


# Physical Deep Learning Based on Optimal Control of Dynamical Systems

Genki Furuhashi,<sup>1</sup> Tomoaki Niiyama,<sup>2</sup> and Satoshi Sunada<sup>2,3,\*</sup>

<sup>1</sup>*Graduate School of Natural Science and Technology, Kanazawa University Kakuma, Kanazawa, Ishikawa 920-1192, Japan*

<sup>2</sup>*Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University Kakuma-machi Kanazawa, Ishikawa 920-1192, Japan*

<sup>3</sup>*Japan Science and Technology Agency (JST), PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan*

 (Received 22 December 2020; revised 18 February 2021; accepted 19 February 2021; published 31 March 2021)

Deep learning is the backbone of artificial-intelligence technologies, and it can be regarded as a kind of multilayer feedforward neural network. An essence of deep learning is information propagation through layers. This suggests that there is a connection between deep neural networks and dynamical systems in the sense that information propagation is explicitly modeled by the time evolution of dynamical systems. In this study, we perform pattern recognition based on the optimal control of continuous-time dynamical systems, which is suitable for physical hardware implementation. The learning is based on the adjoint method to optimally control dynamical systems, and the deep (virtual) network structures based on the time evolution of the systems are used for processing input information. As a key example, we apply the dynamics-based recognition approach to an optoelectronic delay system and demonstrate that the use of the delay system allows for image recognition and nonlinear classifications using *only a few* control signals. This is in contrast to conventional multilayer neural networks, which require a large number of weight parameters to be trained. The proposed approach provides insight into the mechanisms of deep network processing in the framework of an optimal control problem and presents a pathway for realizing physical computing hardware.

DOI: [10.1103/PhysRevApplied.15.034092](https://doi.org/10.1103/PhysRevApplied.15.034092)

## I. INTRODUCTION

The recent rapid progress of information technologies, including machine learning, has led to studies on alternative computing concepts and hardware, such as neuromorphic processing [1–6], reservoir computing [7–12], and deep learning [13–16]. In particular, deep learning has become a groundbreaking tool for data processing owing to its high-level performance [13]. Furthermore, energy-efficient computing for deep learning is gaining value with the rising need for processing large amounts of data [17]. An underlying key factor of deep learning is its high expressive power, which is the result of the layer-to-layer propagation of information in the deep network. This expressive power enables the representation of extremely complex functions in a manner that cannot be achieved using shallow networks with the same number of neurons [18,19]. Interestingly, recent studies have reported that the information propagation in multilayer systems can be expressed as the time evolution of dynamical systems [20–23]. From the point of view of dynamical systems, the learning process of networks can be regarded

as the optimal control of the dynamical systems [20,21]. This viewpoint suggests that there is a connection between deep neural networks and dynamical systems and indicates the possibility of using dynamical systems as physical deep-learning machines.

In this paper, we reveal the potential of dynamical systems with optimal control for the physical implementation. We propose a deep neural-network-like architecture using dynamical systems with delayed feedback and show that delayed feedback allows for the virtual construction of a deep network structure in a *physically single node* using a time-division multiplexing method. In the proposed approach, the virtual deep network for information propagation comes from the time evolution of delay systems; the systems are optimally controlled such that information processing, including classification, is facilitated. The significant difference between our deep network and ordinary deep neural-network architectures is that the learning via optimal control is realized by only a few control signals and minimal weight parameters, whereas the learning by conventional deep neural networks requires a large number of weight parameters [15,16]. The proposed approach using optimal control is applicable for a wide variety of experimentally controllable systems; it allows for simple

\*sunada@se.kanazawa-u.ac.jp

but large-scale deep networks in physical systems with a few control parameters.

## II. MULTILAYER NEURAL NETWORKS AND DYNAMICAL SYSTEMS

First, we briefly discuss the relationship between multilayer neural networks and dynamical systems. Let a dataset to be learned be composed of  $K$  inputs,  $\mathbf{x}_k \in \mathbb{R}^M$  and their corresponding target vectors,  $\mathbf{t}_k \in \mathbb{R}^L$ , where  $k \in \{1, 2, \dots, K\}$ .  $M$  and  $L$  are dimensions of the inputs and target vectors, respectively. The goal of supervised learning is to find a function that maps inputs onto corresponding targets,  $\mathbf{G} : \mathbf{x}_k \rightarrow \mathbf{t}_k$ . To this end, we consider an output function,  $\mathbf{y} = \tilde{\mathbf{G}}(\mathbf{x}, \mathbf{w}) \in \mathbb{R}^L$ , parameterized by the  $M_w$ -dimensional vector,  $\mathbf{w} \in \mathbb{R}^{M_w}$ , and the following loss function:

$$J = \sum_{k=1}^K \Psi(\mathbf{t}_k, \mathbf{y}_k), \quad (1)$$

where  $\Psi(\mathbf{t}_k, \mathbf{y}_k)$  is a function of the distance between the target  $\mathbf{t}_k$  and the output,  $\mathbf{y}_k = \tilde{\mathbf{G}}(\mathbf{x}_k, \mathbf{w})$ .  $\mathbf{w}$  is determined such that loss function  $J$  is minimized, i.e., output  $\mathbf{y}_k$  corresponds to target  $\mathbf{t}_k$ .

It is well known that a neural-network model with an appropriate activation function is a good candidate for representing function  $\mathbf{G}$  owing to its universal approximation capability [24–26]. In multilayer neural networks, the output,  $\mathbf{y}_k = (y_{0,k}, y_{1,k}, \dots, y_{L-1,k})^T$ , is given by the layer-to-layer propagation of an input,  $\mathbf{x}_k$  [Fig. 1(a)]. The layer-to-layer propagation based on multilayer network structures plays a role in increasing expressivity [18] and enhancing learning performance. In this study, instead of standard multilayer networks, we utilize information propagation in a continuous-time dynamical system,

$$\frac{d\mathbf{r}(t)}{dt} = \mathbf{F}[\mathbf{r}(t), \mathbf{u}(t)], \quad (2)$$

where  $\mathbf{r}(t) \in \mathbb{R}^M$  is the state vector at time  $t$  and  $\mathbf{u}(t) \in \mathbb{R}^{M_u}$  represents a control signal vector. Based on the correspondence between a multilayer network [Fig. 1(a)] and a dynamical system [Fig. 1(b)], we suppose that an input  $\mathbf{x}_k$  is set as an initial state  $\mathbf{r}(0)$ . Additionally, the corresponding output,  $\mathbf{y}_k$ , is given by the time evolution (feedforward propagation) of the state vector,  $\mathbf{r}_k(T_e) = \mathbf{r}(T_e, \mathbf{x}_k)$ , up to the end time  $t = T_e$ , i.e.,  $\mathbf{y}_k = \mathbf{y}[\mathbf{r}_k(T_e), \boldsymbol{\omega}]$ , where  $\boldsymbol{\omega} \in \mathbb{R}^{L \times M}$  is a parameter matrix determined in the training process. Loss function  $J$  is obtained by repeating the aforementioned feedforward propagation for all training data instances and using their outputs. The goal of the learning is to find an optimal control vector,  $\mathbf{u}^*(t)$ , and a parameter vector,  $\boldsymbol{\omega}^*$ , such that  $J$  is minimized, i.e.,  $\mathbf{w}^* = \{[\mathbf{u}^*(t)]_{0 < t \leq T_e}, \boldsymbol{\omega}^*\}$

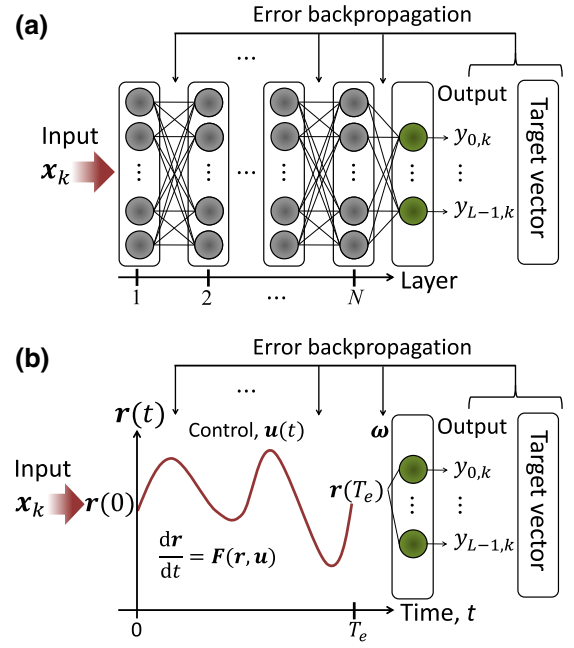


FIG. 1. Schematics of (a) multilayer neural network and (b) dynamical system. In (b), the output,  $\mathbf{y}_k = (y_{0,k}, y_{1,k}, \dots, y_{L-1,k})^T$ , is given by the end state  $\mathbf{r}(T_e)$  and weight parameter  $\boldsymbol{\omega}$ .  $\mathbf{u}(t)$  and  $\boldsymbol{\omega}$  can be updated using a gradient-based optimization algorithm.

$= \operatorname{argmin}_{\mathbf{w}} J$ . It should be noted that learning using a discretized version of Eq. (2) directly corresponds to that using a residual network (ResNet) [21,22].

One strategy for finding optimal controls and parameters is to compute the gradients of the loss function, i.e., the direction of the steepest descent, and to update control and parameter vectors in an iterative manner, i.e.,  $\mathbf{u}(t) \rightarrow \mathbf{u}(t) + \delta \mathbf{u}(t)$  and  $\boldsymbol{\omega} \rightarrow \boldsymbol{\omega} + \delta \boldsymbol{\omega}$ , respectively. In a simple gradient descent method,  $\delta \mathbf{u}(t)$  and  $\delta \boldsymbol{\omega}$  are chosen such that the largest local decrease of  $J$  is obtained.  $\delta \boldsymbol{\omega}$  is simply selected in the opposite direction of the gradient  $dJ/d\boldsymbol{\omega}$ , e.g.,  $\delta \boldsymbol{\omega} = -\alpha_\omega dJ/d\boldsymbol{\omega} = -\alpha_\omega \sum_k (\partial \Psi / \partial \mathbf{y}_k \partial \mathbf{y}_k / \partial \boldsymbol{\omega})$ , where  $\alpha_\omega$  is the learning rate, which is usually a small positive number.  $\delta \mathbf{u}(t)$  is obtained using the adjoint method developed in the context of optimal control problems [27,28] as follows:

$$\delta \mathbf{u}(t) = -\alpha_u \sum_{k=1}^K \left( \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \right)^T, \quad (3)$$

where  $\alpha_u$  is a small positive number,  $\mathbf{F}_k = \mathbf{F}[\mathbf{r}_k(t), \mathbf{u}(t)]$  and  $\mathbf{r}_k(t) = \mathbf{r}(t, \mathbf{x}_k)$ .  $\mathbf{p}_k(t) \in \mathbb{R}^M$  is the adjoint state vector that satisfies the end condition at  $t = T_e$ ,  $\mathbf{p}_k(T_e) = \partial \Psi[\mathbf{t}_k, \mathbf{y}_k(\boldsymbol{\omega}, \mathbf{r}_k)] / \partial \mathbf{r}_k|_{t=T_e}$ . For  $0 \leq t < T_e$ , the time evolution of  $\mathbf{p}_k(t)$  is given by

$$\frac{d\mathbf{p}_k^T(t)}{dt} = -\mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k}. \quad (4)$$

The derivation of Eqs. (3) and (4) is shown in Appendix A. We note that integrating Eq. (4) in the backward direction (from  $t = T_e$  to  $t = 0$ ) corresponds to backpropagation in neural networks. In summary, the algorithm for computing optimal  $\mathbf{u}^*$  and  $\boldsymbol{\omega}^*$  is as follows:

- (i) Set the input,  $\mathbf{x}_k$ , for the  $k$ th data instance as an initial state, i.e.,  $\mathbf{r}(0) = \mathbf{x}_k$ .
- (ii) Forward propagation: starting from initial state  $\mathbf{x}_k$ , integrate Eq. (2) and obtain the end state,  $\mathbf{r}_k(T_e) = \mathbf{r}(T_e, \mathbf{x}_k)$ . Then, compute the output,  $\mathbf{y}_k = \mathbf{y}[\mathbf{r}_k(T_e), \boldsymbol{\omega}]$ .
- (iii) Repeat the forward propagation for all data instances and compute loss function  $J$ .
- (iv) Backpropagation: integrate adjoint Eq. (4) for  $\mathbf{p}_k(t)$  in the backward direction from  $t = T_e$  with  $\mathbf{p}_k(T_e) = \partial\Psi/\partial\mathbf{r}_k(T_e)$ .
- (v) Compute  $\delta\boldsymbol{\omega}$  and  $\delta\mathbf{u}(t)|_{0 < t < T_e}$  using Eq. (3) with appropriate learning rates,  $\alpha_\omega$  and  $\alpha_u$ .
- (vi) Update control signal  $\mathbf{u}(t)$  and parameter  $\boldsymbol{\omega}$ . For the updates, one can use different optimization algorithms [29].

### A. Binary classification problem

Here, we use an abstract dynamical system for solving a typical fundamental problem, the binary classification problem. The goal of the binary classification is to classify a given dataset into two categories labeled as, for example, “0” or “1.” For this, we here consider a simple dynamical model,  $\dot{\mathbf{r}} = \tanh[\mathbf{a}(t)\mathbf{r} + \mathbf{b}(t)]$ , where the state vector is two dimensional,  $\mathbf{r} = (\xi, \eta)^T$ . Weight  $\mathbf{a}(t) \in \mathbb{R}^{2 \times 2}$  and bias  $\mathbf{b}(t) \in \mathbb{R}^2$  are used as control signals. We apply this model to the binary classification problem for a spiral dataset,  $\{\mathbf{x}_k, c_k\}_{k=1}^K$ , where  $\mathbf{x}_k \in \mathbb{R}^2$  is distributed around one of two spirals in the  $\xi\eta$  plane, as shown in Fig. 2(a), and  $\mathbf{x}_k$  is labeled by  $c_k$  as “0” or “1” according to the classes. For the classification, we use one-hot encoding, i.e., target  $\mathbf{t}_k$  corresponding to input  $\mathbf{x}_k$  is set as  $\mathbf{t}_k = (t_{0,k}, t_{1,k})^T = (1, 0)^T$  if  $c_k = 0$  and  $\mathbf{t}_k = (0, 1)^T$  if  $c_k = 1$ . For the output,  $\mathbf{y}_k = (y_{0,k}, y_{1,k})^T$ , the softmax function is used:  $y_{l,k} = \exp(z_{l,k}) / \sum_l \exp(z_{l,k})$ , where  $z_{l,k} = \boldsymbol{\omega}_l^T \mathbf{r}_k(T_e) + \omega_l^{\text{bias}}$ ,  $\boldsymbol{\omega}_l \in \mathbb{R}^2$ , and  $\omega_l^{\text{bias}} \in \mathbb{R}$ . If  $z_{0,k} \gg z_{1,k}$ ,  $\mathbf{y}_k$  approaches  $(1, 0)^T$ , whereas if  $z_{0,k} \ll z_{1,k}$ ,  $\mathbf{y}_k$  approaches  $(0, 1)^T$ .  $J$  is selected as a cross-entropy loss function,  $J = -1/K \sum_{k=1}^K \sum_{l=0,1}^1 t_{l,k} \ln y_{l,k}$ . A training set of  $K = 1000$  data points is used to train  $\mathbf{u}(t) = \{\mathbf{a}(t), \mathbf{b}(t)\}$  and  $\boldsymbol{\omega} = \{\boldsymbol{\omega}_l, \omega_l^{\text{bias}}\}_{l=0,1}$ . The classification accuracy is evaluated using a test set of 1000 data points. For the gradient-based optimization, we use the Adam optimizer [30] with a batch size of  $K$ .

Figures 2(b) and 2(c) show the learning curve and classification accuracy for the training and test datasets, respectively. The loss function monotonically decreases, and the classification accuracy approaches 100%. For sufficient

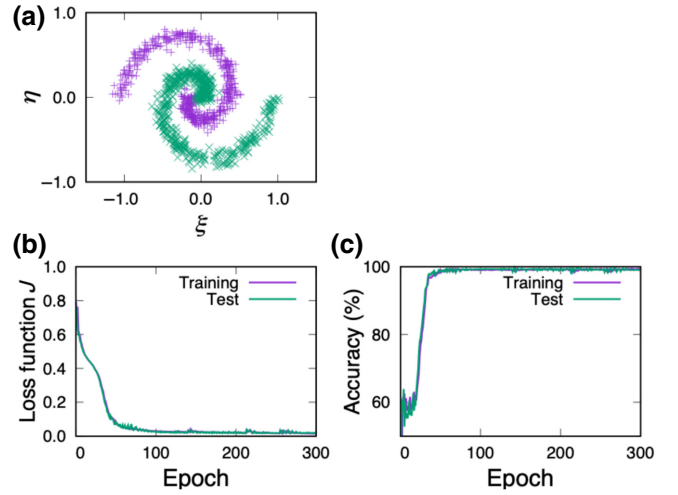


FIG. 2. (a) Spiral dataset for binary classification. The dataset consists of the two data groups labeled as “0” or “1,” colored purple or green, respectively. (b) Loss function  $J$  and (c) classification accuracy as a function of the training (test) epoch.

training over 300 training epochs, the classification accuracy is over 99% when end time  $T_e$  is set as  $200\Delta t$ , where  $\Delta t \approx 0.01$  is the time step used in the simulation. Figure 3 shows the time evolution of the two distributions constituting the spiral dataset. During the evolution, the distributions of the initial states are disentangled [Figs. 3(a)–3(d)] and become linearly separable at end time  $T_e$  [Fig. 3(d)] to aid the classification at the softmax output layer. As a result, any input state can be classified into either of two classes [Fig. 3(e)]. We note that the classification based on disentanglement is different from that of other schemes utilizing dynamical systems, e.g., reservoir computing, whose classification is based on the mapping of input information onto a high-dimensional feature space [9]. In addition, we note that the disentanglement is facilitated as end time  $T_e$  increases, i.e., the number of layers increases, and high classification accuracy is achieved for  $T_e \leq 400\Delta t$ , as shown in Fig. 3(f). A slight decrease in classification accuracy at  $T_e = 600\Delta t$  is attributed to slowdown of the training due to a local plateau of loss function  $J$ , which is occasionally caused in a nonconvex optimization problem [31]. At  $T_e = 600\Delta t$ , we confirmed that a classification accuracy of over 99% is obtained when the number of training epochs is extended to 400.

### III. PHYSICAL IMPLEMENTATION IN DELAY SYSTEMS

In the previous section, we showed binary classification based on optimal control in a two-dimensional dynamical model, where weight  $\mathbf{a}(t) \in \mathbb{R}^{2 \times 2}$  and bias  $\mathbf{b}(t) \in \mathbb{R}^2$  were used as control signals. The excellent performance of this demonstration suggests the realization of deep learning

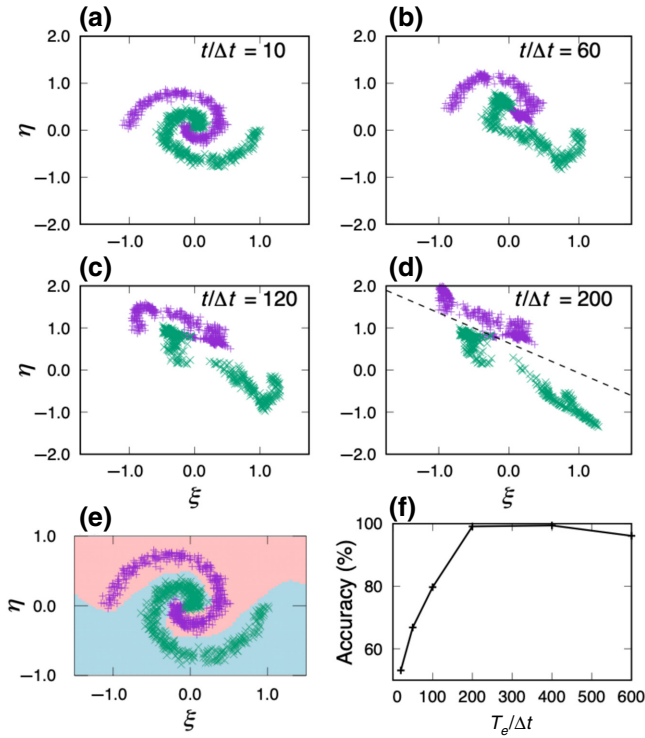


FIG. 3. (a)–(d) Configuration of the states constituting the two spirals over the time evolution in the trained system up to the end time  $T_e = 200\Delta t$ .  $t/\Delta t$  effectively represents the number of layers from the viewpoint of a neural network. The initial spiral distribution is disentangled according to the time evolution (layer-to-layer propagation) and becomes linearly separable at end time  $t = T_e$ . In (d), the dotted line represents the decision boundary for separating the two distributions. (e) Result of binary classification. The inputs can be classified into two regions indicated by pink and blue colors. (f) Classification accuracy for a test dataset as a function of end time  $T_e$ .

in various physical systems, such as coupled oscillators, fluids, and elastic bodies. However, for processing high-dimensional data, a number of signals must be used to control the high-dimensional degrees of freedom in the systems; this may be difficult in terms of physical implementation in an actual system. To overcome the difficulty, we propose the use of delay systems to achieve feasible optimal control of numerous degrees of freedom with limited control signals. It is known that delay systems can be regarded as infinite-dimensional dynamical systems, as visualized in a time-space representation [32]. Furthermore, delay systems can support numerous virtual neurons using a time-division multiplexing method [10]. In addition, they can exhibit various dynamical phenomena, including stable motion, periodic motion, and high-dimensional chaos, with experimentally controllable parameters, e.g., delay time and feedback strength [33,34]. Thus, their high expressivity as well as controllability are promising.

### A. Learning by optimal control in a delay system

We introduce a training method based on the optimal control of a delay system, the time evolution of which is governed by the following equation:

$$\frac{d\mathbf{r}(t)}{dt} = \mathbf{F}[\mathbf{r}(t), \mathbf{r}(t - \tau), \mathbf{u}(t)], \quad (5)$$

where  $\mathbf{r}(t) \in \mathbb{R}^{M_r}$  and  $\mathbf{u}(t) \in \mathbb{R}^{M_u}$  represent the state vector and control signal vector at time  $t$ , respectively, and  $\tau$  is the delay time. The aforementioned equation can be integrated by setting  $\mathbf{r}(t)$  for  $-\tau \leq t \leq 0$  as an initial condition. The information dynamics can intuitively be interpreted by a space-time representation [32] based on the time discretization of Eq. (5),  $\mathbf{r}_n^{j+1} = \mathbf{r}_n^j + \Delta t \mathbf{F}(\mathbf{r}_n^j, \mathbf{r}_{n-1}^j, \mathbf{u}_n^j)$ , where  $t = n\tau + j\Delta t$ ,  $\mathbf{r}_n^j = \mathbf{r}(n\tau + j\Delta t)$ ,  $n \in \{-1, 0, 1, \dots, N-1\}$ ,  $j \in \{0, 1, \dots, M_\tau - 1\}$ , and  $M_\tau = \tau/\Delta t$ . In this representation,  $\mathbf{r}_n^j$  can be regarded as the  $j$ th network node in the  $n$ th layer, which is affected by an adjacent node,  $\mathbf{r}_n^{j-1}$ , and node  $\mathbf{r}_{n-1}^j$  in the  $(n-1)$ th layer, as shown in Fig. 4. Feedforward propagation is carried out as follows: First, an input,  $\mathbf{x}_k = (x_{1,k}, \dots, x_{M_r,k})^T$ , is encoded as  $\mathbf{r}_{-1}^j = \mathbf{r}_{-1}^j(\mathbf{x}_k)$  for  $j \in \{0, 1, \dots, M_\tau\}$  in the initial condition. Then, Eq. (5) is numerically solved to obtain  $\mathbf{r}_{N-1}^j$  in the  $(N-1)$ th layer [corresponding to  $\{\mathbf{r}_k(t)\}_{T_e - \tau \leq t < T_e}$  in continuous time]. The output,  $\mathbf{y}_k \in \mathbb{R}^L$ , is computed using the nodes in the  $(N-1)$ th layer,  $\{\mathbf{r}_{N-1}^j\}_{j=0}^{M_\tau-1}$ , as shown in Fig. 4. In the continuous time representation, the output is defined as  $\mathbf{y}_k = \mathbf{y}(\mathbf{z}_k)$ , where  $\mathbf{z}_k = \int_{T_e - \tau}^{T_e} \omega(t) \mathbf{r}_k(t) dt + \mathbf{b}$ .  $\omega(t) \in \mathbb{R}^{L \times M_r}$  and  $\mathbf{b} \in \mathbb{R}^L$  are the weight and bias parameters to be trained, respectively, and  $\mathbf{r}_k(t)$  is the state vector starting from the initial state  $\{\mathbf{r}(t, \mathbf{x}_k)\}_{-\tau \leq t \leq 0}$ . Finally, loss function  $J$  [Eq. (1)] is computed. To minimize loss function  $J$ , a gradient-based optimization is used, where  $\mathbf{u}(t)$ ,  $\omega(t)$ ,  $\mathbf{b}$  are updated in an iterative manner. In a gradient descent method, the update variations,  $\delta \mathbf{u}(t)$ ,  $\delta \omega(t)$ , and  $\delta \mathbf{b}$ , can be chosen as follows:

$$\delta \mathbf{u}(t) = -\alpha_u \sum_{k=1}^K \left( \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \right)^T, \quad (6)$$

$$\delta \omega = -\alpha_\omega \sum_{k=1}^K \left( \mathbf{r}_k \frac{\partial \Psi}{\partial \mathbf{z}_k} \right)^T, \quad \delta \mathbf{b}_l = -\alpha_b \sum_{k=1}^K \left( \frac{\partial \Psi}{\partial \mathbf{z}_k} \right)^T. \quad (7)$$

The details of these derivations are provided in Appendix B. In the aforementioned equations,  $\mathbf{F}_k = \mathbf{F}[\mathbf{r}_k(t), \mathbf{r}_k(t - \tau), \mathbf{u}(t)]$ ,  $\Psi = \Psi[\mathbf{t}_k, \mathbf{y}(\mathbf{z}_k)]$  is a function of  $\mathbf{z}_k$ , and  $\alpha_i$  for  $i \in \{u, \omega, b\}$  is the learning rate, which is a small positive number. In Eq. (6),  $\mathbf{p}_k(t)$  is the adjoint state vector, which satisfies  $\mathbf{p}_k(T_e) = 0$ .  $\mathbf{p}_k(t)$  can be obtained



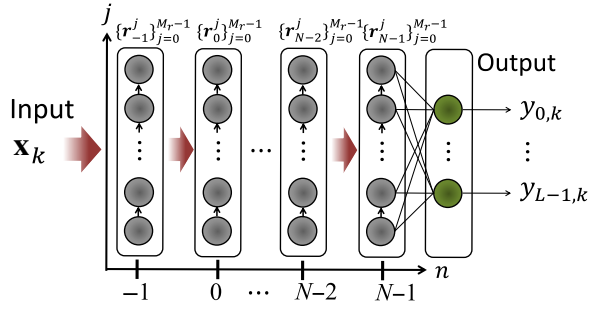


FIG. 4. Schematic of virtual network of a delay system.

by solving the following adjoint equations in the backward direction:

$$\frac{\partial \mathbf{p}_k^T(t)}{\partial t} = -\frac{\partial \Psi}{\partial \mathbf{z}_k} \boldsymbol{\omega}(t) - \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k}, \quad (8)$$

for  $T_e - \tau \leq t < T_e$ , and

$$\frac{d\mathbf{p}_k^T(t)}{dt} = -\mathbf{p}^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} - \mathbf{p}_k^T(t + \tau) \frac{\partial \mathbf{F}_k(t + \tau)}{\partial \mathbf{r}_k}, \quad (9)$$

for  $0 \leq t < T_e - \tau$ .

## B. Optoelectronic delay system

As an effective and feasible example, we consider the use of an optoelectronic delay system, as shown in Fig. 5. The delay system is composed of a laser, optoelectronic intensity modulator, photodetector, and electrical filter to construct a time-delay feedback loop. The time evolution of the system state,  $\mathbf{r}(t) = [\xi(t), \eta(t)]^T$ , is given by the following equations [35]:

$$\tau_L \frac{d\xi}{dt} = -\left(1 + \frac{\tau_L}{\tau_H}\right) \xi - \eta + \beta \cos^2 [u_1(t)\xi(t - \tau) + u_2(t)], \quad (10)$$

$$\tau_H \frac{d\eta}{dt} = \xi, \quad (11)$$

where  $\xi(t)$  is the normalized voltage, and  $\tau_H$  and  $\tau_L$  are the time constants of the low-pass and high-pass filters, respectively.  $\beta$  represents the feedback strength.  $u_1(t)$  and  $u_2(t)$  are electronic signals added to the feedback loop, which are used as control signals in the system.

## C. Results

### 1. Binary classification

We demonstrate binary classification for a spiral dataset, as shown in Fig. 2(a), with the aforementioned optoelectronic delay system. The goal of binary classification is to classify two categories labeled as “0” or “1” for the

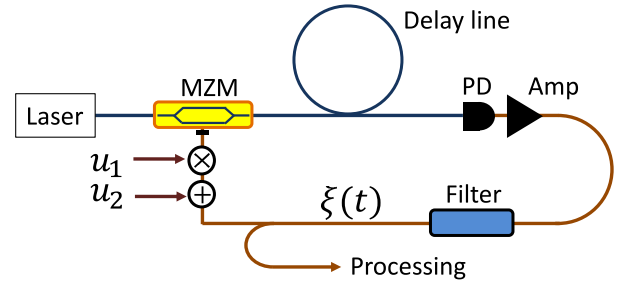


FIG. 5. Schematic of optoelectronic delay system. MZM, Mach-Zehnder modulator; Delay line, optical fiber delay line; PD, photodetector; Amp, electric amplifier; Filter, a two-pole band-pass filter (consisting of low-pass and high-pass filters).

spiral dataset,  $\{\mathbf{x}_k, c_k\}_{k=1}^K$ . In the same manner as that described in Sec. A, target  $\mathbf{t}_k$  is set as  $\mathbf{t}_k = (1, 0)^T$  if  $c_k = 0$ , whereas  $\mathbf{t}_k = (0, 1)^T$  if  $c_k = 1$ . Output  $y_{l,k}$  is set as the softmax function, i.e.,  $y_{l,k} = \exp z_{l,k} / \sum_{l'} \exp z_{l',k}$ , where  $z_{l,k} = \int_{T_e - \tau}^{T_e} \omega_l(t) \xi_k(t) dt + b_l$ ,  $\omega_l(t) \in \mathbb{R}$  and  $b_l \in \mathbb{R}$ . Then,  $J$  is defined as the cross-entropy loss function,  $-1/K \sum_{k=1}^K \sum_{l=0}^{L-1} t_{l,k} \log y_{l,k}$ . In this simulation, the follow-

ing parameters settings are applied:  $\tau_H = 1.59$  ms,  $\tau_L = 15.9$   $\mu$ s, and  $\tau = 230$   $\mu$ s. The input,  $\mathbf{x}_k = (x_{1,k}, x_{2,k})^T$ , is encoded as the initial state of  $\xi$ , i.e.,  $\xi_k(t) = x_{1,k}$  for  $-\tau \leq t < -\tau/2$  and  $\xi_k(t) = x_{2,k}$  for  $-\tau/2 \leq t \leq 0$ . We set  $u_1(t) = 1.0$  and  $u_2(t) = -\pi/4$  as the initial control signals and  $\omega_l(t) = 0$  and  $b_l = 0$  ( $l \in \{0, 1\}$ ) as the initial weight and bias parameters. We use the Adam optimizer [30] with a batch size of  $K$  for the gradient-based optimization. The update equations for  $u_1(t)$ ,  $u_2(t)$ ,  $\omega_l$ , and  $b_l$  are shown in Appendix C. In the aforementioned conditions, the classification accuracy at training epoch 100 is 99.1% when the feedback strength is  $\beta = 3.0$  and the end time is  $T_e = 5\tau$ . To gain insight into the classification mechanism, we investigated the effect of the control signals,  $u_1(t)$  and  $u_2(t)$ , and weights,  $\omega_l(t)$ ,  $l \in \{0, 1\}$ , on the delay dynamics. Figures 6(a)–6(e) show the trained control signals,  $u_1(t)$  and  $u_2(t)$ , weights,  $\omega_0(t)$  and  $\omega_1(t)$ , and four instances of  $\xi_k(t)$  at training epoch 100.  $\xi_k(t)$  in a range of  $T_e - \tau \leq t \leq T_e$  is used for computing the (softmax) outputs,  $y_{l,k}$ , which represents the probability that the input  $\mathbf{x}_k$  is classified as class  $l \in \{0, 1\}$ . Considering that  $y_{l,k}$  is a function of  $z_{l,k} = \int_{T_e - \tau}^{T_e} \omega_l(t) \xi_k(t) dt + b_l$ , we computed  $\tilde{z}_{l,k,l'} = \int_{T_e - \tau}^{T_e} \omega_l(t) \xi_{k,l'}(t) dt$ .  $\tilde{z}_{l,k,l'}$  corresponds to the correlation between (softmax) weights  $\omega_l(t)$  used for the classification as class  $l$  and the  $k_{l'}$ th instance,  $\xi_{k,l'}(t)$ , which starts from the initial states labeled as  $l' \in \{0, 1\}$ . Figures 6(f) and 6(g) show the histograms of the correlation values,  $\tilde{z}_{l,k,l'}$ , for 500 instances. The correlation values are positive for  $l = l'$  in most cases, whereas they are negative for  $l \neq l'$ . In other words,  $\tilde{z}_{l,k,l} > \tilde{z}_{l,k,l'} (l \neq l')$  in most cases. Thus, the softmax output  $y_{l,k}$  for classification as  $l$  is

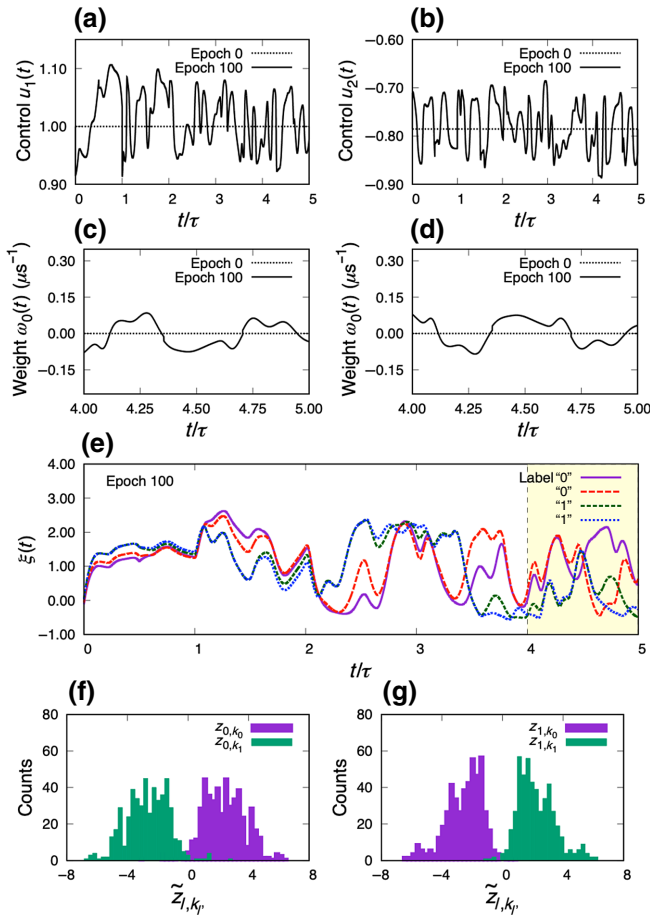


FIG. 6. (a),(b) Control signals,  $u_1(t)$  and  $u_2(t)$  at training epoch 100. (c),(d) Weight parameters,  $\omega_0(t)$  and  $\omega_1(t)$ , at training epoch 100. (e) Four instances of  $\xi_k(t)$  starting from different initial states labeled as “0” or “1,” which are within a distance  $|\mathbf{x}_k - \mathbf{x}_{k'}| < 0.1$ . The end time is set as  $T_e = 5\tau$ .  $\xi_k(t)$  for  $4\tau \leq t < 5\tau$  (indicated by light yellow color) is used to obtain  $z_{l,k} = \int_{T_e-\tau}^{T_e} \omega_l(t)\xi_k(t)dt + b_l$ . (f),(g) Histograms of correlation values,  $\tilde{z}_{l,k_l}$ , between softmax weights,  $\omega_l(t)$ , and the  $k_l$ th instance,  $\xi_{k_l}$ .

activated by  $\xi_{k_l}(t)$  starting from the initial states with the same class  $l$ . These results reveal that the trained control signals control each trajectory such that it is positively correlated to the weight  $\omega_l(t)$  and  $y_{l,k_l}$  is maximized. Figure 7 shows the classification accuracy at training epoch 100 as a function of feedback strength  $\beta$  and end time  $T_e/\tau$ . As seen in this figure, classification performance is low for  $\beta < 2.0$ . In this regime, the system exhibits transient behavior to stable limit cycle motion, which is insensitive to external perturbations [Figs. 8(a) and 8(b)]. This means that it is difficult to control the system. When  $\beta$  increases ( $\beta > 2.5$ ), the system starts to exhibit complex behavior and becomes sensitive to the control signals for a large end time,  $T_e$ , as shown in Figs. 8(c)–8(f). Sensitivity plays a role in aiding the classification. However, extremely high sensitivity makes it difficult to control the system and

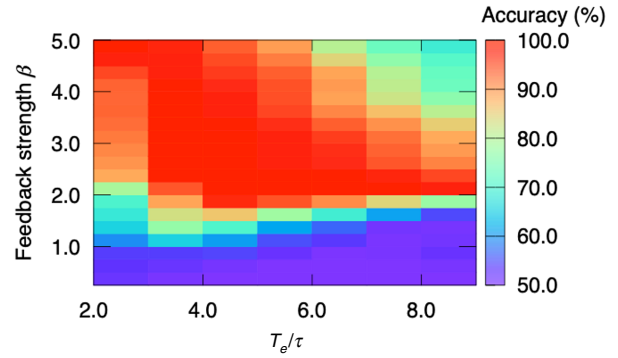


FIG. 7. Classification accuracy as a function of feedback strength  $\beta$  and end time  $T_e$ .

decreases classification performance, as observed for  $\beta > 4.0$  and  $T_e > 7.0\tau$  in Fig. 7. Consequently, a high classification performance of over 99% is achieved with moderate values of  $T_e$  and  $\beta$ , suggesting that the transient behavior around the edge of chaos plays a role in classification.

### 2. MNIST handwritten digit classification

To investigate the classification performance for a higher dimensional dataset, we use the MNIST handwritten digit dataset, commonly used as a standard benchmark for learning [36,37]. The dataset has a training set of 60 000  $28 \times 28$  pixel grayscale images of ten handwritten digits, along with a test set of 10 000 images. To

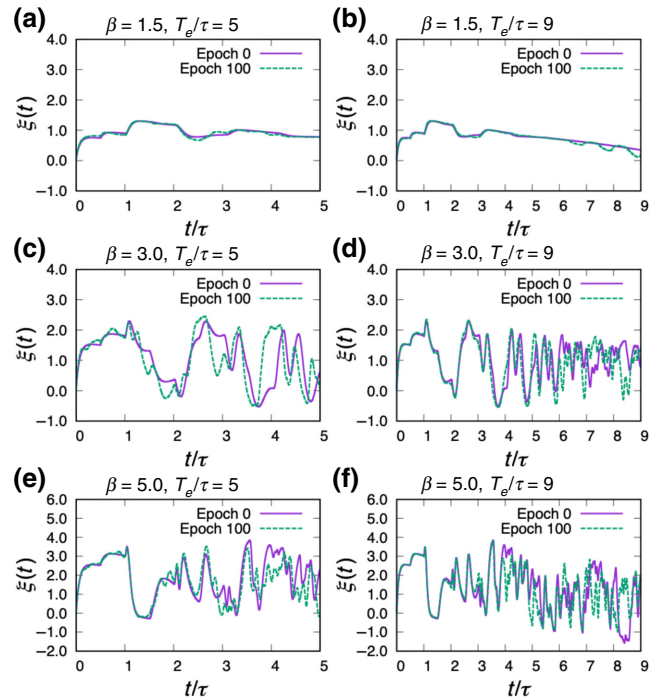


FIG. 8. Typical examples of  $\xi_k(t)$  starting from an initial state at training epochs 0 and 100 for various values of  $\beta$  and  $T_e$ .

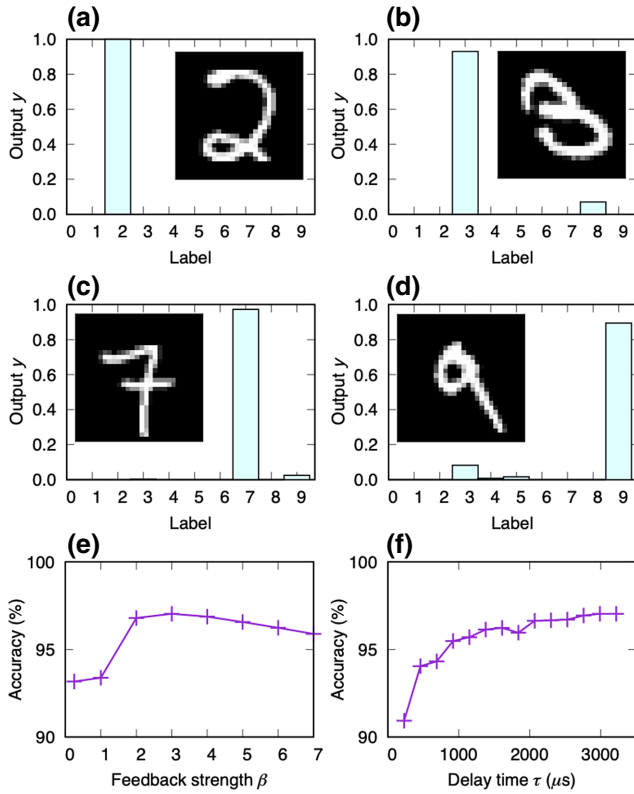


FIG. 9. (a)–(d) Examples of softmax outputs,  $\{y_{l,k}\}_{l=0}^9$ , for the MNIST handwritten images. Each inset shows the input handwritten image corresponding to (a) “2,” (b) “3,” (c) “7,” and (d) “9.” The softmax output,  $y_{l,k}$ , represents the probability that the input image  $k$  belongs to class  $l$ . In (a)–(d),  $\beta = 3.0$ ,  $\tau = 1610 \mu\text{s}$ , and  $T_e = 3\tau$ . (e)–(f) Classification accuracy for the MNIST dataset (10 000 test images) as a function of (e) feedback strength  $\beta$  and (f) delay time  $\tau$  in the delay system. In (e),  $\tau = 3220 \mu\text{s}$  and  $T_e = 3\tau$ . In (f),  $\beta = 3.0$  and  $T_e = 3\tau$ .

set an initial state of the system state, an input image of  $28 \times 28$  pixels is enlarged to double its length and width; it is transformed to a  $m$ -dimensional vector, where  $m = (28 \times 2)^2$ . The vector components are sequentially set as  $\xi_k(t_j)$  at time  $t_j = -\tau + j \Delta t$  with a time interval of  $\Delta t = \tau/M_\tau \approx 0.07 \mu\text{s}$ . The input process is repeated  $M_\tau/m$  times to encode the information of the input image as  $\xi_k(t_j)$  for all  $j \in \{0, 1, \dots, M_\tau\}$ . The training of the delay system is based on the gradient-based optimization using the Adam optimizer with a batch size of 100. The maximum number of epochs to train is set as 50 to ensure the convergence of the training process. As a demonstration of the classifications at epoch 50, we show four examples of the softmax outputs, which represents the probability that the input image belongs to one of the ten classes, in Figs. 9(a)–9(d). Figures 9(e) and 9(f) show the classification accuracy for various values of feedback strength  $\beta$  and delay time  $\tau$ , where  $T_e/\tau = 3$  is fixed. For this image dataset, the delay system with  $\beta = 3.0$  exhibits

relatively high classification performance. The best classification accuracy for this system is 97%. We emphasize that accurate classification is achieved with two training signals,  $u_1(t), u_2(t)$ , and minimal weight parameters,  $\omega_l(t)$ , and  $b_l$ ,  $l \in \{0, 1, \dots, 9\}$ , owing to the time-division multiplexing encoding method based on the delay structure, as discussed in Sec. A. This is in contrast to conventional neural networks, where more than hundreds or thousands of weight parameters need to be trained. We can see that classification accuracy improves as delay time  $\tau$  increases [Fig. 9(f)]. As discussed in Sec. A, the effective number of the network nodes,  $M_\tau$ , depends on  $\tau$  in the delay system, i.e.,  $M_\tau \approx \tau/\Delta t$ . This suggests that larger-scale networks, i.e., systems with a longer delay, play a role in achieving better classification for this large dataset.

#### IV. CONCLUSION

We discussed the applicability of optimally controlled dynamical systems to information processing. The dynamics-based processing provides insight into the mechanism of information processing based on deep network structures, and it can be easily implemented in physical systems. As a particular example, we introduced an optoelectronic delay system. The delay system can be trained to perform nonlinear classification and image recognition with only a few control signals and classification weights based on the time-division multiplexing method. This feature of delay systems is an advantage to hardware implementation of the systems and is distinctively different from conventional neural networks, which require a large number of training parameters. The dynamics-based processing based on optimal control can be applied to various physical systems to construct not only feedforward networks but also recurrent neural networks, including reservoir computing. This provides an alternative direction for physics-based computing.

#### ACKNOWLEDGMENTS

This work was supported, in part, by JSPS KAKENHI (Grant No. 20H042655) and JST PRESTO (Grant No. JPMJPR19M4). The authors thank Professor Kazutaka Kanno and Professor Atsushi Uchida for valuable discussions on optoelectronic delay systems.

#### APPENDIX A: ADJOINT METHOD FOR EQ. (2)

We derive the adjoint equations used to find an optimal control vector  $\mathbf{u}^*(t)$ . The first step is to incorporate the constraint of Eq. (2),  $d\mathbf{r}/dt - \mathbf{F}(\mathbf{r}, \mathbf{u}) = \mathbf{0}$ , into loss function  $J$  with the Lagrangian multiplier (adjoint) vector  $\mathbf{p}_k(t) \in \mathbb{R}^M$  as follows:

$$J_L = \sum_{k=1}^K \left[ \Psi(\mathbf{t}_k, \mathbf{y}_k) + \int_0^{T_e} \mathbf{p}_k^T (\mathbf{F}_k - \dot{\mathbf{r}}_k) dt \right], \quad (\text{A1})$$

where  $\mathbf{r}_k \in \mathbb{R}^M$  is the state vector starting from the initial state,  $\mathbf{r}_k(0) = \mathbf{x}_k \in \mathbb{R}^M$ , and  $\mathbf{F}_k = \mathbf{F}[\mathbf{r}_k, \mathbf{u}(t)]$ . Considering that the second term of Eq. (A1) is rewritten as

$$\int_0^{T_e} \mathbf{p}_k^T (\mathbf{F}_k - \dot{\mathbf{r}}_k) dt = -\mathbf{p}_k^T(T_e) \mathbf{r}_k(T_e) + \mathbf{p}_k^T(0) \mathbf{r}_k(0) + \int_0^{T_e} (\mathbf{p}_k^T \mathbf{F}_k + \dot{\mathbf{p}}_k^T \mathbf{r}_k) dt, \quad (\text{A2})$$

we can obtain

$$J_L = \sum_{k=1}^K [\Psi(\mathbf{t}_k, \mathbf{y}_k) - \mathbf{p}_k^T(T_e) \mathbf{r}_k(T_e) + \mathbf{p}_k^T(0) \mathbf{r}_k(0)] + \sum_{k=1}^K \int_0^{T_e} (\mathbf{p}_k^T \mathbf{F}_k + \dot{\mathbf{p}}_k^T \mathbf{r}_k) dt. \quad (\text{A3})$$

Let  $\delta \mathbf{r}_k$  and  $\delta J_L$  be the variations of  $\mathbf{r}_k$  and loss function  $J_L$  in terms of the variation  $\delta \mathbf{u}$ , respectively. Then, variation  $\delta J_L$  is computed as follows:

$$\begin{aligned} \delta J_L &= \sum_{k=1}^K \left( \left. \frac{\partial \Psi}{\partial \mathbf{r}_k} \right|_{t=T_e} - \mathbf{p}_k^T(T_e) \right) \delta \mathbf{r}_k(T_e) \\ &+ \sum_{k=1}^K \int_0^{T_e} \left[ \mathbf{p}_k^T \left( \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} \delta \mathbf{r}_k + \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \delta \mathbf{u} \right) + \dot{\mathbf{p}}_k^T \delta \mathbf{r}_k \right] dt \\ &= \sum_{k=1}^K \left( \left. \frac{\partial \Psi}{\partial \mathbf{r}_k} \right|_{t=T_e} - \mathbf{p}_k^T(T_e) \right) \delta \mathbf{r}_k(T_e) \\ &+ \sum_{k=1}^K \int_0^{T_e} \left( \mathbf{p}_k^T \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} + \dot{\mathbf{p}}_k^T \right) \delta \mathbf{r}_k dt + \sum_{k=1}^K \int_0^{T_e} \mathbf{p}_k^T \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \delta \mathbf{u} dt, \end{aligned} \quad (\text{A4})$$

where  $\partial \Psi / \partial \mathbf{r}_k = \partial \Psi[\mathbf{t}_k, \mathbf{y}(\mathbf{r}_k, \boldsymbol{\omega})] / \partial \mathbf{r}_k$ , and  $\delta \mathbf{r}_k(0) = 0$  is used in the aforementioned derivation because the initial state  $\mathbf{r}_k(0)$  is fixed as  $\mathbf{r}_k(0) = \mathbf{x}_k$ . As the Lagrangian multiplier,  $\mathbf{p}_k$ , can be set freely, we select  $\mathbf{p}_k$  such that it satisfies the following equation:

$$\mathbf{p}_k^T(T_e) = \left. \frac{\partial \Psi_k}{\partial \mathbf{r}_k} \right|_{t=T_e}, \quad \text{for } t = T_e \quad (\text{A5})$$

$$\frac{d\mathbf{p}_k^T}{dt} = -\mathbf{p}_k^T \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k}, \quad \text{for } 0 < t < T_e. \quad (\text{A6})$$

In this case, we obtain a simple form of  $\delta J_L$  as follows:  $\delta J_L = \sum_k \int_0^{T_e} \mathbf{p}_k^T \partial \mathbf{F}_k / \partial \mathbf{u} \delta \mathbf{u} dt$ . Accordingly, when the variation  $\delta \mathbf{u}$  is set as

$$\delta \mathbf{u} = -\alpha \sum_{k=1}^K \left( \mathbf{p}_k^T \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \right)^T, \quad (\text{A7})$$

with a positive constant  $\alpha$ ,  $\delta J_L = -1/\alpha \int_0^{T_e} \delta \mathbf{u}^2 dt \leq 0$  is satisfied. Thus,  $J$  monotonically decreases when  $\mathbf{u}(t)$  is updated as  $\mathbf{u}(t) \rightarrow \mathbf{u}(t) + \delta \mathbf{u}$ .

## APPENDIX B: ADJOINT METHOD FOR EQ. (5)

We consider the update variations,  $\delta \mathbf{u}(t)$ ,  $\delta \boldsymbol{\omega}(t)$ , and  $\delta \mathbf{b}$ , for the optimization of the delay system that obeys Eq. (5). In the same manner as that shown in Appendix A, we consider the augmented loss function  $J_L$  incorporating Eq. (5) as follows:

$$J_L = \sum_{k=1}^K \Psi[\mathbf{t}_k, \mathbf{y}_k(\mathbf{z}_k)] + \sum_{k=1}^K \int_0^{T_e} \mathbf{p}_k^T (\mathbf{F}_k - \dot{\mathbf{r}}_k) dt, \quad (\text{B1})$$

where  $\mathbf{z}_k = \int_{T_e-\tau}^{T_e} \boldsymbol{\omega} \mathbf{r}_k dt + \mathbf{b}$ ,  $\mathbf{F}_k = \mathbf{F}[\mathbf{r}_k(t), \mathbf{r}_k(t-\tau), \mathbf{u}(t)]$ , and  $\mathbf{p}_k$  is the Lagrangian multiplier (adjoint vector). Variation  $\delta J$  in terms of the variation  $\delta \mathbf{u}$  is expressed as

$$\begin{aligned} \delta J_L &= \sum_{k=1}^K \left( \frac{\partial \Psi}{\partial \mathbf{z}_k} \int_{T_e-\tau}^{T_e} \boldsymbol{\omega}(t) \delta \mathbf{r}_k dt - \mathbf{p}_k^T(T_e) \delta \mathbf{r}_k(T_e) \right) \\ &+ \sum_{k=1}^K \int_0^{T_e} \mathbf{p}_k^T \left( \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} \delta \mathbf{r}_k + \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_{k,\tau}} \delta \mathbf{r}_{k,\tau} + \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \delta \mathbf{u} \right) dt \\ &+ \sum_{k=1}^K \int_0^{T_e} \dot{\mathbf{p}}_k^T \delta \mathbf{r}_k dt, \end{aligned} \quad (\text{B2})$$

where  $\mathbf{r}_{k,\tau} = \mathbf{r}_k(t-\tau)$ . In the aforementioned equation, considering  $\int_0^{T_e} \mathbf{p}_k^T(t) \partial \mathbf{F}_k / \partial \mathbf{r}_{k,\tau} \delta \mathbf{r}_{k,\tau} dt = \int_0^{T_e-\tau} \mathbf{p}_k^T(t+\tau) \partial \mathbf{F}_k(t+\tau) / \partial \mathbf{r}_k \delta \mathbf{r}_k dt$ , variation  $\delta J_L$  is rewritten as

$$\begin{aligned} \delta J_L &= - \sum_{k=1}^K \mathbf{p}_k^T(T_e) \delta \mathbf{r}_k(T_e) + \sum_{k=1}^K \int_{T_e-\tau}^{T_e} \left( \frac{\partial \Psi}{\partial \mathbf{z}_k} \boldsymbol{\omega}(t) + \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} + \dot{\mathbf{p}}_k^T(t) \right) \delta \mathbf{r}_k dt \\ &+ \sum_{k=1}^K \int_0^{T_e-\tau} \left( \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k} + \mathbf{p}_k^T(t+\tau) \frac{\partial \mathbf{F}_k}{\partial \mathbf{r}_k}(t+\tau) + \dot{\mathbf{p}}_k^T(t) \right) \delta \mathbf{r}_k dt + \sum_{k=1}^K \int_0^{T_e} \mathbf{p}_k^T(t) \frac{\partial \mathbf{F}_k}{\partial \mathbf{u}} \delta \mathbf{u} dt. \end{aligned} \quad (\text{B3})$$



When  $\mathbf{p}_k(t)$  is selected such that  $\mathbf{p}_k(T_e) = 0$  and the following equations are satisfied:

$$\frac{d\mathbf{p}_k^T(t)}{dt} = -\frac{\partial\Psi}{\partial\mathbf{z}_k}\boldsymbol{\omega}(t) - \mathbf{p}_k^T(t)\frac{\partial\mathbf{F}_k}{\partial\mathbf{r}_k}, \quad (\text{B4})$$

for  $T_e - \tau \leq t < T_e$ , and

$$\frac{d\mathbf{p}_k^T(t)}{dt} = -\mathbf{p}^T(t)\frac{\partial\mathbf{F}_k}{\partial\mathbf{r}_k} - \mathbf{p}_k^T(t+\tau)\frac{\partial\mathbf{F}_k(t+\tau)}{\partial\mathbf{r}_k}, \quad (\text{B5})$$

for  $0 \leq t < T_e - \tau$ , we can obtain a simple form of  $\delta J_L$ , as  $\sum_{k=1}^K \int_0^{T_e} \mathbf{p}_k^T(t) \partial\mathbf{F}_k / \partial\mathbf{u} d\mathbf{u} dt$ . When  $\delta\mathbf{u}$  is selected as

$$\delta\mathbf{u} = -\alpha \sum_{k=1}^K \left( \mathbf{p}_k^T(t) \frac{\partial\mathbf{F}_k}{\partial\mathbf{u}} \right)^T, \quad (\text{B6})$$

$\delta J_L = -1/\alpha \int \delta\mathbf{u}^2 dt \leq 0$  is always satisfied. Then, we consider variations  $\delta_{\omega} J_L$  and  $\delta_{\mathbf{b}} J_L$  in terms of the variations of weights and bias parameters, respectively. In the same manner shown earlier, we obtain,  $\delta_{\omega} J_L = \sum_{k=1}^K \partial\Psi / \partial\mathbf{z}_k \delta\mathbf{z}_k = \sum_{k=1}^K \partial\Psi / \partial\mathbf{z}_k \int_{T_e-\tau}^{T_e} \delta\boldsymbol{\omega} \mathbf{r}_k dt$  in terms of the weight variation  $\delta\boldsymbol{\omega}$  and  $\delta_{\mathbf{b}} J_L = \sum_{k=1}^K \partial\Psi / \partial\mathbf{z}_k \delta\mathbf{b}$  in terms of the weight variation  $\delta\mathbf{b}$ . Clearly, when  $\delta\boldsymbol{\omega}$  and  $\delta\mathbf{b}$  are set as follows:

$$\delta\boldsymbol{\omega} = -\alpha_{\omega} \sum_{k=1}^K \left( \frac{\partial\Psi}{\partial\mathbf{z}_k} \right)^T \mathbf{r}_k^T, \quad (\text{B7})$$

and

$$\delta\mathbf{b} = -\alpha_b \sum_{k=1}^K \left( \frac{\partial\Psi}{\partial\mathbf{z}_k} \right)^T, \quad (\text{B8})$$

with small positive constants,  $\alpha_{\omega}$  and  $\alpha_b$ ,  $\delta_{\omega} J_L \leq 0$  and  $\delta_{\mathbf{b}} J_L \leq 0$ .

### APPENDIX C: UPDATE EQUATIONS FOR $u_1$ , $u_2$ , $\omega_l$ , AND $b_l$ IN THE OPTOELECTRONIC DELAY SYSTEM

In this study, we set the loss function as  $J = -1/K \sum_{k=1}^K \sum_{l=0}^{L-1} t_{l,k} \log y_{l,k}$ , where  $y_{l,k} = e^{z_{l,k}} / \sum_{l=0}^{L-1} e^{z_{l,k}}$ , and  $z_{l,k} = \int_{T_e-\tau}^{T_e} \omega_l(t) \xi_k(t) dt + b_l$ . In this case, the adjoint equations for  $\mathbf{p}_k = (p_{\xi,k}, p_{\eta,k})^T$  are given as follows for

$T_e - \tau \leq t < T_e$ ,

$$\frac{dp_{\xi,k}}{dt} = -\frac{1}{K} \sum_{l=0}^{L-1} (y_{l,k} - t_{l,k}) \omega_l + g p_{\xi,k} - g_H p_{\eta,k}, \quad (\text{C1})$$

$$\frac{dp_{\eta,k}}{dt} = g_L p_{\xi,k}, \quad (\text{C2})$$

and for  $T_e < t \leq T_e - \tau$ ,

$$\begin{aligned} \frac{dp_{\xi,k}}{dt} &= g p_{\xi,k} - g_H p_{\eta,k} \\ &+ \tilde{\beta} u_1(t+\tau) \sin \delta_k(t+\tau) p_{\xi,k}(t+\tau), \end{aligned} \quad (\text{C3})$$

$$\frac{dp_{\eta,k}}{dt} = g_L p_{\xi,k}, \quad (\text{C4})$$

where  $g = 1/\tau_H + 1/\tau_L$ ,  $g_H = 1/\tau_H$ ,  $g_L = 1/\tau_L$ ,  $\tilde{\beta} = \beta/\tau_L$ , and  $\delta_k(t) = 2[u_1(t)\xi_k(t-\tau) + u_2(t)]$ . Then, the update variables of control signal vectors are given as follows:

$$\delta u_1(t) = \alpha_u \tilde{\beta} \sum_k^K p_{\xi,k} \sin \delta_k(t) \xi_k(t-\tau), \quad (\text{C5})$$

$$\delta u_2(t) = \alpha_u \tilde{\beta} \sum_k^K p_{\xi,k} \sin \delta_k(t), \quad (\text{C6})$$

and the update variables of weights and bias parameters are

$$\delta\omega_l = -\frac{\alpha_{\omega}}{K} \sum_k^K (y_{l,k} - t_{l,k}) \xi_k, \quad (\text{C7})$$

$$\delta b_l = -\frac{\alpha_b}{K} \sum_k^K (y_{l,k} - t_{l,k}). \quad (\text{C8})$$

- [1] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, Physics for neuromorphic computing, *Nat. Rev. Phys.* **2**, 499 (2020).
- [2] M. I. Rabinovich, P. Varona, A. I. Selverston, and H. D. I. Abarbanel, Dynamical principles in neuroscience, *Rev. Mod. Phys.* **78**, 1213 (2006).
- [3] S. Furber, Large-scale neuromorphic computing systems, *J. Neural Eng.* **13**, 051001 (2016).
- [4] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, *et al.*, A million spiking-neuron integrated circuit with a scalable communication network and interface, *Science* **345**, 668 (2014).
- [5] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, All-optical spiking neurosynaptic

- networks with self-learning capabilities, *Nature* **569**, 208 (2019).
- [6] A. N. Tait, T. Ferreira de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks, *Sci. Rep.* **7**, 7430 (2017).
- [7] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt, An experimental unification of reservoir computing methods, *Neural Netw.* **20**, 391 (2007).
- [8] H. Jaeger and H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, *Science* **304**, 78 (2004).
- [9] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, Recent advances in physical reservoir computing: A review, *Neural Netw.* **115**, 100 (2019).
- [10] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, Information processing using a single dynamical node as complex system, *Nat. Commun.* **2**, 468 (2011).
- [11] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, Parallel photonic information processing at gigabyte per second data rates using transient states, *Nat. Commun.* **4**, 1364 (2013).
- [12] M. Inubushi and S. Goto, Transfer learning for nonlinear dynamics and its application to fluid turbulence, *Phys. Rev. E* **105**, 043301 (2020).
- [13] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature* **521**, 436 (2015).
- [14] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, Scaling for edge inference of deep neural networks, *Nat. Electron.* **1**, 216 (2018).
- [15] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, *Science* **361**, 1004 (2018).
- [16] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Deep learning with coherent nanophotonic circuits, *Nat. Photonics* **11**, 441 (2017).
- [17] X. Chen and X. Lin, Big data deep learning: Challenges and perspectives, *IEEE Access* **2**, 514 (2014).
- [18] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, *Adv. Neural Inf. Process. Syst.* **29**, 3360 (2016).
- [19] F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, On the number of linear regions of deep neural networks, *Adv. Neural Inf. Process. Syst.* **27**, 2924 (2014).
- [20] G.-H. Liu and E. A. Theodorou, [arXiv:1908.10920v2](https://arxiv.org/abs/1908.10920v2) (2019).
- [21] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, *Adv. Neural Inf. Process. Syst.* **31**, 6572 (2018).
- [22] M. Benning, E. Celledoni, M. J Ehrhardt, B. Owren, and C.-B. Schönlieb, [arXiv:1904.05657](https://arxiv.org/abs/1904.05657) (2019).
- [23] E. Haber and L. Ruthotto, Stable architectures for deep neural networks, *Inverse Probl.* **34**, 014004 (2017).
- [24] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal Syst.* **2**, 303 (1989).
- [25] K. Funahashi, On the approximate realization of continuous mappings by neural networks, *Neural Netw.* **2**, 183 (1989).
- [26] S. Sonoda and N. Murata, Neural network with unbounded activation functions is universal approximator, *Appl. Comput. Harm. Anal.* **43**, 233 (2017).
- [27] D. E. Kirk, *Optimal Control Theory: An Introduction* (Dover Publications, Inc., Mineola, NY, 2004).
- [28] A. P. Sage and C. C. White, III, *Optimum Systems Control* (Prentice-Hall, Englewood Cliffs, New Jersey, 1977).
- [29] S. Ruder, [arXiv:1609.04747](https://arxiv.org/abs/1609.04747).
- [30] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2015).
- [31] Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, *Adv. Neural Inf. Process. Syst.* **27**, 2933 (2014).
- [32] F. T. Arecchi, G. Giacomelli, A. Lapucci, and R. Meucci, Two-dimensional representation of a delayed dynamical system, *Phys. Rev. A* **45**, R4225z(R) (1992).
- [33] A. Uchida, *Optical Communication with Chaotic Lasers* (Wiley-VCH, 2012).
- [34] M. C. Soriano, J. Garcia-Ojalvo, C. R. Mirasso, and I. Fischer, Complex photonic: Dynamics and applications of delay-coupled semiconductor lasers, *Rev. Mod. Phys.* **85**, 421 (2013).
- [35] T. E. Murphy, A. B. Cohen, B. Ravoori, K. R. B. Schmitt, A. V. Setty, F. Sorrentino, C. R. S. Williams, E. Ott, and R. Roy, Complex dynamics and synchronization of delayed-feedback nonlinear oscillators, *Phil. Trans. R. Soc. A* **368**, 343 (2010).
- [36] Y. LeCun, C. Cortes, and C. J. C. Burges, The MNIST database of handwritten digits, <http://yann.lecun.com/exdb/mnist>.
- [37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**, 2278 (1998).