

## Developing constrained neural units over time

Questa è la versione sottoposta a revisione paritaria (postprint) della seguente opera:

*Original*

Developing constrained neural units over time / Betti, Alessandro; Gori, Marco; Marullo, Simone; Melacci, Stefano. - (2020), pp. 1-8. ( IJCNN 2020 - International Joint Conference on Neural Networks Glasgow, UK 19-24/072020) [10.1109/IJCNN48605.2020.9207028].

*Availability:*

This version is available at: 20.500.11771/34879

*Publisher:*

*Published*

DOI:10.1109/IJCNN48605.2020.9207028

*Terms of use:*

This publication is made accessible in accordance with the terms for deposit in the institutional repository, as defined by the IMT School for Advanced Studies Lucca's Open Access Policy. ([https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib\\_0.pdf](https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib_0.pdf)).

Si prega di consultare le pagine informative dell'editore relative alle politiche di autoarchiviazione.

(Article begins on next page)

# Developing Constrained Neural Units Over Time

Alessandro Betti, Marco Gori, Simone Marullo, Stefano Melacci

*Department of Information Engineering and Mathematics*

*University of Siena*

Siena, Italy

alessandro.betti2@unisi.it, marco@diism.unisi.it, simone.marullo@student.unisi.it, mela@diism.unisi.it

**Abstract**—In this paper we present a foundational study on a constrained method that defines learning problems with Neural Networks in the context of the principle of least cognitive action, which very much resembles the principle of least action in mechanics. Starting from a general approach to enforce constraints into the dynamical laws of learning, this work focuses on an alternative way of defining Neural Networks, that is different from the majority of existing approaches. In particular, the structure of the neural architecture is defined by means of a special class of constraints that are extended also to the interaction with data, leading to “architectural” and “input-related” constraints, respectively. The proposed theory is cast into the time domain, in which data are presented to the network in an ordered manner, that makes this study an important step toward alternative ways of processing continuous streams of data with Neural Networks. The connection with the classic Backpropagation-based update rule of the weights of networks is discussed, showing that there are conditions under which our approach degenerates to Backpropagation. Moreover, the theory is experimentally evaluated on a simple problem that allows us to deeply study several aspects of the theory itself and to show the soundness of the model.

**Index Terms**—Cognitive Action Laws, Constrained Neural Networks, Learning Over Time

## I. INTRODUCTION

When a large amount of supervised data is available, (Deep) Neural Networks have shown to yield impressive results in several real-world tasks [1]–[3]. The classic formulation of supervised learning in Neural Networks consists in optimizing the values of the weights attached to a pre-designed neural architecture in order to fit the given training data under some regularity conditions or, in any case, to control the learning process to avoid overfitting and gain generalization skills [4]. Stochastic gradient descent [5] is commonly regarded as the *de facto* schema for the optimization of the network weights. According to this approach, at each time step only one sample (or a mini-batch of samples) is considered, and the training data are shuffled before the beginning of any training epoch, completely ignoring any information eventually available in the data ordering. Whenever the system is designed to learn from data as soon as they become available over time, we

move a step toward a real online learning setting. Neural Networks are also exploited in this setting, even if there are several other challenging issues than are not present in the static batch case [6], and stochastic gradient descent can still be applied, updating the network weights after having processed each newly received sample.

In this paper we follow the ideas behind the principle of least cognitive action which very much resembles the principle of least action in mechanics, and that was investigated in the context of Neural Networks and Computer Vision in [7], developing the so-called *Cognitive Action Laws of Learning*. Such learning framework naturally deals with learning problems in which the time component plays a crucial role, so that it is well-suited to approach in a principled way those cases in which data become available over time, that is the setting of this paper.

Starting from the seminal work of [8], optimization problems on the weights of Neural Networks can be equivalently formulated by extending the space of the learnable parameters, introducing a constrained optimization problem on the product space of weights and other neuron-related variables, as also investigated in more recent works [9]–[11]. This point of view allows us to describe the structure of the architecture of the network in terms of constraints involving such extended set of learnable parameters. Moreover, a constraint-based description has implications in the way the network operations can be parallelized and, potentially, in the way the structure of the network is progressively developed, thus offering a very generic perspective on which to build a foundational study.

In this paper we develop this idea by introducing the so-called *architectural constraints* among neurons, and also *input constraints* between each input neuron and the corresponding feature of the considered input example, thus fully breaking down the description of the network in terms of constraints. We show how these constraints can be applied within the context of the principle of least cognitive action to model a temporal evolution of the variables of the network. This approach allows us to formulate a learning problem that is intrinsically linked, from the beginning, to the idea of dealing with a temporal trajectory of the weights, as opposed to the classical approaches where the definition of the learning problem and the choice of the optimization algorithm consist of two separate and conceptually orthogonal aspects.

Our foundational study is developed using the formalism of calculus of variations with subsidiary conditions. We show

Accepted for publication at the IEEE International Joint Conference on Neural Networks (IJCNN) 2020 (DOI: TBA).

2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

how it is possible to formulate the theory using Lagrange multipliers and how those multipliers can be explicitly found by solving a linear system in as many unknowns as the number of neurons in the network. We also discuss how this theory, in a specific regime of the regularization parameters and in certain conditions of dissipation, exactly reproduces the Backpropagation rule for the computations of the gradients.

There are many reasons to pursue this *dynamical* approach to learning [7], [12], especially in those learning problems where the data comes as a (temporally) coherent signal. Dynamical constraints, such as invariance under motion [12], can be valuable in guiding the learning process, and the theory presented in this paper is developed as a building block to enable the implementation and the analysis of such temporal constraints for deep architectures (for example, in order to overcome the layer-wise training limitations of [12]).

This paper is organized as follows. Section II describes the proposed theoretical study, that is experimentally assessed focusing on a simple and easy-to-understand problem in Section III. Finally, Section IV concludes the paper with our ideas about future work.

## II. LEARNING FRAMEWORK

Our learning framework is rooted on the idea of describing a Neural Network by means of constraints among neural units. Before going into a formal definition and describing all the details, we illustrate the basic idea using a running example. Then, in Section II-A we will inject the network into the time domain, while the connection with Backpropagation is formalized in Section II-B.

The diagram on the right shows a feed-forward network with 2 input units (triangles), 2 hidden and 1 output neurons (circles). The graph of the architecture makes explicit the way in which the values associated with the input nodes (vertices 0 and 1) are propagated up until the output of the network (vertex 4). In particular, if we denote with  $x^i$  the output of the  $i$ -th neuron and with  $w_{ij}$  the weight associated with the arch  $j \rightarrow i$ , then the diagram in the figure implies  $x^2 = \sigma(w_{20}x^0 + w_{21}x^1)$ ,  $x^3 = \sigma(w_{30}x^0 + w_{31}x^1)$  and  $x^4 = \sigma(w_{42}x^2 + w_{43}x^3)$ , being  $\sigma$  the activation function. Therefore, in the  $(w, x)$  space these compositional relations between the nodes variables  $x^i$ ,  $i = 0, \dots, 4$  can be regarded as *architectural constraints*, namely  $G^2 = G^3 = G^4 = 0$ , where:

$$G^2 = x^2 - \sigma(w_{20}x^0 + w_{21}x^1), \quad G^3 = x^3 - \sigma(w_{30}x^0 + w_{31}x^1), \\ G^4 = x^4 - \sigma(w_{42}x^2 + w_{43}x^3).$$

Moreover the way in which input signals are provided to the network can be regarded as additional constraints. Let us suppose we want to compute the value of the network on the input  $x^0 = e^0$  and  $x^1 = e^1$ , where  $e^0$  and  $e^1$  are two scalar values; this two assignments can be interpreted as two additional *input constraints*  $G^0 = G^1 = 0$  where

$$G^0 = x^0 - e^0, \quad G^1 = x^1 - e^1.$$

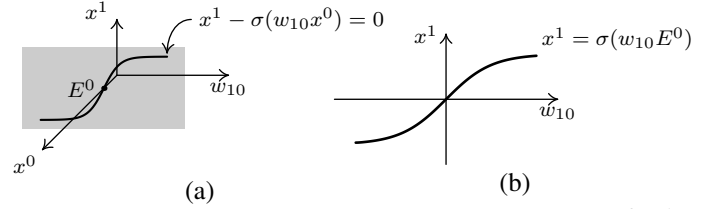


Fig. 1. Visualization of the neural constraints for the neural network  $\mathbf{0} \rightarrow \mathbf{1}$  (two neurons  $\{0, 1\}$ ). Constraint  $G^1(x, W) = 0$ , restricted to the plane  $x^0 = E^0$ , is shown in (a). In (b), such restriction is represented in the  $w_{10}$ - $x^1$  plane.

We can now formally describe the models that we consider in this paper, that in Section II-A will be used to formulate the learning problem in the context of the time domain. Given a simple digraph  $D = (N, A)$  of order  $\nu$ , where  $N$  is the set of vertices, while  $A$  is the set of edges. Without loss of generality, we can assume  $N = \{0, 1, \dots, \nu - 1\}$  and  $A \subseteq N \times N$ . A neural network associated with  $D$  consists of a set of maps  $i \in N \mapsto x^i \in \mathbf{R}$  and  $(i, j) \in A \mapsto w_{ij} \in \mathbf{R}$  together with  $\nu$  constraints  $G^j(x, W) = 0$ ,  $j = 0, 1, \dots, \nu - 1$  where  $W$  is the weight matrix (i.e.,  $(W)_{ij} = w_{ij}$ ) and  $x$  the vector of the outputs of the neurons. Let  $\mathcal{M}_\nu(\mathbf{R})$  be the set of all  $\nu \times \nu$  real matrices and  $\mathcal{M}_\nu^\downarrow(\mathbf{R})$  the set of all  $\nu \times \nu$  strictly lower triangular matrices over  $\mathbf{R}$ . In the rest of the paper we will always assume that  $W \in \mathcal{M}_\nu^\downarrow(\mathbf{R})$ , i.e., a directed acyclic graph. The relations  $G^j = 0$  for  $j = 0, \dots, \nu - 1$  specify the computational scheme with which the information diffuses through the network.

In a typical network with  $\omega$  inputs, the structure of these constraints are defined as follows (see also Fig. 1). For any vector  $\xi \in \mathbf{R}^\omega$ , for any matrix  $M \in \mathcal{M}_\nu(\mathbf{R})$  with entries  $m_{ij}$  and for any given  $\mathcal{C}^1$  map (i.e., a differentiable map whose derivative is continuous)  $e: [0, +\infty) \rightarrow \mathbf{R}^\omega$  we define the constraint function on neuron  $j$  when the example  $e(\tau)$  is presented to the network as

$$G^j(\tau, \xi, M) := \begin{cases} \xi^j - e^j(\tau), & \text{if } 0 \leq j < \omega; \\ \xi^j - \sigma(m_{jk}\xi^k) & \text{if } \omega \leq j < \nu, \end{cases} \quad (1)$$

where  $\sigma: \mathbf{R} \rightarrow \mathbf{R}$  is of class  $\mathcal{C}^2(\mathbf{R})$  (i.e., it is differentiable and its derivative is of class  $\mathcal{C}^1(\mathbf{R})$ ), and we used the Einstein notation (i.e., there is a sum on the index  $k$ ), that we will exploit throughout the whole paper to simplify the notation.<sup>1</sup> With this choice it is clear that the set of all constraints  $G^j(\tau, x, W) = 0$ ,  $j = 0, \dots, \nu - 1$  is the usual computational scheme of a feed-forward network with input  $e(\tau)$  whenever we choose  $W \in \mathcal{M}_\nu^\downarrow(\mathbf{R})$ . Throughout the paper we will use the notation  $G_\tau^j$ ,  $G_\xi^j$ ,  $G_M^j$  for the partial derivatives with respect to the first, second and third arguments of  $G^j$ , respectively. Notice that the dependence of the constraints on  $\tau$  reflects the fact that the computations of a neural network should be based on external inputs.

<sup>1</sup>Summation is intended over all repeated indices of an expression.

### A. Principle of Least Cognitive Action

Human cognitive processes do not emerge with a well-defined distinction between training and test set. As time goes by, humans react efficiently to new stimuli, which suggests us to look for alternative foundations of learning by embedding the agent into its own time-driven learning environment.

Following the ideas of [7], we establish a link with mechanics by paralleling the weights  $W$ , along with the neuronal outputs  $x$ , to the Lagrangian coordinates of a system of particles. We define the temporal trajectories of the variables of the learning problem by laws that come from stationarity conditions of a functional, as it happens for canonical coordinates in classical mechanics.

In particular, we indicate with  $t \in [0, +\infty)$  the time variable and we assume that the updates rules of the parameters are obtained from the stationarity conditions of the functional

$$\mathcal{A}(x, W) := \int \frac{1}{2}(m_x|\dot{x}(t)|^2 + m_W|\dot{W}(t)|^2) \varpi(t) dt + \mathcal{F}(x, W), \quad (2)$$

where  $m_x$  and  $m_W$  are positive scalars (also referred to as *masses*, in analogy with physics),  $\varpi(t)$  is a time-dependent positive weighing term, and  $\mathcal{F}(x, W) := \int F(t, x, \dot{x}, \ddot{x}, W, \dot{W}, \ddot{W}) dt$  (where  $F$  is related to the *potential energy* of the system). Notice that the kinetic energy, which appears in  $\mathcal{A}(x, W)$ , is a sort of temporal regularization term that, once minimized, leads to develop weights that settle to constant values. The functional in Eq. (2) is subject to the previously introduced constraints

$$G^j(t, x(t), W(t)) = 0, \quad 0 \leq j < \nu, \quad (3)$$

where the map  $G^j(\cdot, \cdot, \cdot)$  is taken as in Eq. (1). Notice that, in what follows, we will sometimes drop the dependency on the variables to simplify the notation.

Let  $(\begin{smallmatrix} G^\xi \\ G^M \end{smallmatrix})$  be the  $(\nu + \nu^2) \times \nu$  Jacobian matrix of the constraints  $G$  with respect to its second and third argument, respectively ( $\xi$  and  $M$ ):

$$\left( \begin{smallmatrix} G_\xi \\ G_M \end{smallmatrix} \right)_{ij} := \begin{cases} G_{\xi^i}^j & \text{for } 0 \leq i < \nu, \\ G_{\text{vec}(M)_{i-\nu}}^j & \text{for } \nu \leq i < \nu + \nu^2, \end{cases}$$

where, for any matrix  $M$ ,  $\text{vec}(M)$  is a vectorization of  $M$ . Variational problems with subsidiary conditions can be tackled using the method of Lagrange multipliers to convert the constrained problem into an unconstrained one (see [13]). In order to use this method, it is necessary to verify the independence hypothesis between the constraints; in this case we should check that the matrix  $(\begin{smallmatrix} G^\xi \\ G^M \end{smallmatrix})$  is full rank. Indeed the following proposition holds true:

**Proposition 1.** *The matrix  $(\begin{smallmatrix} G^\xi \\ G^M \end{smallmatrix}) \in \mathcal{M}_{(\nu^2+\nu) \times \nu}(\mathbf{R})$  is full rank.*

*Proof.* First of all notice that if  $(G_\xi)_{ij} = G_{\xi^i}^j$  is full rank also  $(\begin{smallmatrix} G^\xi \\ G^M \end{smallmatrix})$  has this property. Then, since

$$G_{\xi^i}^j(\tau, \xi, M) = \begin{cases} \delta_{ij}, & \text{if } 0 \leq j < \omega; \\ \delta_{ij} - \sigma'(m_{jk}\xi^k)m_{ji} & \text{if } \omega \leq j < \nu, \end{cases}$$

we immediately notice that  $G_{\xi^i}^i = 1$  and that for all  $i > j$  we have  $G_{\xi^i}^j = 0$ . This means that

$$(G_{\xi^i}^j(\tau, \xi, M)) = \begin{pmatrix} 1 & * & \cdots & * \\ 0 & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is clearly full rank.  $\square$

This result is sufficient for the existence of Lagrange multipliers  $\lambda_j(t)$  such that the (weak) extremals of (2) subject to constraints (3) are extremals of the following functional:

$$\begin{aligned} \mathcal{A}^*(x, W) &= \int \frac{1}{2}(m_x|\dot{x}(t)|^2 + m_W|\dot{W}(t)|^2) \varpi(t) dt \\ &\quad - \int \lambda_j(t) G^j(t, x(t), W(t)) dt + \mathcal{F}(x, W). \end{aligned} \quad (4)$$

The equations ruling the update of model parameters  $(x, W)$  come from the Euler-Lagrange equations for this variational problem:

$$\begin{aligned} -m_x \varpi(t) \ddot{x}(t) - m_x \dot{\varpi}(t) \dot{x}(t) - \lambda_j(t) G_\xi^j(x(t), W(t)) \\ + L_F^x(x(t), W(t)) &= 0; \\ -m_W \varpi(t) \ddot{W}(t) - m_W \dot{\varpi}(t) \dot{W}(t) - \lambda_j(t) G_M^j(x(t), W(t)) \\ + L_F^W(x(t), W(t)) &= 0, \end{aligned} \quad (5)$$

where  $L_F^x = F_x - d(F_{\dot{x}})/dt + d^2(F_{\ddot{x}})/dt^2$  and  $L_F^W = F_W - d(F_{\dot{W}})/dt + d^2(F_{\ddot{W}})/dt^2$  are the functional derivatives of  $F$  with respect to  $x$  and  $W$  respectively (see [13], [14]).

An expression to compute the Lagrange multipliers is derived by differentiating two times the constraints with respect to time and using the obtained quantities to substitute the second-order terms in the Euler-Lagrange equations:

$$\begin{aligned} \left( \frac{G_{\xi^a}^i G_{\xi^a}^j}{m_x} + \frac{G_{m_{ab}}^i G_{m_{ab}}^j}{m_W} \right) \lambda_j &= \varpi(G_{\tau\tau}^i + 2(G_{\tau\xi^a}^i \dot{x}^a \\ &\quad + G_{\tau m_{ab}}^i \dot{w}_{ab} + G_{\xi^a m_{bc}}^i \dot{x}^a \dot{w}_{bc}) \\ &\quad + G_{\xi^a \xi^b}^i \dot{x}^a \dot{x}^b + G_{m_{ab} m_{cd}}^i \dot{w}_{ab} \dot{w}_{cd}) \\ &\quad - \dot{\varpi}(\dot{x}^a G_{\xi^a}^i + \dot{w}_{ab} G_{m_{ab}}^i) \\ &\quad + \frac{L_F^x G_{\xi^a}^i}{m_x} + \frac{L_F^{w_{ab}} G_{m_{ab}}^i}{m_W}, \end{aligned} \quad (7)$$

where  $G_\tau^i$ ,  $G_{\tau\tau}^i$ ,  $G_{\xi^a}^i$ ,  $G_{\xi^a \xi^b}^i$ ,  $G_{m_{ab}}^i$  and  $G_{m_{ab} m_{cd}}^i$  are the gradients and the Hessians of constraint (3). In order to compute the multipliers  $\lambda_j$ , Eq. (7) must be solved. We conjecture that an approximate solution scheme can be implemented, by using an iterative procedure seeded with the multipliers computed at the previous time instant. However, in this paper we focus on the exact solution of Eq. (7).

**Initial conditions.** Now let us suppose that we want to design an iterative update algorithm for  $x$  and  $W$  over time,

starting from an arbitrary initial point and going on according to the learning theory so far stated. In this framework, the update algorithm comes straight from the numerical solution of the differential Eqs. (5)–(6) with Cauchy initial conditions. Clearly, we have to choose  $W(0)$  and  $x(0)$  such that  $g_i(0) \equiv 0$ , where we define  $g_i(t) := G^i(t, x(t), W(t))$ , for  $i = 0, \dots, \nu - 1$ . However, since the constraints must hold also for all  $t \geq 0$ , we also need  $g'_i(0) = 0$ . These conditions, written explicitly, mean:

$$G_\tau^i(0, x(0), W(0)) + G_{\xi^a}^i(0, x(0), W(0))\dot{x}^a(0) + G_{m_{ab}}^i(0, x(0), W(0))\dot{w}_{ab}(0) = 0.$$

If the constraints do not depend explicitly on time, it is sufficient to choose  $\dot{x}(0) = 0$  and  $\dot{W}(0) = 0$ , while for time dependent constraints this condition leaves:

$$G_\tau^i(0, x(0), W(0)) = 0,$$

which is an additional constraint on the initial conditions  $x(0)$  and  $W(0)$  to be satisfied. Therefore, one possible consistent way to impose Cauchy conditions is

$$\begin{aligned} G^i(0, x(0), W(0)) &= 0, \quad i = 0, \dots, \nu - 1; \\ G_\tau^i(0, x(0), W(0)) &= 0, \quad i = 0, \dots, \nu - 1; \\ \dot{x}(0) &= 0; \\ \dot{W}(0) &= 0. \end{aligned} \quad (8)$$

Once we fix  $x(0)$ ,  $W(0)$ ,  $\dot{x}(0)$  and  $\dot{W}(0)$  as above, higher derivatives of the variables  $x$  and  $W$  are determined by the differential equations (5)–(6).

### B. Reduction to Backpropagation

Let us consider a supervised problem described by the potential (a loss function)  $V(t, x(t))$ , which measures the error on the example presented at time  $t$  when the outputs of the neurons are  $x(t)$ ; in particular choose

$$F(t, x(t), \dot{x}(t), \ddot{x}(t), W(t), \dot{W}(t), \ddot{W}(t)) = -e^{\vartheta t} V(t, x(t)). \quad (9)$$

where  $\vartheta$  is a damping factor. We will now show that Eqs. (5)–(7) in the limit  $m_x \rightarrow 0$ ,  $m_W \rightarrow 0$ ,  $m_x/m_W \rightarrow 0$  reproduce first-order dynamics, where the updates of  $W$  are performed as prescribed by Backpropagation. In order to see this, choose  $\gamma = m_W \vartheta$  and multiply both sides of Eqs. (5)–(7) by  $\exp(-\vartheta t)$ , then take the limit  $m_x \rightarrow 0$ ,  $m_W \rightarrow 0$ ,  $m_x/m_W \rightarrow 0$ . In this limit, Eq. (6) and Eq. (7) become, respectively:

$$\dot{W}_{ij} = -\frac{1}{\gamma} \sigma'(w_{ik} x^k) \delta_i x^j; \quad (10)$$

$$G_{\xi^a}^i G_{\xi^a}^j \delta_j = -V_{x^a} G_{\xi^a}^i, \quad (11)$$

where  $\delta_j$  is the limit of  $\exp(-\vartheta t) \lambda_j$  and  $V_x$  is the partial derivative of  $V$  with respect to its second variable ( $x$ ). Because the matrix  $G_{\xi^a}^i$  is invertible Eq. (11) actually becomes

$$T\delta = -V_x, \quad (12)$$

where  $T_{ij} := G_{\xi^i}^j$ . This matrix is upper triangular, thus explicitly showing the backward structure of the propagation of the delta error of the Backpropagation algorithm. In supervised problems,  $V$  depends only on the output value of the network. Hence, whenever  $i$  is not an output neuron, computing  $\delta_i$  by solving Eq. (12) corresponds to:

$$\delta_i = \sigma'(w_{jk} x^k) w_{ji} \delta_j.$$

On the other hand, regarding the output units, Eq. (12) enforces the initialization of the  $\delta$ s based on the value of the supervision:

$$\delta_i = -V_{x^i}.$$

It is clear from these remarks that in Eq. (10) the Lagrange multiplier-related term  $\delta$  plays the role of the delta error of Backpropagation.

### III. EXPERIMENTAL ANALYSIS

We investigated an easy-to-understand task in order to provide an experimental assessment of the proposed theory. The task consists in classifying 2-dimensional data according to the well-known non-linearly separable XOR Boolean function. However, this task is defined in the time domain, so that data is provided to the system in an online fashion, according to a function that basically models a trajectory in the input space, passing through the vertices of the Boolean hypercube. In particular, samples are collected along a circular trajectory, as shown in Fig. 2. At each time instant, a sample is provided to the system. Differently, supervision is provided to the system only when the coordinates  $(e^0, e^1)$  of the input sample belong to small circular regions ( $R = 0.2$ ) centered around the vertices of the hypercube; in this case, the target of the nearest vertex is considered. In detail, the equations of the trajectory are

$$\begin{aligned} e^0(t) &= \phi(t)(u^0 + R \cos(\Omega t)) \\ e^1(t) &= \phi(t)(v^0 + R \sin(\Omega t)) \\ \phi(t) &= \frac{1}{2} \left( 1 + \tanh\left(\frac{t-a}{b}\right) \right) \end{aligned}$$

where  $u^0 = 0.5$ ,  $v^0 = 0.5$ ,  $a = 6$ ,  $b = 0.9$ ,  $\Omega = 0.25$ . The smoothing function  $\phi$  is used to get a null derivative of the signals at  $t = 0$ .

We used a model similar to the one that we introduced as running example in Section II, i.e. a one-hidden-layer feed-forward neural network with hyperbolic tangent as activation function  $\sigma$  and biases, as shown in Fig. 3. We used a potential (loss) term that is structured as Eq. (9) and the loss function  $V$  is the classic Mean Squared Error (MSE). Model accuracy and the loss value are computed either on the Boolean hypercube corners or on a set of points sampled from uniform distributions over the supervision regions. In particular, we used the notation *Acc*, *Loss* to indicate the former case, while *Acc2*, *Loss2* are about the latter. Each experiment consists of 10 runs with different random initializations of  $W \sim U(-\sqrt{2}, \sqrt{2})$ , being  $U$  the uniform distribution. Results are averaged over the runs ( $\pm$  standard deviation).

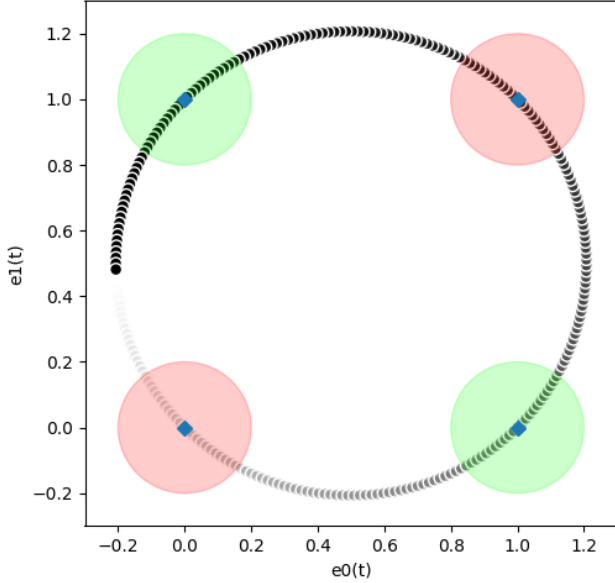


Fig. 2. Trajectory along which samples are fed to the system. When we sample from within the red/green regions, a supervision is also provided. The blue diamonds are the vertices of the Boolean hypercube.

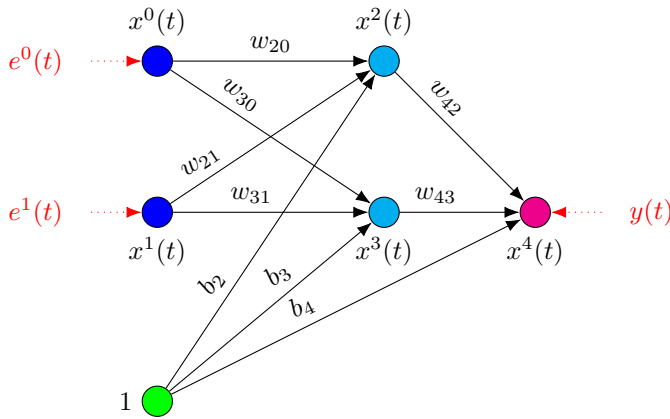


Fig. 3. Neural architecture:  $e^0(t)$  and  $e^1(t)$  are the input signals, while  $y(t)$  is the supervision signal.

We compare the proposed approach with a *baseline* system that exploits stochastic gradient descent to optimize the loss function  $V$  only, operating in the same online learning setting described above. The considered task generates samples that have a very high temporal correlation, making the task not trivial. Table I reports the results of the baseline model, as a function of the choice of the learning rate  $\eta$ . It is interesting to notice that such model fails to correctly classify the data in most of the cases, being it extremely sensitive to the choice of  $\eta$ , as expected.

The proposed learning approach involves solving differential equations of a *second-order model* (see Section II-A,

Eq. 5, Eq. 6). We used the LSODA solver, available in the ODEPACK Fortran library. The algorithm, originally proposed in [15], implements Adams and BDF methods with automatic stiffness detection and switching. We set  $\Delta t = 0.1$  and error control performed by the solver is determined by  $rtol = atol = 1.49 \cdot 10^{-8}$ . We also consider a *first-order model* that is implemented by assuming to have reached the limit conditions of (Section II-B). On the other hand, for 1st-order equations (Section II-B) we used the Euler method, in order to reproduce the small weight updates typical of gradient descent methods. In all the following experiments, we ensured that the learning rate of the baseline model,  $\eta$ , is chosen coherently with the scaling factors of Eq. 10 and the step-size  $\Delta t$ . In particular, we set  $\eta = \Delta t / \gamma = \Delta t / (m_W \vartheta)$ .

TABLE I  
PERFORMANCE OF THE BASELINE MODEL AS A FUNCTION OF THE LEARNING RATE  $\eta$ .

$\eta$	Acc	Loss	Acc2	Loss2
0.001	0.650 $\pm$ 0.122	0.906 $\pm$ 0.133	0.651 $\pm$ 0.117	0.911 $\pm$ 0.132
0.0025	0.700 $\pm$ 0.150	0.759 $\pm$ 0.209	0.697 $\pm$ 0.147	0.767 $\pm$ 0.205
0.01	0.825 $\pm$ 0.115	0.475 $\pm$ 0.283	0.823 $\pm$ 0.116	0.491 $\pm$ 0.278
0.03	0.875 $\pm$ 0.125	0.398 $\pm$ 0.388	0.873 $\pm$ 0.128	0.410 $\pm$ 0.392
0.1	0.900 $\pm$ 0.166	0.305 $\pm$ 0.472	0.902 $\pm$ 0.161	0.313 $\pm$ 0.476
0.125	0.900 $\pm$ 0.166	0.315 $\pm$ 0.492	0.901 $\pm$ 0.163	0.321 $\pm$ 0.496
0.3	0.900 $\pm$ 0.166	0.349 $\pm$ 0.560	0.900 $\pm$ 0.166	0.353 $\pm$ 0.562
0.5	0.900 $\pm$ 0.166	0.369 $\pm$ 0.601	0.900 $\pm$ 0.166	0.371 $\pm$ 0.602
0.8	0.950 $\pm$ 0.100	0.197 $\pm$ 0.393	0.950 $\pm$ 0.101	0.199 $\pm$ 0.397
1	0.975 $\pm$ 0.075	0.096 $\pm$ 0.288	0.974 $\pm$ 0.076	0.102 $\pm$ 0.293
2	0.550 $\pm$ 0.100	1.800 $\pm$ 0.400	0.550 $\pm$ 0.099	1.802 $\pm$ 0.395
3	0.525 $\pm$ 0.075	1.900 $\pm$ 0.300	0.524 $\pm$ 0.071	1.904 $\pm$ 0.287

#### A. Comparison between First-Order Model and Baseline

In Section II-B we discussed the relationships between a special instance of our model, that is the first-order case, and Backpropagation.

We compared the trajectories of the weights and biases comparing such model with the baseline case, setting  $\eta = \Delta t / \gamma = 0.3$  using the previously described criterion, and reporting them in Fig. 4. For each parameter (i.e., weights and biases) we have two trajectories (same color, different style), one from our model, one from the baseline system. It is interesting to see that the two dynamics converge to almost identical values (differences are due to round-off errors), confirming the correctness of the results of Section II-B.

#### B. Comparison between Second and First-Order Model

Section II-B suggests that the second-order model can be moved very close to a first-order model when using small masses and strong damping. We compared the weights (and biases) dynamics of the second-order model, under the aforementioned conditions, with the first-order model. In particular, we set  $\vartheta = 333.3$ ,  $m_x = 1.0 \cdot 10^{-5}$ ,  $m_W = 1.0 \cdot 10^{-3}$  and  $\gamma = \vartheta \cdot m_W$ . Fig. 5 (that follows the same organization of Fig. 4) shows that the two algorithms exhibit almost identical behaviour.

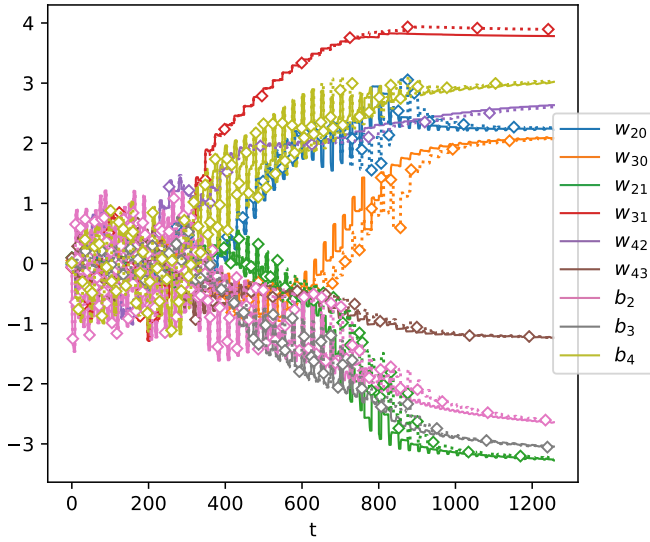


Fig. 4. Comparison between the dynamics of the weights and biases of the first-order model (lines) and the baseline system (dotted lines with markers). The value of the weights (or biases) is plotted against time. Trajectories with the same colors are about the same parameter. The two dynamics converge to almost identical values, indistinguishable in the first time instants (differences for larger  $t$  are due to numerical errors).

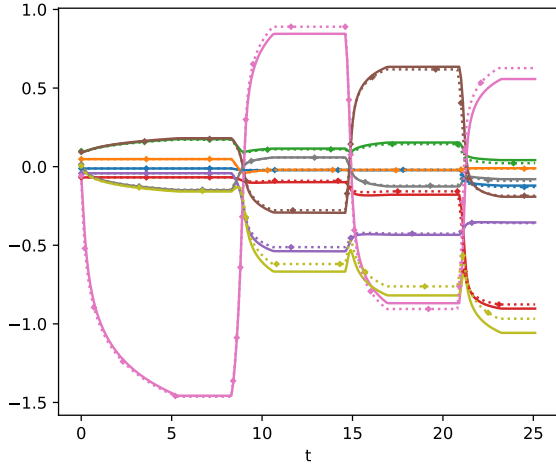


Fig. 5. Comparison between the dynamics of the weights and biases of the second-order model (lines) and the first-order one (dotted lines with markers). The value of the weights (or biases) is plotted against time. Trajectories with the same colors are about the same parameter (refer to legend of Fig. 4). The two algorithms exhibit almost identical behavior.

Notice that to obtain equivalent differential problem formulations it is required to have coherent initial conditions. In the case of the second-order model, the initial value of derivatives of variables is arbitrary, while in the first-order model it is given by the differential equation itself. An easy way to ensure consistency between the formulations is to multiply, in both the models, the potential term  $V$ , with a weighing function  $(1 - e^{-\vartheta t})$  that, under the assumption of strong damping, gets to 1 very quickly. Moreover, with this choice we obtain  $\dot{W}(0) = 0$ , which implies coherence of initial conditions of the two problems and easier convergence of the solution of

the second-order problem to the first-order one. The latter is essentially a continuous-time online learning gradient descent algorithm (as the baseline system), with the only difference being the time-dependent factor  $(1 - e^{-\vartheta t})$ . When altering the potential term as suggested, we get a stronger convergence of the derivatives at the left boundary. On the other hand, the overall slight difference between the weights trajectories is motivated by the non-zero inertial properties of the second-order formulation.

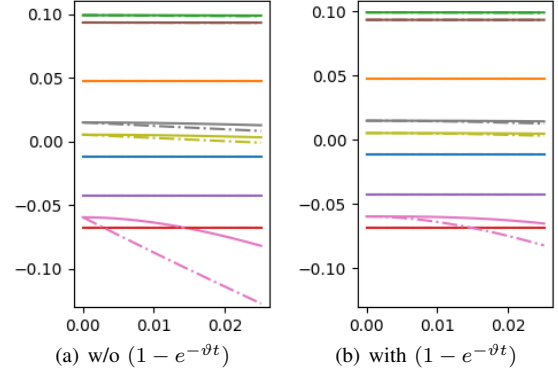


Fig. 6. Comparison between the dynamics of the weights and biases of the second-order model (lines) and the first-order one (dotted lines with markers). The value of the weights (or biases) is plotted against time. Trajectories with the same colors are about the same parameter (refer to legend of Fig. 4). (a) Without altering the potential term with the weighing function  $(1 - e^{-\vartheta t})$ , (b) using the altered potential term. Notice that (b) shows a stronger coherence close to the boundary  $t = 0$ .

### C. Comparison between Second-Order Model and Baseline

We performed an extended comparison between the proposed second-order model and the baseline system. We considered different sets of values for the key parameters of our theory, that are the masses  $m_x, m_W$  and the damping term  $\vartheta$ . The learning rate  $\eta$  of the baseline model is computed as a function of such parameters:  $\eta = \Delta t / \gamma = \Delta t / (m_W \vartheta)$  (further details in the introduction of Section III).

Results are reported in Tab. II. When small masses and large damping coefficients are used, the compared systems perform in a very similar way (rows 1, 2, 4). Otherwise, the behaviours depart one from each other, since the dynamics of the second-order model are clearly more structured than in the vanilla baseline. We also measured the norms of the constraining functions involving the neural units at the last time instant  $T$

$$|g(T)| = \left( \sum_{j=0}^{\nu-1} \left( G^j(T, x(T), W(T)) \right)^2 \right)^{1/2},$$

and we report them in the fifth column of Tab. II. Overall, constraints are fulfilled in all the evaluated settings, confirming the soundness of the proposed update scheme for the Lagrange multipliers of Eq. (7). Interestingly, the differences between  $Acc$  and  $Acc2$  (and also  $Loss$  and  $Loss2$ ) are minimal, suggesting that the classifier is keeping some margin among

TABLE II

COMPARISON OF THE SECOND-ORDER MODEL WITH THE BASELINE SYSTEM, WITH DIFFERENT CONFIGURATIONS OF THE MODEL PARAMETERS. ROWS 1,2,4 ARE ASSOCIATED WITH SMALL MASSES AND STRONG DAMPING. BASELINE RESULTS ARE TAKEN FROM TAB. I.

$\eta$	$m_x$	$m_W$	$\vartheta$	SECOND-ORDER MODEL					BASELINE	
				$ g(T) $	Acc	Loss	Acc2	Loss2	Acc	Loss
0.3	$1.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	33.3	$0.030 \pm 0.007$	$0.900 \pm 0.166$	$0.348 \pm 0.557$	$0.899 \pm 0.167$	$0.354 \pm 0.557$	$0.900 \pm 0.166$	$0.349 \pm 0.560$
0.03	$1.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	333	$0.034 \pm 0.003$	$0.850 \pm 0.166$	$0.436 \pm 0.436$	$0.848 \pm 0.167$	$0.450 \pm 0.436$	$0.875 \pm 0.125$	$0.398 \pm 0.388$
0.3	$1.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-1}$	0.667	$0.030 \pm 0.007$	$0.625 \pm 0.125$	$1.499 \pm 0.499$	$0.624 \pm 0.124$	$1.504 \pm 0.494$	$0.900 \pm 0.166$	$0.349 \pm 0.560$
0.125	$1.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	80.0	$0.031 \pm 0.002$	$0.850 \pm 0.200$	$0.466 \pm 0.591$	$0.851 \pm 0.199$	$0.475 \pm 0.590$	$0.900 \pm 0.166$	$0.315 \pm 0.492$
0.8	$1.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	12.5	$0.031 \pm 0.004$	$0.950 \pm 0.100$	$0.197 \pm 0.393$	$0.950 \pm 0.100$	$0.200 \pm 0.394$	$0.950 \pm 0.100$	$0.197 \pm 0.393$
0.0025	1.0	5.0	8.0	$0.030 \pm 0.001$	$0.450 \pm 0.187$	$1.156 \pm 0.231$	$0.467 \pm 0.179$	$1.150 \pm 0.227$	$0.700 \pm 0.150$	$0.759 \pm 0.209$

the vertices of the Boolean hypercube. We further investigated this aspect by visualizing the decision boundaries developed by the model, reported in Fig. 7 (related to the setting of row 5 in Tab. II), which confirms the previous considerations. The whole supervision-related areas are correctly classified, even if the sampling trajectory only intercepts a small portion of such areas, i.e. the arcs of circumference contained within the supervision regions. For completeness, we also report the same picture in the case of the baseline model, Fig. 8, that leads to a similar result in terms of margin (we randomly selected one of the 10 runs).

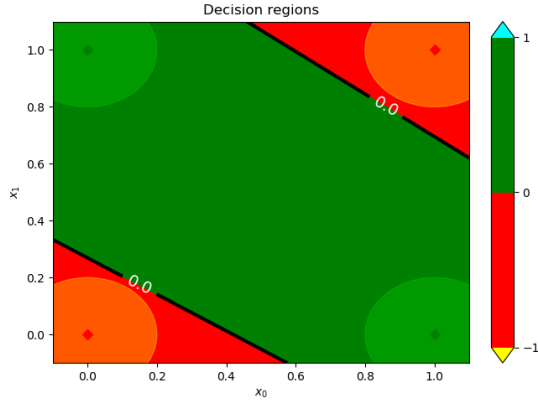


Fig. 7. Second-order model. Decision regions of the network resulting from one of the runs using parameters in row 5 of Tab. II (random initialization).

We analyzed the weight dynamics in three configurations that we selected in order to move from (a) a uniformly weighed case ( $m_x = m_W = \vartheta = 1$ ), (b) an intermediate setting ( $m_x = 0.001$ ,  $m_W = 0.1$ ,  $\vartheta = 5$ ) and (c) the already mentioned small-masses-large-damping case ( $m_x = 0.0001$ ,  $m_W = 0.001$ ,  $\vartheta = 33$ ). Fig. 9, top-row, show the dynamics of the second-order model, while Fig. 9, bottom-row, is about the baseline model (having only parameter  $\eta$ , chosen as previously mentioned in order to produce comparable behavior). Oscillations are significant (case (a)), unless damping is remarkably high (case (c)). As expected, when inertial properties of the variables are not negligible, the dynamics are smoother and less peaky with respect to signal variations (slower behavior).

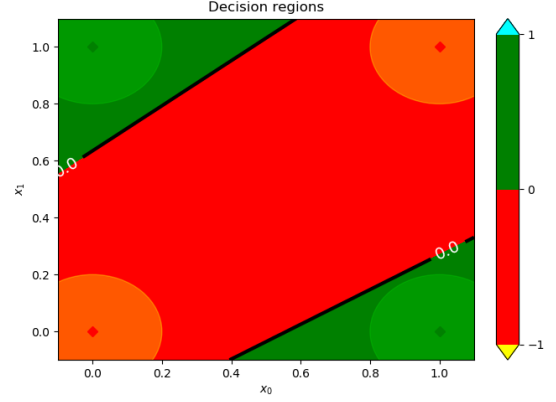


Fig. 8. Baseline model. Decision regions of the network resulting from one of the runs using parameters in row 5 of Tab. II (random initialization).

#### IV. CONCLUSIONS AND FUTURE WORK

We presented a theoretical study of a generic framework that describes the structure of a Neural Network by using constraints among the neural units, and that we injected into the time domain. When data are gradually presented to the system in an online fashion, the proposed framework allowed us to devise the trajectory of the values of the weights that is optimal with respect to the Least Action Principle. We described the connection of the ideas of this paper and Backpropagation, showing that the latter can be obtained by some choices on the parameters of our model, as confirmed by an experimental analysis in which different aspects of the theory were evaluated. We plan to exploit the outcome of this work in order to handle those dynamic constraints that enforce coherence over time, such as motion coherence [12], coherence on predictions over groups data points [16] or on space regions [17], [18].

#### ACKNOWLEDGMENT

This work was partly supported by the PRIN 2017 project RexLearn, funded by the Italian Ministry of Education, University and Research (grant no. 2017TWNMH2).

#### REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

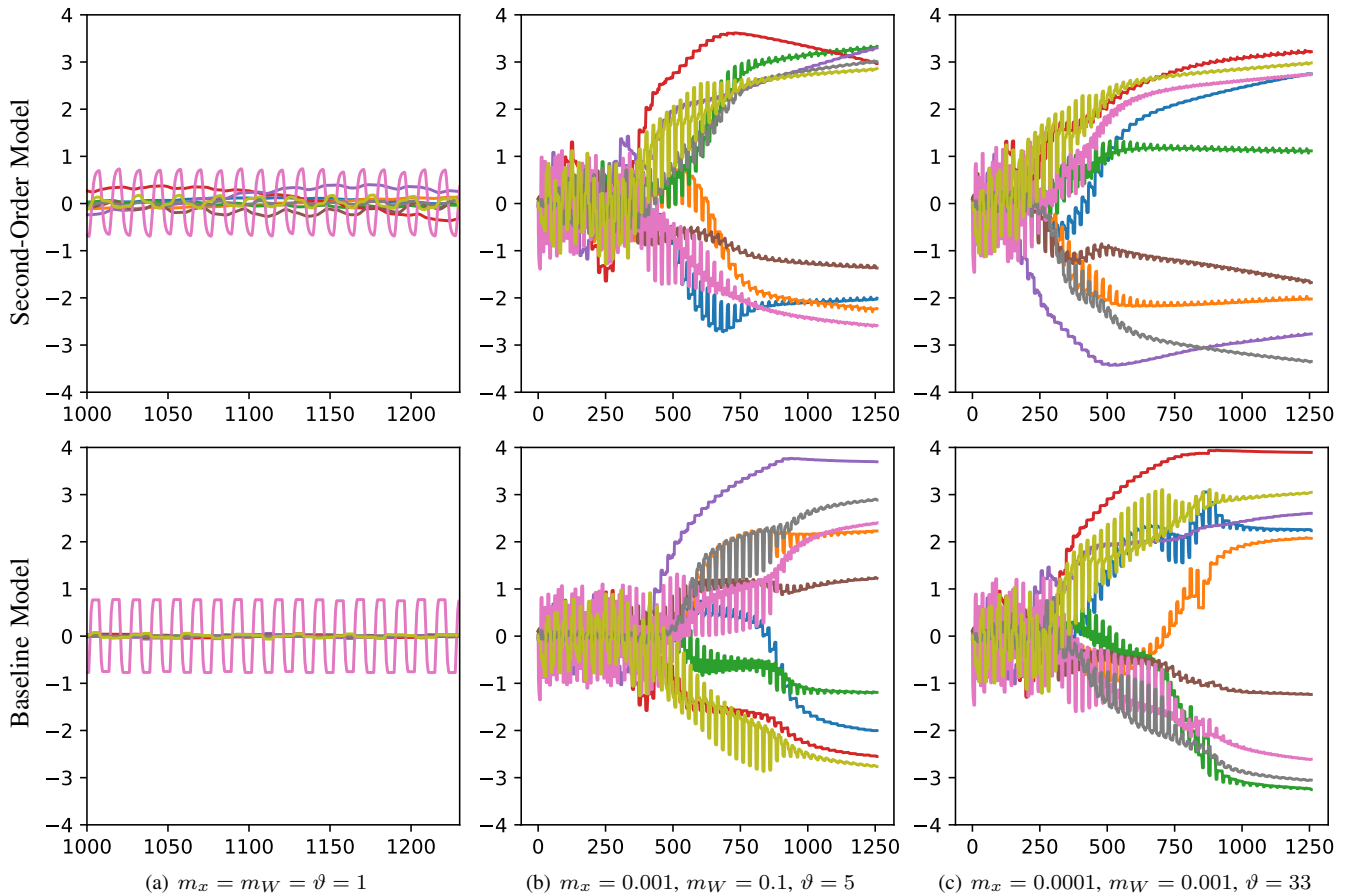


Fig. 9. Evolution of the weights (and biases) for different configurations of the parameters (one for each column, (a),(b),(c)). Each curve shows the value (ordinate) of a network parameter versus time (abscissa). Refer to legend of Fig. 4. Top row: second-order model. Bottom row: baseline model (the learning rate  $\eta$  is a function of  $m_W, \vartheta$  - see the text of the paper for further details).

- [2] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.
- [3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.
- [4] M. Gori, *Machine Learning: A constraint-based approach*. Morgan Kaufmann, 2017.
- [5] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436.
- [6] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, 2019.
- [7] A. Betti, M. Gori, and S. Melacci, "Cognitive action laws: The case of visual features," *IEEE transactions on neural networks and learning systems*, 2019.
- [8] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, "A theoretical framework for back-propagation," in *Proceedings of the 1988 connectionist models summer school*, vol. 1. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988, pp. 21–28.
- [9] M. Carreira-Perpinan and W. Wang, "Distributed optimization of deeply nested systems," in *Artificial Intelligence and Statistics*, 2014, pp. 10–19.
- [10] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein, "Training neural networks without gradients: A scalable admm approach," in *International conference on machine learning*, 2016, pp. 2722–2731.
- [11] A. Gotmare, V. Thomas, J. Brea, and M. Jaggi, "Decoupling back-propagation using constrained optimization methods," in *Workshop on Efficient Credit Assignment in Deep Learning and Deep Reinforcement Learning, ICML 2018.*, 2018, pp. 1–11.
- [12] A. Betti, M. Gori, and S. Melacci, "Motion invariance in visual environments," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 2009–2015. [Online]. Available: <https://doi.org/10.24963/ijcai.2019/278>
- [13] M. Giaquinta and S. Hildebrandt, *Calculus of variations I*. Springer Science & Business Media, 2004, vol. 310.
- [14] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. Interscience Publ., New York and London, 1962, vol. 1, p. 190.
- [15] L. Petzold, "Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations," *SIAM Journal on Scientific and Statistical Computing*, vol. 4, no. 1, pp. 136–148, 1983. [Online]. Available: <https://doi.org/10.1137/0904010>
- [16] S. Melacci, L. Sarti, M. Maggini, and M. Bianchini, "A neural network approach to similarity learning," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2008, pp. 133–136.
- [17] S. Melacci and M. Gori, "Learning with box kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2680–2692, 2013.
- [18] G. Gnecco, M. Gori, S. Melacci, and M. Sanguineti, "A theoretical framework for supervised learning from regions," *Neurocomputing*, vol. 129, pp. 25–32, 2014.