact of each extension to the NEF, we simulate our network under five conditions: using standard Principle 3, accounting for second-order dynamics, accounting for the pulse-extender, accounting for the transistor mismatch in \( \tau_1 \), and our full extension (Fig. 5). We find that the last reduces the error by 63%, relative to the first, across a wide range of input frequencies (5–50 Hz).

B. Controlled 2D Oscillator

Consider a controllable two-dimensional oscillator,

\[
\dot{x} = f(x) + u, \ y = x,
\]

\[
f(x_1, x_2, x_3) = [-\omega x_3 x_2, \ \omega x_3 x_1, 0]^T,
\]

where \( \omega = 3 \). We now describe the silicon synapse and soma circuits (Fig. 3) that we use to validate our extensions to the NEF. An incoming pulse discharges \( C_r \), which \( I_s \) subsequently charges [9]. As a result, \( M_{R2} \) turns on momentarily, producing an output current-pulse \( I_r \) with width and height:

\[
\epsilon = C_r V_f / I_r, \quad \gamma = I_r,
\]

where \( V_f \) is the gate-voltage at which \( M_{R2} \) can no longer pass \( I_r \). This current-pulse is filtered to obtain the synapse’s output, \( I_{\text{syn}} \), whose dynamics obey:

\[
\tau_1 \frac{dI_{\text{syn}}}{dt} + I_{\text{syn}} = A I_{\text{PE}},
\]

where \( \tau_1 = C_{L1} U_T / I_{r1}, \ A = I_{r2} / I_{r1}, \) and \( U_T \) is the thermal voltage. The above assumes all transistors operate in the subthreshold region and ignores all parasitic capacitances [10]. If we include the parasitic capacitance \( C_{L2} \), a small-signal analysis reveals second-order dynamics:

\[
\tau_1 \tau_2 \frac{d^2I_{\text{syn}}}{dt^2} + (\tau_1 + \tau_2) \frac{dI_{\text{syn}}}{dt} + I_{\text{syn}} = A I_{\text{PE}},
\]

where \( \tau_2 = C_{L2} U_T / \kappa I_{r2} \), and \( \kappa \) is the subthreshold-slope coefficient. This second-order LPF and the aforementioned pulse-extender are modelled together in the Laplace domain by (8) after scaling by \( A \).

The dynamics of the soma are described by:

\[
\frac{C_s U_T}{\kappa I_m} \frac{dI_m}{dt} = I_m + I_{\text{syn}},
\]

where \( I_m \) is the current in the positive-feedback loop, assuming all transistors operate in the subthreshold region [13], [14].

In general, the coefficients \( 1, \epsilon / 2, \epsilon^2 / 12 \), and so on, correspond to the \([0/2] \) Padé approximants of \( \sum_{n=0}^q \frac{x^n}{(n+1)!} \).
Parameter variability validated in SPICE, for two fundamental examples: an integrator and a controlled oscillator. When compared to the previous standard approach, our extension is shown to reduce the error by 63% and 73% for the integrator and oscillator, respectively. Thus, our theory enables a more accurate mapping of nonlinear dynamical systems onto a recurrently connected neuromorphic architecture using non-ideal silicon synapses. Furthermore, we derive our theory independently of the particular neuron model and encoding. This advance helps pave the way toward understanding how non-ideal physical primitives may be systematically analyzed and then subsequently exploited to support useful computations in neuromorphic hardware.

ACKNOWLEDGEMENTS

This work was supported by CFI and OIT infrastructure, the Canada Research Chairs program, NSERC Discovery grant 261453, ONR grants N000141310419 and N000141512827, and NSERC CGS-D funding. The authors thank Wilten Nicola for inspiring (9) with a derivation for the double-exponential.

REFERENCES