mathematics

MDPI

Article

A Formal Approach to Optimally Configure a Fully Connected Multilayer Hybrid Neural Network

Goutam Chakraborty [1,2,*,†], Vadim Azhmyakov [3,*,†] and Luz Adriana Guzman Trujillo [4,†]

[1] Department of Software & Information Science, Iwate Prefectural University, Iwate Ken, Takizawa 020-0693, Japan
[2] Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle 517325, A.P., India
[3] Department of Mathematics, Madanapalle Institute of Technology & Science, Madanapalle 517325, A.P., India
[4] LARIS, University of Angers, 49000 Angers, France; luzadriguz@gmail.com
* Correspondence: goutam@iwate-pu.ac.jp (G.C.); drvadimazhmyakov@mits.ac.in (V.A.)
† These authors contributed equally to this work.

Abstract: This paper is devoted to a novel formal analysis, optimizing the learning models for feedforward multilayer neural networks with hybrid structures. The proposed mathematical description replicates a specific switched-type optimal control problem (OCP). We have developed an equivalent, optimal control-based formulation of the given problem of training a hybrid feedforward multilayer neural network, to train the target mapping function constrained by the training samples. This novel formal approach makes it possible to apply some well-established optimal control techniques to design a versatile type of full connection neural networks. We next discuss the irrelevance of the necessity of Pontryagin-type optimality conditions for the construction of the obtained switched-type OCP. This fact motivated us to consider the so-called direct-solution approaches to the switched OCPs, which can be associated with the learning of hybrid neural networks. Concretely, we consider the generalized reduced-gradient algorithm in the framework of the auxiliary switched OCP.

Keywords: feedforward network; hybrid activation layers; deep network; discrete dynamic systems; switched optimal control

MSC: 49N99; 90C90; 93C10

check for updates

Academic Editors: Ioannis Tsoulos and Quanxin Zhu

Received: 11 November 2024
Revised: 16 December 2024
Accepted: 25 December 2024
Published: 31 December 2024

Citation: Chakraborty, G.; Azhmyakov, V.; Guzman Trujillo, L.A. A Formal Approach to Optimally Configure a Fully Connected Multilayer Hybrid Neural Network. *Mathematics* **2025**, *13*, 129. https://doi.org/10.3390/math13010129

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Although it has been almost forty years since the multilayer perceptron (MLP) was proposed [1], it is still widely used as a full-connection classifier along with deep convolution neural networks. MLPs constitute fully-connected feedforward neural networks. It was proven that MLPs are universal approximators [2,3]. The importance of MLPs can never be overstated, since they are key classifier models due to their expressive power guaranteed by the universal approximation theorem [3]. However, the important issue is the generalization property of a well-trained network [4]. During the 1990s, balancing between the bias and variance of the network, which is achieved by appropriate choice of the number of layers and the number of nodes, was an important research topic issue (see, e.g., [5] and references therein). This issue was mitigated by regularizing the connection weights.

However, a formal method to address the problem of an optimal configuration of the network architecture, to determine the number of hidden layers and their constituent number of units, was never properly addressed. A generic MLP consists of computational

layers, e.g., C_1, C_2, $C_3, \ldots, C_t, \ldots, C_T$, with the same computational model at every layer, but with different number of hidden units. The connection weights between layers are the parameters of the model. A hybrid MLP with trainable parameters involves layers with different activation functions. Except for the general convention of using a larger number of computing nodes for layers closer to the input, there is no hypothesis on how the layers with different activation functions should be organized in order to obtain an optimal overall performance. Of course, various optimization criteria can be defined. Two basic optimization criteria are a faster convergence and better generalization of the mapping function.

For our analysis, considering layers as C_1, C_2, $C_3, \ldots, C_t, \ldots, C_T$, we assume that the number of layers T is fixed, and that every layer has an equal number of units, also called nodes. A node can be disabled (ineffective) by setting all its connecting weights as zeros. We can drop a layer by making it a identity function, for which the output is the same as the input. If we start with a sufficiently large T, we can always converge to the optimal neural network using our algorithm. We always start with the generic model, and update the network parameters to find an optimal configuration for the hybrid feedforward network for the task.

For a concrete solution to optimize the neural network configuration, i.e., the network architecture, we will use a specific optimal control methodology that was initially developed for the optimal design of a switched control system. A conventional MLP neural network can be interpreted as a discrete-time dynamic system [6,7] (and references therein). Despite the impressive results seen in the real-world applications, some basic formal aspects of the analysis of modern feedforward deep network models remain unsolved. This fact motivates the development of possible equivalent interpretations of generic neural network dynamics in the form of some known mathematical abstractions. This equivalent representation of the hybrid MLP neural network in the form of a discrete-time control system makes it possible to apply the switched optimal control methodology to the main optimal structuring problem mentioned above.

Generally, the optimal control approach is not adequate to optimize feedforwardrd deep neural networks, or their variants as in [8]. In this paper, we reduce the optimal configuration problem for a hybrid multilayer neural network model to a specific switched-type OCP. We then discuss some critical aspect of the Pontryagin optimality principle in the framework of the resulting OCP and develop a reduced-gradient method for this problem.

The novelty of the proposed model is its hybrid computational architecture. For the proposed architecture, all layers are fully connected. At different layers, the activation functions could be different. They are automatically chosen from a set of $Q \leq T$ possibilities to optimize the objective fuction defined in Equation (6). T is the number of layers set initially, and is fixed. Refs. [6,7] proposed training neural networks using tools from optimum control. In their proposals, all layers use the same computational model. This greatly limits the flexibility of the model to optimize. In the proposed hybrid network presented herein, although the number of layers is set to a fixed T, the computation q at different layers automatically converges to optimum, choosing the best one from a variety of Q possibilities. For example, the proposed model has the ability to converge to an optimum number of computational layers by forcing transfer functions of redundant layers to identity functions. The model could achieve better configuration due to greater flexibility, facilitated by the combination of hybrid computation adapted at different layers. In addition, the objective function for optimization (Equation (6)) is designed to regularize the connection weights to ensure better generalization.

The remainder of this paper is organized as follows: Section 2 contains a formal mathematical model of a hybrid MLP neural network and includes its equivalent representation

in the form of an OCP. In Section 3, we study the obtained MLP network dynamics in the context of general switched control systems and consider the corresponding switched-type OCP. The proposed approach involves a constructive description of the permutation of layers in the given hybrid MLP network. Section 4 includes a critical discussion of the applicability of the generic Pontryagin-type optimality conditions to a practical implementation of the results of OCP. Motivated by the above critical idea, in Section 5, we develop an alternative approach to the switched OCP solution associated with the hybrid network optimization problem, i.e., the original optimal configuration problem for a hybrid MLP neural network. This approach involves a specific variant of the conventional reduced-gradient algorithm associated with the switched OCP. Section 6 summarizes the paper with indications to future extensions of the idea.

2. A Discrete-Time Optimal Control Framework for the Feedforward Hybrid Deep Learning Networks

In this section, we develop a mathematically rigorous OCP-based formalism for the optimal training problem associated with a hybrid fully connected neural network. By hybrid, we mean that the different layers can use different activation functions. This novel system-theoretic formalization provides an equivalent modelling approach for the main optimal configuration problem for a hybrid MLP neural network.

Recall that the training of a fully connected feedforward neural network is often considered in the a nonlinear programming framework [7]. In this case, one solves a suitable constrained or unconstrained optimization problem [9]. In supervised learning, the key idea is to find a classifier by estimating function (model) parameters of a known parametric function constrained by a supervised dataset. The formal procedure of finding the optimal parameters in the above setting is usually considered as a generalized regression problem.

An artificial neural network consists of processing units (also called nodes or neurons) that can be described in the framework of system inputs and outputs. Therefore, a suitable formal description of a multilayered neural network naturally leads to an equivalent discrete-time dynamic system, where training imparts dynamic behaviour, and its convergence to an optimal state is the goal. Consequently, the main nonlinear optimization problem of training a feedforward deep neural network can also be studied in the context of a specific constrained OCP, though the convergence mechanisms are different. Neural network trainings are gradient descent algorithms, where the rate of convergence is controlled by "learning rate". In OCP, the convergence, which is controlled by changing the discrete step size, is faster.

Let $T \in \mathbb{N}$ be the total number of layers in the given hybrid neural network and $S \in \mathbb{N}$ be the maximal sample size of the dataset. Given a collection of hybrid network layers, with the inputs being

$$Y_0 := \{y_{s,0} \in \mathbb{R}^{N_0},\ s = 1, \ldots, S\},$$

we obtain the generated network outputs, either real (as a regression problem) or integer (as a classification problem), according to the supervised data, as

$$Y_T := \{y_{s,T} \in \mathbb{R}^{N_T},\ s = 1, \ldots, S\}$$

by the following composition of hybrid-type transfer operators:

$$Y_T = (\phi_T^{q_I} \circ \phi_{T-1}^{q_{I-1}} \circ \ldots \circ \phi_1^{q_1})(Y_0, \theta^1). \tag{1}$$

Here,

$$\phi_{\tau+1}^{q} \circ \phi_{\tau}^{q'} := \phi_{\tau+1}^{q}(\phi_{\tau}^{q'}(Y_{\tau-1}, \theta^{\tau}), \theta^{\tau+1}),$$
$$q, q' \in Q,\ \tau = 1, \ldots, T-1$$

where θ^{τ} is the set of trainable parameters. Each transfer operator

$$\phi_{\tau}^{q_i}(\cdot),\ q_i \in Q,\ i = 1, \ldots, I \in \mathbb{N}$$

(Equation (1)) describes the τ layer of the neural network under consideration. The additional index $q_i \in Q$ for $i = 1, \ldots, I$, where $I < T$, defines a specific selection of the network architecture from a given hybrid collection. The family of given network architectures in the hybrid setting (1) is formalized by a finite index set $Q \subset \mathbb{N}$. This set involves a consistent formulation of the optimal structuring problem for a hybrid neural network. Note that the composition of the successive input-to-output transfer operators in (1) constitute the dynamic system.

Recall that in practical applications, the input collection Y_0 of a fully connected deep feedforward neural network (1) usually consists of some images, time-series, or other suitable input data. We consider the feedforward network architecture in any hybrid setting. The hybrid nature of this network is assumed to be given by the different activation functions

$$\sigma_{\tau}^{q}(\cdot),\ \tau = 1, \ldots, T,\ q_i \in Q,\ i = 1, \ldots, I \in \mathbb{N}$$

associated with the layers.

Note that in a conventional MLP framework, the trainable network parameter θ^{τ} in (1) constitutes a pair:

$$\theta^{\tau} := \left(W^{\tau}, b^{\tau}\right),\ \tau = 1, \ldots, T, \tag{2}$$

where

$$W^{\tau} \in \mathbb{R}^{N_{\tau-1} \times N_{\tau}},\ \ b^{\tau} \in \mathbb{R}^{N_{\tau}}.$$

We refer to [8] for the necessary technical details and practical implementations of conventional MLPs. We now introduce the following set of all admissible parameters θ^{τ} in (1):

$$\Theta_{\tau} \subset \mathbb{R}^{N_{\tau-1} \times N_{\tau}} \otimes \mathbb{R}^{N_{\tau}},\ \tau = 1, \ldots, T.$$

The definition in (2) of the successive network parameters $\theta^{\tau} \in \Theta_{\tau}$ in a neural network implies a formal characterization of the transfer operators in (1):

$$\phi_{\tau}^{q} : \mathbb{R}^{N_{\tau-1}} \times \Theta_{\tau} \to \mathbb{R}^{N_{\tau}}\ \ \tau = 1, \ldots, T.$$

We next implement this abstract operator-based description (1) of the dynamics and consider the family

$$F := \{f_{\tau}^{q}(\cdot, \cdot)\}_{\tau=1,\ldots,T}^{q \in Q} \tag{3}$$

of the state transformation functions

$$f_{\tau}^{q} : \mathbb{R}^{N_{\tau-1}} \times \Theta_{\tau} \to R^{N_{\tau}}.$$

In the layered network framework mentioned above, these state transformation functions are defined as follows:

$$f_{\tau}^{q}(y_{s,\tau-1}, \theta^{\tau}) = \sigma_{\tau}^{q}\left(W^{\tau} y_{s,\tau-1} + b^{\tau}\right),$$
$$\sigma_{\tau}^{q}(\cdot) \in \chi,\ \theta^{\tau} \in \Theta_{\tau},\ \tau = 1, \ldots, T \tag{4}$$

for $q \in Q$ and $s = 1, \ldots, S$. By

$$\chi := \{\sigma_\tau^q(\cdot)\}_{\tau=1,\ldots,T}^{q\in Q},$$

we next denote in (4) a family of admissible activation functions $\sigma_\tau^q(\cdot)$ for the hybrid neural network. For every $q \in Q$, the activation function $\sigma_\tau^q(\cdot)$ acts component-wise on the corresponding N_τ-dimensional vector-argument:

$$(W^\tau y_{s,\tau-1} + b^\tau).$$

Therefore, an activation function can formally be determined by a scalar mapping for every $q \in Q$. Recall that various choices of an activation function are available in the literature (see, e.g., [6,7]). Clearly, the hybrid deep neural network states $y_{s,\tau-1}$ in (4) imply the corresponding hybrid state collection:

$$Y_{\tau-1} := \{y_{s,\tau-1} \in \mathbb{R}^{N_{\tau-1}},\ s = 1, \ldots, S\},\ \tau = 1, \ldots, T.$$

in accordance with the operator description (1).

For the given network, the basic relation (4) implies a trainable state transformation of the following type:

$$y_{s,\tau} = f_\tau^{q_i}(y_{s,\tau-1}, \theta^\tau),\ \tau = 1, \ldots, T,\ i = 1, \ldots, I,\ q_i \in Q \tag{5}$$

where $y_{s,0} \in Y_0$ are the given network inputs. The dynamics of a fully connected feedforward hybrid neural network is given by (5). This relation describes the interplay of layers in a hybrid MLP model and can be naturally interpreted as a discrete-time-controlled dynamic system. For abbreviation, we refer to a hybrid deep neural network system as HDNNS.

We also note that the resulting discrete-time system (5) constitutes a specific example of a dynamics with the changing state dimensionality. We refer to [10] for some optimization approaches to systems with variable state dimensionality.

As mentioned above, the problem of training a fully connected feedforward neural network can now be considered as a specific nonlinear optimization problem with an additional structural optimization. For conventional (non-hybrid, where all layers have the same computational properties) multilayer neural network learning models, the optimization variable problem consists of the network parameters $\theta^\tau \in \Theta_\tau$, where $\tau = 1, \ldots, T$. We refer to [6–8] for further concepts and formal details. In contrast to the conventional learning, we now need to enlarge the set of optimization variables for the hybrid network learning models and include the additional discrete index variable $q \in Q$ into the optimization framework. Let

$$\vartheta := \left(\theta^\tau\right)_{\tau=1,\ldots,T},\ \Theta := \Theta_1 \otimes \ldots \otimes \Theta_T,$$

and

$$\rho := \left(q_i\right)_{i=1,\ldots,I}.$$

We have the parameter admissibility condition $\vartheta \in \Theta$. Let

$$\rho \in Q^I = \underbrace{Q \otimes \ldots \otimes Q}_{I \text{ times}}.$$

Now, we consider an objective functional

$$\tilde{J} : \Theta \otimes Q^I \to \mathbb{R}.$$

associated with the optimal training design of the hybrid MLP neural network. The main training problem, namely, the optimal structuring problem for a hybrid MLP network (see Introduction) can now be written in the form of the following OCP:

$$\begin{aligned} &\text{minimize } \tilde{J}(\vartheta,\rho)\\ &\text{subject to } y_{s,\tau} = f_{\tau}^{q_i}(y_{s,\tau-1},\theta^{\tau}),\\ &y_{s,0} \in Y_0,\ \theta^{\tau} \in \Theta_{\tau},\ q_i \in Q,\\ &\tau = 1,\dots,T,\ s = 1,\dots,S,\ i = 1,\dots,I. \end{aligned} \tag{6}$$

The network parameter and the hybrid indices, namely, the pair

$$(\vartheta,\rho) \in \Theta \otimes Q^I$$

can now be interpreted as a "control input" of the dynamic system (5). We use the control theoretic notation that is naturally motivated by the resulting OCP (6). The control input in problem (6) expresses the trainable design of a feedforward neural network. This design also includes the optimal structuring of the given layers. The goal of the resulting network training is to determine optimal parameters $\vartheta \in \Theta$ and the network structure (the hybrid indices $\rho \in Q^I$) such that the objective function $\tilde{J}(\cdot,\cdot)$ is minimized. In the case of learning models, the objective functional usually includes differences between the final network output $x_{s,T} \in X_T$ and some known targets, called "training labels".

We now introduce an additional necessary notation. By $y_{s,T}^{\vartheta,\rho} \in Y_T$, we next denote the final network output, namely, a solution of the HDNNS (5) for a concrete control pair $(\vartheta,\rho) \in \Theta \otimes Q^I$ and for an input of $y_{s,0} \in Y_0$. Generally, the objective functional $\tilde{J}(\vartheta,\rho)$ in (6) can be defined as follows:

$$\tilde{J}(\vartheta,\rho) := \frac{1}{S}\sum_{s=1}^{S} \Psi_s(y_{s,T}^{\vartheta,\rho}). \tag{7}$$

Here,

$$\Psi_s : \mathbb{R}^{N_T} \to \mathbb{R},\ s = 1,\dots,S$$

is a sufficiently smooth function. Let $c_i \in \mathbb{R}^{N_T}$, $i = 1,\dots,d$ be the vectors of training labels. Here, $d \in \mathbb{N}$ is the number of labels. In the framework of an MLP network optimization, one usually considers the following function $\Psi_s(\cdot,\cdot)$:

$$\Psi_s(y_{s,T}^{\vartheta,\rho}) = \sum_{i=1}^{d} ||\mathcal{C}(Ky_{s,T}^{\vartheta,\rho} + b^T) - c_i||_2^2\,, \tag{8}$$

where $\mathcal{C} : \mathbb{R} \to \mathbb{R}$ is a so-called classifier applied component-wise to a vector, $K \in \mathbb{R}^{N_T \times N_T}$, and $||\cdot||_2$ is the Euclidean norm. Recall that $b^T \in \mathbb{R}^{N_T}$. Let us observe that the generic DL optimization models usually involve an average objective functional of the type (7).

The main optimization problem (6) associated with a hybrid MLP neural network can be characterized as a discrete-time OCP with a switched structure. We next assume that this main OCP (6) possesses an optimal solution (optimal control):

$$(\vartheta^{\text{opt}},\rho^{\text{opt}}) \in \Theta \otimes Q^I.$$

The obtained main OCP (6) involves a so-called "terminal objective functional" (also called the Mayer functional). The control input for this system is an admissible pair (ϑ,ρ). An optimal control design formalized by problem (6) implies an adequate training procedure of the given MLP neural network with a hybrid structure. The process of finding the optimal

parameters $(\vartheta^{\mathrm{opt}}, \rho^{\mathrm{opt}})$ in optimization problem (6) can also be interpreted as a generalized (dynamic) regression problem. We refer to [11] for the necessary mathematical details and some important control theoretical results related to a class of dynamic optimization problems of the type (6).

The practically oriented DL models often include a regularized version of the initially given objective functional $\tilde{J}(\cdot,\cdot)$. In such a case, the objective functional $\tilde{J}(\cdot,\cdot)$ in (6) is replaced by the corresponding regularization:

$$\tilde{J}_{reg}(\vartheta,\rho) := \tilde{J}(\vartheta,\rho) + \frac{1}{S}\sum_{s=1}^{S}\sum_{\tau=1}^{T} R_\tau(\theta^\tau). \tag{9}$$

where $R_\tau : \Theta_\tau \to \mathbb{R}$ is a suitable regularization function. For the generic regularization framework (9), one can consider the celebrated Tikhonov–Phillips concept [7]. In a modern ML, this regularization method is also known as a Weight Decay approach (see, e.g., [6]). Note that depending on the concrete application, many different choices of a suitable regularization function $R_\tau(\cdot)$ are possible. We refer to [12] for a detailed discussion related to the proximal-point regularization methodology. Let us also note that in the case of a regularized functional $\tilde{J}_{\mathrm{reg}}(\vartheta,\rho)$, the resulting OCP constitutes a Bolza-type problem.

3. Application of a Switched System Methodology to the Hybrid Deep Learning Model

The HDNNS (5) introduced in the previous section describes a learning process of the hybrid MLP network. It exhibits both discrete-time and combinatorial dynamic behaviour. The combinatorial structure of system (5) and that of the main OCP (6) is given by the hybrid indices $\rho \in Q^I$ with $q_i \in Q$. In order to develop a constructive analytic description of this combinatorial part of the system, we next consider (5) in the context of switched control systems [11,13,14].

Recall that the switched system methodology has been established as a powerful modelling and solution approach to a wide class of complex real-world dynamic systems. We refer to [11] for the necessary theoretical concepts and practical engineering applications of switched systems.

We now use the generic definition of a switched control system from [11] and propose a novel formal description adapted to the HDNNS abstraction (5).

Definition 1. *A switched dynamic system associated with the* HDNNS *model* (5) *is a collection of* $\{Q, F, l, \Gamma\}$, *where the following apply:*

- $Q \subset \mathbb{N}$ *is a finite index set;*
- F *is a family of vector fields introduced in* (3)*;*
- $l = \{t_i\}$, $i = 1, \ldots, I$ *is a sequence of switching times associated with an implemented sequence of layers in the* HDNNS *such that*

$$0 = t_0 < t_1 < \ldots < t_{I-1} < t_I \leq T;$$

- $\Gamma \subset \Xi := \{(q, q') : q, q' \in Q\}$ *is a reset set for the network layers.*

Note that in the conventional hybrid/switched systems and control theory, the index set Q in Definition 1 is sometimes called a "set of locations". A "location" of the system under consideration is specified by a concrete variable $q_i \in Q$.

The switched dynamic system from Definition 1 is assumed to be determined on an interval $[0, T]$. A concrete sequence l of switching times

$$t_i,\ i = 1, \ldots, I$$

defines a partition of the interval $[0, T]$ by some adjoint subintervals $[t_{i-1}, t_i)$ associated with every hybrid layer index

$$q_i \in Q,\ i = 1, \ldots, I.$$

We now use the results of [11,12,14] and rewrite the complete dynamics of a switched system from Definition 1 in the following compact form:

$$\begin{aligned} & y_{s,\tau} = \sum_{i=1}^{I} \beta_{[t_{i-1},t_i)}(\tau) f_{\tau}^{q_i}(y_{s,\tau-1}, \theta^{\tau}), \\ & y_{s,0} \in X_0,\ \theta^{\tau} \in \Theta_{\tau},\ q_i \in Q, \\ & \tau = 1, \ldots, T,\ s = 1, \ldots, S,\ i = 1, \ldots, I. \end{aligned} \tag{10}$$

Here, $\beta_{[t_{i-1},t_i)}(\cdot),\ i = 1, \ldots, I$ is the characteristic function of the interval $[t_{i-1}, t_i)$.

The concept of a switched dynamic system from Definition 1 applied to the HDNNS model (5) leads to a constructive (non-combinatorial) representation of the control variable $\rho \in Q^I$ in system (5) and in the main OCP (6). The following vector-function

$$\beta(\cdot) := \left(\beta_{[t_{i-1},t_i)}(\cdot)\right)_{i=1,\ldots,I}$$

makes it possible for a function-based representation of the combinatorial control variable $\rho \in Q^I$. We next define a set $\mathcal{B}$ of all admissible characteristic functions $\beta(\cdot)$ for the discrete-time switched system (10):

$$\mathcal{B} := \{\beta(\cdot) \mid \beta[t_{i-1}, t_i)(\tau) \in \{0,1\}^I,\ \sum_{i=1}^{I} \beta_{[t_{i-1},t_i)}(\tau) = 1\},$$

where $\tau = 1, \ldots, T$. From Definition 1, it is easy to see that set $\mathcal{B}$ is in a one-to-one correspondence to the set of all admissible sequences l of switching times. These switching times correspond to the possible changes in the hybrid MLP network structure according to the main optimal structuring problem that we consider.

As mentioned above, the developed switched system approach to the HDNNS makes it possible to replace the control variable $\rho \in Q^I$ with an equivalent expression of $\beta(\cdot)$ in system (5). Let $y_{s,T}^{\vartheta,\beta}$ be a solution of the switched dynamic system (10) for a concrete selection of the control input

$$(\vartheta, \beta(\cdot)) \in \Theta \otimes \mathcal{B}.$$

Using the switched system's formalism, we now rewrite the main OCP (6) in the following form:

$$\begin{aligned} & \text{minimize } J(\vartheta, \beta(\cdot)) \\ & \text{subject to } y_{s,\tau} = \sum_{i=1}^{I} \beta_{[t_{i-1},t_i)}(\tau) f_{\tau}^{q_i}(y_{s,\tau-1}, \theta^{\tau}), \\ & y_{s,0} \in Y_0,\ \theta^{\tau} \in \Theta_{\tau},\ \beta(\cdot) \in \mathcal{B},\ q_i \in Q, \\ & \tau = 1, \ldots, T,\ s = 1, \ldots, S,\ i = 1, \ldots, I. \end{aligned} \tag{11}$$

The objective functional $J(\cdot,\cdot)$ in Equation (11) can be rewritten as

$$J(\vartheta,\beta(\cdot)) := \frac{1}{S}\sum_{s=1}^{S}\Psi_s(y_{s,T}^{\vartheta,\beta}). \tag{12}$$

Evidently, the objective $J(\vartheta,\beta(\cdot))$ in (12) is a combined functional.

For the given hybrid MLP network model, the state $y_{s,T}^{\vartheta,\beta}$ represents the final output of the neural network learning process. In accordance with the originally given OCP (6), we next assume that the switched-type problem (11) possesses an optimal solution

$$(\vartheta^{opt},\beta^{opt}(\cdot)) \in \Theta\otimes\mathcal{B}.$$

The corresponding optimal solution of the switched dynamic system (10) is next defined by $y_{s,\tau}^{opt}$.

The switched-type problem (11) can be considered as a "constructive" version of the initially presented OCP stated in (6). It contains a numerically tractable formalization $\beta(\cdot)\in\mathcal{B}$ of the combinatorial control component $\rho\in Q^I$ in (11). Similarly to the case of the original OCP regularization (9), we also introduce a regularized version $J_{\mathrm{reg}}(\cdot,\cdot)$ of the objective functional $J(\cdot,\cdot)$ of the problem in (11):

$$J_{reg}(\vartheta,\beta(\cdot)) := J(\vartheta,\beta(\cdot)) + \frac{1}{S}\sum_{s=1}^{S}\sum_{\tau=1}^{T}R_\tau(\theta^\tau). \tag{13}$$

The selection of a concrete regularization functional $R_\tau(\cdot)$ in (13) can be made in the same way as in problem (9).

4. A Critical Analysis of the Necessity of Optimality Conditions of Pontryagin Type

This section is devoted to the necessary optimality conditions for the switched-type OCP in Equation (11). These optimality conditions are given in the form of a Pontryagin Maximum Principle (PMP) for the general optimal control processes governed by switched dynamics. We refer to [15,16] for the necessary technical details, examples, and analytical results. The applicability analysis of the PMM for a concrete class of OCPs (11) involving the switched control systems (10) shows the inconsistency of a possible application of this PMP. As will be shown in this section, the switched systems' approach to the hybrid feedforward network learning involves the non-effectiveness of the celebrated PMP for problem (6). Since OCP (11) constitutes a mathematically equivalent representation of the hybrid MLP neural network learning problem (6), the same inconsistency conclusion will be true for the initially given OCP (6). This important fact provides the main motivation for the further development of some alternative solution schemes for the network learning problem under consideration.

Recall that the celebrated PMP expresses the necessary optimality conditions in many types of the conventional, hybrid, and switched-type OCPs. Consider now the switched-type OCP (6) and introduce the corresponding Hamiltonian function:

$$\begin{aligned} &H : \mathbb{N}\times\mathbb{R}^{N_{\tau-1}}\times\mathbb{R}^{N_\tau}\times\Theta_\tau\times\mathcal{B}\to\mathbb{R},\\ &H(\tau,y,p,\theta,\beta(\cdot)) := \langle\, p, \sum_{i=1}^{I}\beta_{[t_{i-1},t_i)}(\tau)f_\tau^{q_i}(y,\theta)\,\rangle_{\mathbb{R}^{N_\tau}} \end{aligned} \tag{14}$$

By $p \in \mathbb{R}^{N_\tau}$, we denote in (14) the generic adjoint variable, and $\langle \cdot, \cdot \rangle_Z$ is a scalar product in an Euclidean space Z. For every location

$$q_i \in Q,\ i = 1, \ldots, I$$

of the switched control system (10), we can define the corresponding partial Hamiltonian:

$$H^{q_i}(\tau, y, p, \theta, \beta_{[t_{i-1}, t_i)}(\cdot)) := \langle\ p, \beta_{[t_{i-1}, t_i)}(\tau) f_\tau^{q_i}(y, \theta)\ \rangle_{\mathbb{R}^{N_\tau}}. \tag{15}$$

Evidently,

$$H(\tau, y, p, \theta, \beta(\cdot)) = \sum_{i=1}^{I} H^{q_i}(\tau, y, p, \theta, \beta_{[t_{i-1}, t_i)}(\cdot)).$$

Recall that the classical PMP for a discrete-time OCP expresses the necessary optimality condition in terms of a Hamiltonian function. We now present an advanced version of the optimality conditions, namely, the PMP for the concrete discrete-time switched OCP (11).

Theorem 1. *Consider the discrete-time switched* OCP (11) *and assume that it has an optimal solution*

$$(\vartheta^{opt}, \beta^{opt}(\cdot)) \in \Theta \otimes \mathcal{B}$$

and is Lagrange*-regular. Assume that the right-hand side f_τ^q in* HDNSN (5) *is given by* (4)*, and the concrete function Ψ_s in* (12) *is defined by* (8)*. Let the classifier $\mathcal{C}$ be a continuously differentiable function. Assume further that for each $\tau = 1, \ldots, T$ and $y \in \mathbb{R}^{N_{\tau-1}}$, the orientor field*

$$\mathcal{F} := \{\sum_{i=1}^{I} \beta_{[t_{i-1}, t_i)}(\tau) f_\tau^{q_i}(y, \theta)\ :\ (\vartheta, \beta(\cdot)) \in \Theta \otimes \mathcal{B}\}$$

is convex. Then, there exists an adjoint process

$$P_\tau := \{p_{s,\tau} \in \mathbb{R}^{N_\tau},\ s = 1, ..., S\},\ \tau = 1, \ldots, T$$

such that the following relations are satisfied for $s = 1, ..., S$ and $\tau = 1, \ldots, T$:

$$\begin{aligned}
&y_{s,\tau}^{opt} = \sum_{i=1}^{I} \frac{\partial}{\partial p} H^{q_i}(\tau, y_{s,\tau-1}^{opt}, p_{s,\tau}, \theta^{\tau,opt}, \beta_{[t_{i-1}, t_i)}^{opt}(\cdot)),\\
&y_{s,0}^{opt} \in Y_0,\\
&p_{s,\tau-1} = \sum_{i=1}^{I} \frac{\partial}{\partial y} H^{q_i}(\tau, y_{s,\tau-1}^{opt}, p_{s,\tau}, \theta^{\tau,opt}, \beta_{[t_{i-1}, t_i)}^{opt}(\cdot)),\\
&p_{s,T} = -\frac{1}{S} \frac{\partial}{\partial y} \Psi_s(y_{s,T}^{opt}),\\
&\sum_{s=1}^{T} \sum_{i=1}^{I} H^{q_i}(\tau, y_{s,\tau-1}^{opt}, p_{s,\tau}, \theta^{\tau,opt}, \beta_{[t_{i-1}, t_i)}^{opt}(\cdot)) \geq\\
&\sum_{s=1}^{T} \sum_{i=1}^{I} H^{q_i}(\tau, y_{s,\tau-1}^{opt}, p_{s,\tau}, \theta, \beta_{[t_{i-1}, t_i)}(\cdot))\ \ \forall (\vartheta, \beta(\cdot)) \in \Theta \otimes \mathcal{B}.
\end{aligned} \tag{16}$$

We refer to [11,17], and some references therein, for the formulation and proof of the PMP for OCP (11). The classic PMP for the conventional discrete-time OCPs can be found in [16]. Let us note that the derivative

$$\frac{\partial}{\partial y} \Psi_s(y_{s,T}^{opt})$$

in (16) can be explicitly calculated in the framework of an MLP network. One uses the concrete expression (8) of function $\Psi_s(\cdot)$ for this purpose. Moreover, in the case of a regularized functional $J_{\text{reg}}(\vartheta, \beta(\cdot))$ determined by (13), the corresponding Hamiltonian of the resulting regularized OCP can be defined as follows:

$$H_{reg}(\tau, y, p, \theta, \beta(\cdot)) := H(\tau, y, p, \theta, \beta(\cdot)) - \frac{1}{S} R_\tau(\theta) =$$
$$\langle\, p, \sum_{i=1}^{I} \beta_{[t_{i-1}, t_i)}(\tau) f_\tau^{q_i}(y, \theta) \,\rangle_{\mathbb{R}^{N_\tau}} - \frac{1}{S} R_\tau(\theta).$$

Evidently, in this case, the regularization functional $R_\tau(\cdot)$ only appears in the last relation (inequality) of the generic system (16) of optimality conditions.

Versions of the PMP related to concrete OCPs provide a theoretical foundation for an important class of numerical methods in optimal control. This class involves the so-called indirect computational methods for OCPs. We refer to [11,12] for an overview. However, the PMP mentioned above includes a specific assumption, namely, the convexity condition of the orientor field $\mathcal{F}$. This convexity assumption constitutes the most crucial condition of Theorem 1. A possible violation of this assumption will imply an incorrectness of the PMP optimality conditions for the specific OCP (11). The same is also true with respect to the initially given hybrid MLP neuronal network learning problem (6). Note that in the case of the regularized OCP with the cost functional $J_{\text{reg}}(\vartheta, \beta(\cdot))$ and with the correspondingly regularized Hamiltonian $H_{\text{reg}}(\tau, y, p, \theta, \beta(\cdot))$, the conditions of Theorem 1 needs to be extended by the additional convexity assumption of the regularization function $R_\tau(\cdot)$ (see, e.g., [11]).

It is necessary to stress that the convexity assumption for the orientor field $\mathcal{F}$ has no relations to the convexity of the right-hand side of the switched dynamic system (10). We refer to [11,15,16] for various examples of the specific convex and non-convex orientor fields.

In the case of MLS trainable layers, the functions

$$f_\tau^q(\cdot, \cdot) \in F$$

have a quasi-affine structure with respect to

$$\theta^\tau \in \Theta_t$$

(see (4)). This special case includes many important types of layers, for example, the fully connected layers, convolution layers, and batch normalization layers (see, e.g., [6,7]). One can assume that the convexity of the parameter is set as Θ. However, the non-convexity of the set $\mathcal{B}$ implies the non-convexity of the complete (product) control set

$$\Theta \otimes \mathcal{B}$$

even in the case of a convex Θ_τ. This above property, the nonlinearity of the activation function σ_τ^q and the specific bilinear (non-convex) structure of the summands

$$\beta_{[t_{i-1}, t_i)}(\tau) f_\tau^{q_i}(x, \theta)$$

in the right-hand side of (10) exclude the necessary convexity property of the orientor field $\mathcal{F}$ required in Theorem 1. When $\mathcal{F}$ is a non-convex set then, in general, it is not true that the PMP from Theorem 1 constitutes the necessary optimality conditions for the switched-type OCP (11).

We now discuss an additional counter-argument for a possible consideration of the PMP (Theorem 1) for the numerical treatment of the switched OCP (11). It is well known that the necessary optimality condition in the form of a PMP provides a numerically consistent algorithm in a full space (for example, in a Lebesgue space) of admissible control functions. On the other hand, a direct application of the corresponding PMP to problem (11) does not guarantee admissibility of the numerical optimal solution $(\vartheta^*, \beta^*(\cdot))$ in the sense of problem (11), i.e., it does not guarantee the required condition:

$$\beta^*(\cdot) \in \mathcal{B}.$$

The above counter-argument shows that the celebrated PMP and the corresponding computational solution procedures cannot be applied directly to the specific problem (11). The "admissibility problem" mentioned above is a direct consequence of the following pivotal theoretic observation related to a formal mathematical proof of the PMP: the set of possible needle variations associated with the characteristic functions $\beta_{[t_{i-1},t_i)}(\cdot)$ in problem (11) constitutes a very "poor" set of variations (see, e.g., [15]). As a consequence, one cannot derive the generic adjoint equation that guarantees the belonging of the numerically optimal value $\beta^*(\cdot)$ to the set of $\mathcal{B}$ admissible switched controls. The possible numerical application of the optimality condition (system) (16), including the adjoint equation, namely, the difference equation for variable $p_{s,\tau-1}$ in (16), generally implies the inadmissibility $\beta^*(\cdot) \notin \mathcal{B}$ of the resulting numerically optimal variable $\beta^*(\cdot)$. Since $\mathcal{B}$ is a non-convex set, we cannot effectively apply any suitable projection approach for $\beta^*(\cdot)$.

The two main critical arguments discussed above reflect the conceptual applicability problem of the generic PMP in the context of numerical treatments of the OCP (11). This fact significantly restricts the possible numerical application of the presented Theorem 1 and the corresponding indirect solution algorithms. This situation provides a main motivation for the development of novel, direct-solution techniques for OCPs associated with the optimal learning of hybrid fully connected neural networks.

5. A Reduced-Gradient Approach to the Switched-Type OCP for Hybrid Neural Networks

This section presents a novel solution scheme for the switched-type OCP (11) related to the optimal structuring problem for a hybrid MLP neural network. As mentioned in Section 5, the possible application of the indirect (PMP-based) numerical method for the switched-type OCP (11) involves some conceptual difficulties. On the other hand, the specific structure of the optimization problem under consideration makes it possible to derive a constructive expression for the so-called "reduced gradient" of the objective functional $J(\cdot,\cdot)$ in OCP (11). The explicit characterization of the reduced gradient discussed in the next theorem will be used as an analytic basis for a specific first-order solution algorithm that we propose. This algorithm can be applied directly to the original switched OCP (11) as well as to the regularized version of the problem.

Consider OCP (11) associated with the given hybrid neural network of the MLP type with the continuously differentiable functions $\Psi_s(\cdot)$ from (8). Since the concrete functions $f_\tau^q(\cdot,\cdot)$ in (3) are also continuously differentiable, we derive the Fréchet differentiability of the objective functional $J(\cdot,\cdot)$ in OCP (11). The corresponding gradient of $J(\cdot,\cdot)$ is next denoted by $\nabla J(\cdot,\cdot)$.

We now apply the generalized reduced-gradient method that is comprehensively discussed in [11,12] to the concrete OCP (11) and obtain the following formal results.

Theorem 2. *Consider the switched* OCP (11) *and assume that all conditions of* Theorem 1 *are satisfied. The reduced gradient* $\nabla J(\cdot,\cdot)$ *of the objective functional* $J(\cdot,\cdot)$ *in problem* (11) *can be computed as a solution of the following system of equations:*

$$\begin{aligned}
&y_{s,\tau} = \frac{\partial}{\partial p} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)),\\
&y_{s,0} \in Y_0,\\
&p_{s,\tau-1} = \frac{\partial}{\partial y} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)),\\
&p_{s,T} = -\frac{1}{S} \nabla \Psi_s(y_{s,T}),\\
&\nabla J(\theta, \beta(\cdot)) = -\begin{pmatrix} \nabla_{\theta} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \\ \nabla_{\beta} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \end{pmatrix},\\
&\tau = 1, ..., T,\ s = 1, \ldots, S,\ i = 1, \ldots, I,\ q_i \in Q,
\end{aligned} \tag{17}$$

where $p_{s,\tau} \in \mathbb{R}^{N_\tau},\ \tau = 1, \ldots, T,\ s = 1, \ldots, S$ *is an adjoint variable.*

By $\nabla_{\theta} H$ and $\nabla_{\beta} H$, we denote here the partial derivative of H with respect to θ and β, respectively. A complete proof of Theorem 2 in the case of a fixed $s = 1, \ldots, S$ can be found in [11,12]. Some similar results for concrete classes of discrete-time and continuous-time OCPs are obtained in [17].

Using expression (14) for the system Hamiltonian $H(\tau, y, p, \theta, \beta(\cdot))$, we can calculate the elements $\nabla_{\theta} H$ and $\nabla_{\beta} H$ of the gradient vector $\nabla J(\theta, \beta)$ in (17):

$$\begin{aligned}
&\nabla_{\theta}^{T} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) =\\
&\langle\ p_{s,\tau}, \sum_{i=1}^{I} \beta_{[t_{i-1}, t_i)}(\tau) \frac{\partial}{\partial \theta} f_{\tau}^{q_i}(y_{s,\tau-1}, \theta^{\tau})\ \rangle_{\mathbb{R}^{N_\tau}},
\end{aligned}$$

and

$$\begin{aligned}
&\nabla_{\beta}^{T} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) =\\
&\left(\langle p_{s,\tau}, f_{\tau}^{q_1}(y_{s,\tau-1}, \theta^{\tau})\rangle_{\mathbb{R}^{N_\tau}}, \ldots, \langle p_{s,\tau}, f_{\tau}^{q_I}(y_{s,\tau-1}, \theta^{\tau})\rangle_{\mathbb{R}^{N_\tau}}\right).
\end{aligned}$$

By ∇_{θ}^{T} and ∇_{β}^{T}, we denote here the transpose operation. Let us note that a detailed calculation of the partial derivative

$$\nabla_{\theta}^{T} H(\tau, y, p, \theta, \beta(\cdot))$$

depends on a concrete selection of an activation function σ_{τ}^{q} in (4).

Consider now the regularized functional $J_{reg}(\cdot,\cdot)$ from (13) and the corresponding regularized OCP (11). We assume that the regularizing functions

$$R_{\tau}(\cdot),\ \tau = 1, ..., T$$

are continuously differentiable. In this case, the Hamiltonian $H(\tau, y, p, \theta, \beta(\cdot))$ introduced in (14) needs to be replaced by its regularized version that was introduced in the previous section:

$$H_{reg}(\tau, y, p, \theta, \beta(\cdot)).$$

Using this replacement, the reduced gradient $\nabla J_{reg}(\cdot,\cdot)$ in the regularized OCP can be found from the corresponding extension of the basic system of Equation (17) from Theorem 2. Let us formulate the corresponding result for a regularized OCP (11).

Theorem 3. *Consider a regularized version of the switched* OCP (11) *and assume that all conditions of* Theorem 1 *are satisfied. The reduced gradient* $\nabla J_{reg}(\cdot,\cdot)$ *of the objective functional* $J_{reg}(\cdot,\cdot)$ *can be computed as a solution of the following system of equations:*

$$\begin{aligned}
&y_{s,\tau} = \frac{\partial}{\partial p} H_{reg}(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)),\\
&y_{s,0} \in Y_0,\\
&p_{s,\tau-1} = \frac{\partial}{\partial y} H_{reg}(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)),\\
&p_{s,T} = -\frac{1}{S}\nabla \Psi_s(y_{s,T}),\\
&\nabla J_{reg}(\theta, \beta(\cdot)) = -\begin{pmatrix} \nabla_{\theta} H_{reg}(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \\ \nabla_{\beta} H_{reg}(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \end{pmatrix} =\\
&\begin{pmatrix} \frac{1}{S}\nabla R_{\tau}(\theta^{\tau}) \\ 0_{\mathbb{R}^I} \end{pmatrix} - \begin{pmatrix} \nabla_{\theta} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \\ \nabla_{\beta} H(\tau, y_{s,\tau-1}, p_{s,\tau}, \theta^{\tau}, \beta(\tau)) \end{pmatrix},\\
&\tau = 1, \ldots, T,\ s = 1, \ldots, S,\ i = 1, \ldots, I,\ q_i \in Q,
\end{aligned} \tag{18}$$

where $0_{\mathbb{R}^I}$ *denotes the zero vector in the* Euclidean *space* $\mathbb{R}^I$.

The results of Theorems 2 and 3, namely, the constructive expressions for the reduced gradients in the originally given OCP (11) and in the regularized problem, make it possible to consider various first-order gradient-based computational techniques for a numerical treatment of the switched OCP (11). Let us present here the generic reduced-gradient algorithm for problem (11):

$$\begin{aligned}
&\text{initial step :}\\
&\begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(0)} \in \Theta \otimes \mathcal{B};\\
&\text{iterations :}\\
&\begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(k+1)} = \begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(k)} - \gamma_k \nabla J(\theta^{(k)}, \beta^{(k)}(\cdot)),\ k = 0, 1, \ldots\\
&\text{inclusion condition :}\\
&\begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(k+1)} \in \Theta \otimes \mathcal{B}.
\end{aligned} \tag{19}$$

Here, $k = 0, 1, \ldots$ is the iteration index. Clearly, for the objective functional $J_{\text{reg}}(\cdot,\cdot)$ from the regularized version of problem (11), the gradient in (19) needs to be replaced by the constructive expression of

$$\nabla J_{\text{reg}}(\theta^{(k)}, \beta^{(k)}(\cdot))$$

from Theorem 3 (see (19)).

We now study the numerical stability of the basic gradient method (19). Recall that the convergence properties of the generic reduced-gradient method for the switched system optimization was comprehensively discussed by many authors (see, e.g., [11,17,18] and references therein). Concretely, we can prove the following special convergence result for the reduced-gradient algorithm (19).

Theorem 4. *Consider OCP* (11) *and assume that all conditions of* Theorem 1 *are satisfied. Let*

$$\left\{ \begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(k)} \right\}_{k=0,1,\ldots} \tag{20}$$

be a sequence generated by method (19)*. Then, for an admissible initial point*

$$\begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(0)} \in \Theta \otimes \mathcal{B},$$

the above sequence (20) *is a minimizing sequence for problem* (11)*, i.e.,*

$$\lim_{k\to\infty} J(\vartheta, \beta^k(\cdot)) = J(\vartheta^{opt}, \beta^{opt}(\cdot)).$$

Moreover, this sequence converges weakly to a solution $(\vartheta^{opt}, \beta^{opt}(\cdot))$ *of the switched-type* OCP (11).

Proof. The property of sequence (20) to be a minimizing sequence for problem (11) immediately follows from [9]. Moreover, the concrete functions

$$f_\tau^q(y_{s,\tau-1}, \theta^\tau) \in F,$$

where F is defined in (3), are Lipschitz continuous and possess the Lipschitz continuous derivative for all $q \in Q$. Since the composition of two Lipschitz continuous mappings involves the same property for the resulting mapping, we obtain the Lipschitz continuity of the gradient $\nabla J(\theta, \beta(\cdot))$. The weak convergence of sequence (20) generated by method (19) to the optimal pair

$$(\vartheta^{opt}, \beta^{opt}(\cdot))$$

follows now from [19]. The proof is completed. □

Let us make some observations related to Theorem 4. Since $\vartheta \in \Theta$ and Θ is a subset of a finite-dimensional Euclidean space (see Section 2), the weak convergence of the first component ϑ in (20) to ϑ^{opt} coincides with the norm convergence in this Euclidean space. The weak convergence of the second component $\beta(\cdot)$ in (20) to $\beta^{opt}(\cdot)$ is in fact a weak convergence in the Hilbert space of all square integrable functions of the corresponding dimensionality.

Note that the inclusion condition

$$\begin{pmatrix} \theta \\ \beta(\cdot) \end{pmatrix}^{(k+1)} \in \Theta \otimes \mathcal{B}$$

for the updated iterations in (19) can be implemented using a projection. Assume that set Θ is convex. In that case, one needs to consider the following convexification of the product set $\Theta \otimes \mathcal{B}$:

$$\Theta \otimes \mathrm{conv}\{\mathcal{B}\},$$

where $\mathrm{conv}\{\mathcal{B}\}$ is a convexification of the function set $\mathcal{B}$. We refer to [11,20] for further theoretical and computational details related to the relaxation (convexification) procedures for gradient methods.

Finally, note that the proposed reduced-gradient algorithm provides an adequate alternative to the PMM-based optimality conditions critically discussed in Section 4. It constitutes a novel, OCP-based approach to the optimal structuring problem for a hybrid MLP neural network.

6. Concluding Remarks

In this paper, we developed a novel, optimal control-based approach to an optimization problem associated with the training of the feedforward fully connected hybrid neural network. We use the control theoretical approach and reformulate the main optimal structuring problem for a hybrid MLP neural network in the form of a specific switched-type OCP. This equivalent reformulation makes it possible to apply the advanced optimal control methodology to the training and structural optimization of a given hybrid MLP network. To be precise, we considered the optimal control of discrete-time switched systems and developed a reduced-gradient method for the practical treatment of the resulting switched OCP.

Let us note that the proposed equivalent representation of the trainable network design has a potential to be used in the optimal training of a convolution neural network. This optimistic conclusion follows from the general structure of the CNN (see, e.g., [21,22]) and from the existence of many well-developed theoretical results and computational methods for optimal control.

The equivalent reduction in a given sophisticated fully connected neural network design problem to a semiclassical OCP shows the effectiveness of the proposed optimal control methodology for neural network learning.

This is a preliminary proposal to optimize the connection weights of sequential layers of computing nodes to accomplish a mapping from a domain to a range, which is the task performed by a neural network. This model is capable of accommodating different activation functions at different layers, which prompted us to call it a hybrid (layer) neural network. This work constitutes initial theoretical research on the subject. The advanced optimal control approach discussed here is a complementary technique in the context of existing neural network training algorithms. To establish our approach as a viable alternative to established learning algorithms, we need to prototype and simulate the model. Comprehensive simulation studies for performance analysis and comparison with the traditional method of learning, both in the context of computational efficiency and quality of results, are needed. This is our future work.

Author Contributions: Conceptualization, G.C. and V.A.; Methodology, G.C.; Validation, L.A.G.T.; Formal analysis, V.A.; Writing—original draft, G.C.; Writing—review & editing, V.A., L.A.G.T. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Y_0	The set of inputs of the supervised data, 0 is the label for input
N_0	The dimension of the input data
Y_T	The set of outputs of the supervised data, T is the label for output
N_T	The dimension of the output data
T	Thus, $T \in \mathbb{N}$, the integer T is the number of layers, as well as the output layer and the output data label
S	The number of training data s is used as the index of the training data
$y_{s,0}$	$y_{s,0} \in \mathbb{R}^{N_0}$, $s = 1, \ldots, S$ is the s^{th} input data
$y_{s,T}$	$y_{s,0} \in \mathbb{R}^{N_0}$, $s = 1, \ldots, S$ is the s^{th} output data

τ	Index for network layer number
$N_{\tau-1}$	Number of input nodes (dimension of the input) to τ^{th} layer
N_τ	Number of output nodes (dimension of the output) of τ^{th} layer
θ^τ	$\theta^\tau := \left(W^\tau, b^\tau\right)$, $\tau = 1, \ldots, T$, are the trainable parameters of the τ^{th} network layer
W^τ	The connection weights between input and output nodes of τ^{th} layer
b^τ	The connection weights of the bias node of τ^{th} layer
Q	A finite index set for a family of predefined computational models; this allows the network architecture to accommodate hybrid layers
$q_i \in Q$	q_i is a specific i^{th} option from $Q = q_1, \ldots, q_i, \ldots, q_I$ $I \leq T$, i.e., there could be layers of same computation models
$\phi_\tau^{q_i}$	The transfer function of τ^{th} layer using $q_i \in Q$
$\phi_{\tau+1}^{q} \circ \phi_\tau^{q'}$	The combined transfer function of two consecutive network layers Layer τ's output is the input to layer $\tau + 1$
Θ_τ	$\Theta_\tau \subset \mathbb{R}^{N_{\tau-1} \times N_\tau} \otimes \mathbb{R}^{N_\tau}$ is the set of admissible parameters of θ^τ
ϕ_τ^q	The τ^{th} layer transfer operator, where $\mathbb{R}^{N_{\tau-1}} \times \Theta_\tau \to \mathbb{R}^{N_\tau}$
f_τ^q	$f_\tau^q : \mathbb{R}^{N_{\tau-1}} \times \Theta_\tau \to R^{N_\tau}$ The transfer function from input to output vector
σ_τ^q	The activation of type $q \in Q$; thus, $f_\tau^q(y_{s,\tau-1}, \theta^\tau) = \sigma_\tau^q\left(W^\tau y_{s,\tau-1} + b^\tau\right)$
ϑ, ρ	The variables for θ^τ and q_i, respectively
$\tilde{J}$	The objective function for regularization expressed in Equation (6)

References

1. Rumelhart, D.E.; McClelland, L.J. *Parallel Distributed Processing Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, MA, USA, 1987.
2. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
3. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1997**, *4*, 251–257. [CrossRef]
4. Zhong, S.; Cherkassky, V. Factors controlling generalization ability of MLP networks. In Proceedings of the IJCNN'99 (International Joint Conference on Neural Networks), Washington, DC, USA, 10–16 July 1999.
5. Chakraborty, G.; Murakami, M.; Shiratori, N.; Noguchi, S. A Growing Network that optimizes between undertraining and Overtraining. In Proceedings of the IEEE International Conference on Neural Network, Perth, Australia, 27 November–1 December 1995; Volume 2, pp. 1116–1121.
6. Benning, M.; Calledoni, E.; Ehrhardt, M.J.; Owren, B.; Schoenlieb, C.-B. Deep learning as optimal control problems: Models and numerical methods. *J. Comput. Dyn.* **2015**, *6*, 171–198. [CrossRef]
7. Li, Q.; Hao, S. An optimal control approach to deep learning and applications to discrete-weight neural networks. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
8. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
9. Bertsekas, D.P. *Nonlinear Programming*; Athena Scientific: Belmont, MA, USA, 1999.
10. Galván-Guerra, R.; Azhmyakov, V.; Egerstedt, M. Optimization of multiagent systems with increasing state dimensions: Hybrid LQ approach. In Proceedings of the 2011 American Control Conference, San Francisco, CA, USA, 29 June–1 July 2011; pp. 881–887.
11. Azhmyakov, V. *A Relaxation Based Approach to Optimal Control of Switched Systems*; Elsevier: Oxford, UK, 2019.
12. Azhmyakov, V.; Basin, M.; Raisch, J. A proximal point based approach to optimal control of affine switched systems. *Discret. Event Dyn. Syst.* **2012**, *22*, 61–81. [CrossRef]
13. Atlee Jackson, E. On the control of complex dynamic systems. In *Physica D: Nonlinear Phenomena*; Elsevier: Amsterdam, The Netherlands, 1991; Volume 50, pp. 341–366.
14. Egerstedt, M.; Wardi, Y.; Axelsson, H. Transition-time optimization for switched systems. *IEEE Trans. Autom. Control* **2006**, *AC-51*, 110–115.
15. Boltyanski, V.; Poznyak, A. *The Robust Maximum Principle*; Birkhauser: New York, NY, USA, 2012.
16. Halkin, H. A Maximum Principle of the Pontryagin type for systems described by nonlinear difference equations. *SIAM J. Control* **1966**, *4*, 90–111. [CrossRef]
17. Teo, K.L.; Goh, C.J.; Wong, K.H. *A Unifed Computational Approach to Optimal Control Problems*; Wiley: New York, NY, USA, 1991.
18. Polak, E. *Optimization*; Springer: New York, NY, USA, 1997.
19. Goldstein, A.A. Convex programming in Hilbert space. *Bull. Am. Math. Soc.* **1964**, *70*, 709–710. [CrossRef]

20. Roubicek, T. *Relaxation in Optimization Theory and Variational Calculus*; De Gruyter: Berlin, Germany, 1997.
21. Saleh, A.M.; Hamoud, T. Analysis and best parameters selection for person recognition based on gait model using CNN algorithm and image augmentation. *J. Big Data* **2021**, *8*, 1. [CrossRef] [PubMed]
22. Wang, J.; Lin, J.; Wang, Z. Efficient convolution architectures for convolutional neural network. In Proceedings of the 2016 8th International Conference on Wireless Communications and Signal Processing (WCSP), Yangzhou, China, 13–15 October 2016; pp. 1–5.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.