Deep Neural Network Structures Solving Variational Inequalities *

Patrick L Combettes, Jean-Christophe Pesquet

Deep Neural Network Structures Solving Variational Inequalities*

Patrick L. Combettes[1] and Jean-Christophe Pesquet[2]

[1]North Carolina State University, Department of Mathematics, Raleigh, NC 27695-8205, USA.
plc@math.ncsu.edu

[2]University Paris-Saclay, CentraleSupélec, Center for Visual Computing, OPIS Inria project team, 91190 Gif sur Yvette, France. jean-christophe@pesquet.eu

Abstract. Motivated by structures that appear in deep neural networks, we investigate nonlinear composite models alternating proximity and affine operators defined on different spaces. We first show that a wide range of activation operators used in neural networks are actually proximity operators. We then establish conditions for the averagedness of the proposed composite constructs and investigate their asymptotic properties. It is shown that the limit of the resulting process solves a variational inequality which, in general, does not derive from a minimization problem. The analysis relies on tools from monotone operator theory and sheds some light on a class of neural networks structures with so far elusive asymptotic properties.

Keywords. averaged operator, deep neural network, monotone operator, nonexpansive operator, proximity operator, variational inequality.

1 Introduction

A powerful tool from fixed point theory to analyze and solve optimization and inclusion problems in a real Hilbert space $\mathcal{H}$ is the class of averaged nonexpansive operators, which was introduced in [3]. Let $T\colon \mathcal{H} \to \mathcal{H}$ be a nonexpansive operator, i.e., T is 1-Lipschitzian. Then $\alpha \in \,]0,1]$ is an averagedness constant of T if $\mathrm{Id} + \alpha^{-1}(T - \mathrm{Id})$ remains nonexpansive, in which case we say that T is α*-averaged*; if $\alpha = 1/2$, T is *firmly nonexpansive*. The importance of firmly nonexpansive operators in convex optimization and variational methods has long been recognized [19, 27, 36, 41, 46]. The broader class of averaged operators was shown in [7] to play a prominent role in the analysis of convex feasibility problems. In this context, the underlying problem is to find a common fixed point of averaged operators. In [20], it was shown that many convex minimization and monotone inclusion problems reduce to the more general problem of finding a fixed point of compositions of averaged operators, which provided a unified analysis of various proximal splitting algorithms. Along these lines, several fixed point methods based on various combinations of averaged operators have since been devised, see [1, 2, 5, 9, 11, 13, 14, 17, 18, 24, 25, 38, 43, 47] for recent work. Motivated by deep neural network structures with thus far elusive asymptotic properties, we investigate in the present paper a novel averaged operator model involving a mix of nonlinear and linear operators.

Artificial neural networks have attracted considerable attention as a tool to better understand, model, and imitate the human brain [31, 37, 42]. In a Hilbertian setting [6], an $(n+1)$-layer feed-forward neural network architecture acting on real Hilbert spaces $(\mathcal{H}_i)_{0\leqslant i\leqslant n}$ is defined as the composition of operators $R_n \circ (W_n \cdot + b_n) \circ \cdots \circ R_1 \circ (W_1 \cdot + b_1)$ where, for every $i \in \{1,\ldots,n\}$, $R_i\colon \mathcal{H}_i \to$

*Contact author: P. L. Combettes, plc@math.ncsu.edu, phone: +1 919 515-2671. The work of P. L. Combettes was supported by the National Science Foundation under grant CCF-1715671. The work of J.-C. Pesquet was supported by Institut Universitaire de France.

$\mathcal{H}_i$ is a nonlinear operator known as an activation operator, $W_i\colon \mathcal{H}_{i-1} \to \mathcal{H}_i$ is a linear operator, known as a weight operator, and $b_i \in \mathcal{H}_i$ is a so-called bias parameter. Deep neural networks feature a (possibly large) number n of layers. In recent years, they have been found to be quite successful in a wide array of classification, recognition, and prediction tasks; see [34] and its bibliography. Despite their success, the operational structure and properties of deep neural networks are not yet well understood from a mathematical viewpoint. In the present paper, we propose to analyze them within the following iterative model. We emphasize that our purpose is not to study the training of the network, which consists of optimally setting the weight operators and bias parameters from data samples, but to analyze mathematically such a structure once it is trained. Our model is also of general interest in constructive fixed point theory for monotone inclusion problems.

Model 1.1 Let $m \geqslant 1$ be an integer, let $\mathcal{H}$ and $(\mathcal{H}_i)_{0\leqslant i\leqslant m}$ be nonzero real Hilbert spaces, such that $\mathcal{H}_m = \mathcal{H}_0 = \mathcal{H}$. For every $i \in \{1,\ldots,m\}$ and every $n \in \mathbb{N}$, let $W_{i,n}\colon \mathcal{H}_{i-1} \to \mathcal{H}_i$ be a bounded linear operator, let $b_{i,n} \in \mathcal{H}_i$, and let $R_{i,n}\colon \mathcal{H}_i \to \mathcal{H}_i$. Let $x_0 \in \mathcal{H}$, let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $]0,+\infty[$, set

$$(\forall n \in \mathbb{N})(\forall i \in \{1,\ldots,m\}) \quad T_{i,n}\colon \mathcal{H}_{i-1} \to \mathcal{H}_i\colon x \mapsto R_{i,n}(W_{i,n}x + b_{i,n}), \tag{1.1}$$

and iterate

$$\begin{array}{l}\text{for } n = 0,1,\ldots\\ \left\lfloor \begin{array}{ll} x_{1,n} & = T_{1,n}x_n\\ x_{2,n} & = T_{2,n}x_{1,n}\\ & \vdots\\ x_{m,n} & = T_{m,n}x_{m-1,n}\\ x_{n+1} & = x_n + \lambda_n(x_{m,n} - x_n). \end{array}\right.\end{array} \tag{1.2}$$

In sharp contrast with existing algorithmic frameworks involving averaged operators (see cited works above), the operators involved in Model 1.1 are not necessarily all defined on the same Hilbert space and, in addition, they need not all be averaged. Let us also note that the relaxation parameters $(\lambda_n)_{n\in\mathbb{N}}$ in (1.2) allow us to model skip connections [44], in the spirit of residual networks [33]. If $\lambda_n \equiv 1$, we obtain the standard feed-forward architecture [31].

Our contributions are articulated around the following findings.

- We show that a wide range of activation operators used in neural networks are actually proximity operators, which paves the way to the analysis of such networks via fixed point theory for monotone inclusion problems.
- We provide a new analysis of compositions of proximity and affine operators, establishing mild conditions that guarantee that the resulting operator is averaged.
- We show that, under suitable assumptions, the asymptotic output of the network converges to a point defined via a variational inequality. Furthermore, in general, this variational inequality does not derive from a minimization problem.

The remainder of the paper is organized as follows. In Section 2, we bring to light strong connections between the activation functions employed in neural networks and the theory of proximity operators in convex analysis. In Section 3, we derive new results on the averagedness properties of compositions of proximity and affine operators acting on different spaces. In Section 4, we investigate the asymptotic behavior of a class of deep neural networks structures and show that their fixed points solve a variational inequality. The main assumption on this subclass of Model 1.1 is that the structure of the network is periodic in the sense that a group of layers is repeated. Finally, in Section 5, the same properties are established for a broader class of networks.

Notation. We follow standard notation from convex analysis and operator theory [8, 40]. Thus, the expressions $x_n \rightharpoonup x$ and $x_n \to x$ denote, respectively, weak and strong convergence of a sequence $(x_n)_{n\in\mathbb{N}}$ to x in $\mathcal{H}$, and $\Gamma_0(\mathcal{H})$ is the class of lower semicontinuous convex functions $\varphi\colon \mathcal{H} \to \left]-\infty,+\infty\right]$ such that $\operatorname{dom}\varphi = \big\{x \in \mathcal{H} \;\big|\; \varphi(x) < +\infty\big\} \neq \varnothing$. Now let $\varphi \in \Gamma_0(\mathcal{H})$. The conjugate of φ is denoted by φ^*, its subdifferential by $\partial\varphi$, and its proximity operator is $\operatorname{prox}_\varphi\colon \mathcal{H} \to \mathcal{H}\colon x \mapsto \operatorname{argmin}_{y\in\mathcal{H}}(\varphi(y) + \|x-y\|^2/2)$. The symbols $\operatorname{ran} T$, $\operatorname{dom} T$, $\operatorname{Fix} T$, and $\operatorname{zer} T$ denote respectively the range, the domain, the fixed point set, and the set of zeros of an operator T. The space of bounded linear operators from a Banach space $\mathcal{X}$ to a Banach space $\mathcal{Y}$ is denoted by $\mathcal{B}\,(\mathcal{X},\mathcal{Y})$. Finally, ℓ_+^1 denotes the set of summable sequences in $\left[0,+\infty\right[$.

2 Proximal activation in neural networks

The following facts will be needed.

Lemma 2.1 *Let $\varphi \in \Gamma_0(\mathcal{H})$. Then the following hold:*

(i) [8, Proposition 12.29] $\operatorname{Fix}\operatorname{prox}_\varphi = \operatorname{Argmin}\varphi$.

(ii) [8, Corollary 24.5] *Let $g \in \Gamma_0(\mathcal{H})$ be such that $\varphi = g - \|\cdot\|^2/2$. Then $\operatorname{prox}_\varphi = \nabla g^*$.*

2.1 Activation functions

An activation function is a function $\varrho\colon \mathbb{R} \to \mathbb{R}$ which models the firing activity of neurons. The simplest instance, that goes back to the perceptron machine [42], is that of a binary firing model: the neuron is either firing or at rest. For instance, if the firing level is 1 and the rest state is 0, we obtain the binary step function

$$\varrho\colon \xi \mapsto \begin{cases} 1, & \text{if } \xi > 0;\\ 0, & \text{if } \xi \leqslant 0, \end{cases} \tag{2.1}$$

which was initially proposed in [37]. As this discontinuous activation model may lead to unstable neural networks, various continuous approximations have been proposed. Our key observation is that a vast array of activation functions used in neural networks belong to the following class.

Definition 2.2 The set of functions from $\mathbb{R}$ to $\mathbb{R}$ which are increasing, 1-Lipschitzian, and take value 0 at 0 is denoted by $\mathcal{A}(\mathbb{R})$.

Remarkably, we can precisely characterize this class of activation functions as that of proximity operators.

Proposition 2.3 *Let $\varrho\colon \mathbb{R} \to \mathbb{R}$. Then $\varrho \in \mathcal{A}(\mathbb{R})$ if and only if there exists a function $\phi \in \Gamma_0(\mathbb{R})$, which has 0 as a minimizer, such that $\varrho = \operatorname{prox}_\phi$.*

Proof. The fact that the class of increasing, 1-Lipschitzian functions from $\mathbb{R}$ to $\mathbb{R}$ coincides with that of proximity operators of functions in $\Gamma_0(\mathbb{R})$ is shown in [22, Proposition 2.4]. In view of Lemma 2.1(i) and Definition 2.2, the proof is complete. ☐

To illustrate the above results, let us provide examples of common activation functions $\varrho \in \mathcal{A}(\mathbb{R})$, and identify the potential ϕ they derive from in Proposition 2.3 (see Fig. 1).

Example 2.4 The most basic activation function is $\varrho = \mathrm{Id} = \mathrm{prox}_0$. It is in particular useful in dictionary learning approaches, which correspond to the linear special case of Model 1.1 [45].

Example 2.5 The saturated linear activation function [31]

$$\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \begin{cases} 1, & \text{if } \xi > 1;\\ \xi, & \text{if } -1 \leqslant \xi \leqslant 1;\\ -1, & \text{if } \xi < -1 \end{cases} \tag{2.2}$$

can be written as $\varrho = \mathrm{prox}_\phi$, where ϕ is the indicator function of $[-1, 1]$.

Example 2.6 The rectified linear unit (ReLU) activation function [39]

$$\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \begin{cases} \xi, & \text{if } \xi > 0;\\ 0, & \text{if } \xi \leqslant 0 \end{cases} \tag{2.3}$$

can be written as $\varrho = \mathrm{prox}_\phi$, where ϕ is the indicator function of $[0, +\infty[$.

Example 2.7 Let $\alpha \in \left]0, 1\right]$. The parametric rectified linear unit activation function [32] is

$$\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \begin{cases} \xi, & \text{if } \xi > 0;\\ \alpha\xi, & \text{if } \xi \leqslant 0. \end{cases} \tag{2.4}$$

We have $\varrho = \mathrm{prox}_\phi$, where

$$\phi\colon \mathbb{R} \to \left]-\infty, +\infty\right] : \xi \mapsto \begin{cases} 0, & \text{if } \xi > 0;\\ (1/\alpha - 1)\xi^2/2, & \text{if } \xi \leqslant 0. \end{cases} \tag{2.5}$$

Proof. Let $\xi \in \mathbb{R}$. Then $\phi'(\xi) = 0$ if $\xi > 0$, and $\phi'(\xi) = (1/\alpha - 1)\xi$ if $\xi \leqslant 0$. In turn $(\mathrm{Id} + \phi')\xi = \xi$ if $\xi > 0$, and $(\mathrm{Id} + \phi')(\xi) = \xi/\alpha$ if $\xi \leqslant 0$. Hence, $\varrho = (\mathrm{Id} + \phi')^{-1}$ is given by (2.4). ☐

Example 2.8 The bent identity activation function $\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto (\xi + \sqrt{\xi^2 + 1} - 1)/2$ satisfies $\varrho = \mathrm{prox}_\phi$, where

$$\phi\colon \mathbb{R} \to \left]-\infty, +\infty\right] : \xi \mapsto \begin{cases} \xi/2 - \big(\ln(\xi + 1/2)\big)/4, & \text{if } \xi > -1/2;\\ +\infty, & \text{if } \xi \leqslant -1/2. \end{cases} \tag{2.6}$$

Proof. This follows from [23, Lemma 2.6 and Example 2.18]. ☐

Example 2.9 The inverse square root unit activation function [16] is $\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \xi/\sqrt{1 + \xi^2}$. We have $\varrho = \mathrm{prox}_\phi$, where

$$\phi\colon \mathbb{R} \to \left]-\infty, +\infty\right] : \xi \mapsto \begin{cases} -\xi^2/2 - \sqrt{1 - \xi^2}, & \text{if } |\xi| \leqslant 1;\\ +\infty, & \text{if } |\xi| > 1. \end{cases} \tag{2.7}$$

Proof. Let $\xi \in \left]-1, 1\right[= \mathrm{dom}\,\nabla\phi = \mathrm{dom}\,\partial\phi = \mathrm{ran}\,\mathrm{prox}_\phi$. Then $\xi + \phi'(\xi) = \xi/\sqrt{1 - \xi^2}$ and therefore $\mathrm{prox}_\phi = (\mathrm{Id} + \phi')^{-1}\colon \mu \mapsto \mu/\sqrt{1 + \mu^2}$. ☐

Example 2.10 The inverse square root linear unit activation function [16]

$$\varrho\colon \mathbb{R}\to\mathbb{R}\colon \xi\mapsto \begin{cases} \xi, & \text{if } \ \xi\geqslant 0;\\ \dfrac{\xi}{\sqrt{1+\xi^2}}, & \text{if } \ \xi<0 \end{cases} \tag{2.8}$$

can be written as $\varrho = \operatorname{prox}_\phi$, where

$$\phi\colon \mathbb{R}\to\left]-\infty,+\infty\right]\colon \xi\mapsto \begin{cases} 0, & \text{if } \ \xi\geqslant 0;\\ 1-\xi^2/2-\sqrt{1-\xi^2}, & \text{if } \ -1\leqslant\xi<0;\\ +\infty, & \text{if } \ \xi<-1. \end{cases} \tag{2.9}$$

Proof. Let $\xi\in\left]-1,+\infty\right[=\operatorname{dom}\nabla\phi=\operatorname{ran}\operatorname{prox}_\phi$. Then $\xi+\phi'(\xi)=\xi$ if $\xi\geqslant 0$, and $\xi+\phi'(\xi)=\xi/\sqrt{1-\xi^2}$ if $\xi<0$. Hence, $\varrho=(\operatorname{Id}+\phi')^{-1}$ is given by (2.8). ☐

Example 2.11 The arctangent activation function $(2/\pi)\arctan$ is the proximity operator of

$$\phi\colon \mathbb{R}\to\left]-\infty,+\infty\right]\colon \xi\mapsto \begin{cases} -\dfrac{2}{\pi}\ln\left(\cos\left(\dfrac{\pi\xi}{2}\right)\right)-\dfrac{1}{2}\xi^2, & \text{if } \ |\xi|<1;\\ +\infty, & \text{if } \ |\xi|\geqslant 1. \end{cases} \tag{2.10}$$

Proof. Let $\xi\in\left]-1,1\right[=\operatorname{dom}\nabla\phi=\operatorname{ran}\operatorname{prox}_\phi$. Then $\xi+\phi'(\xi)=\tan(\pi\xi/2)$ and therefore $\varrho=(\operatorname{Id}+\phi')^{-1}=(2/\pi)\arctan$. ☐

Example 2.12 The hyperbolic tangent activation function tanh [35] is the proximity operator of

$$\phi\colon \mathbb{R}\to\left]-\infty,+\infty\right]\colon \xi\mapsto \begin{cases} \dfrac{(1+\xi)\ln(1+\xi)+(1-\xi)\ln(1-\xi)-\xi^2}{2} & \text{if } \ |\xi|<1;\\ \ln(2)-1/2 & \text{if } \ |\xi|=1;\\ +\infty, & \text{if } \ |\xi|>1. \end{cases} \tag{2.11}$$

Proof. Let $\xi\in\left]-1,1\right[=\operatorname{dom}\nabla\phi=\operatorname{ran}\operatorname{prox}_\phi$. Then $\xi+\phi'(\xi)=\operatorname{arctanh}(\xi)$ and therefore $\varrho=(\operatorname{Id}+\phi')^{-1}=\tanh$. ☐

Example 2.13 The unimodal sigmoid activation function [30]

$$\varrho\colon \mathbb{R}\to\mathbb{R}\colon \xi\mapsto \frac{1}{1+e^{-\xi}}-\frac{1}{2} \tag{2.12}$$

is the proximity operator of

$$\begin{aligned}\phi\colon &\mathbb{R}\to\left]-\infty,+\infty\right]\\ &\xi\mapsto \begin{cases} (\xi+1/2)\ln(\xi+1/2)+(1/2-\xi)\ln(1/2-\xi)-\dfrac{1}{2}(\xi^2+1/4) & \text{if } \ |\xi|<1/2;\\ -1/4, & \text{if } \ |\xi|=1/2;\\ +\infty, & \text{if } \ |\xi|>1/2. \end{cases}\end{aligned} \tag{2.13}$$

Proof. Let $\xi\in\left]-1/2,1/2\right[=\operatorname{dom}\nabla\phi=\operatorname{ran}\operatorname{prox}_\phi$. Then $\xi+\phi'(\xi)=\ln((1+2\xi)/(1-2\xi))$ and therefore $\operatorname{prox}_\phi=(\operatorname{Id}+\phi')^{-1}\colon \mu\mapsto(1/2)(e^\mu-1)/(e^\mu+1)=1/(1+e^{-\mu})-1/2$. ☐

Remark 2.14 Examples 2.12 and 2.13 are closely related in the sense that the function of (2.12) can be written as $\varrho = (1/2)\tanh(\cdot/2)$.

Example 2.15 The Elliot activation function is [28] $\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \xi/(1+|\xi|)$ can be written as $\varrho = \operatorname{prox}_\phi$, where

$$\begin{aligned}\phi\colon \mathbb{R} &\to \left]-\infty,+\infty\right]\\ \xi &\mapsto \begin{cases} -|\xi| - \ln(1-|\xi|) - \dfrac{\xi^2}{2}, & \text{if } |\xi| < 1;\\ +\infty, & \text{if } |\xi| \geqslant 1.\end{cases}\end{aligned} \tag{2.14}$$

Proof. Let $\xi \in \left]-1,1\right[= \operatorname{dom}\nabla\phi = \operatorname{ran}\operatorname{prox}_\phi$. Then $\xi + \phi'(\xi) = \xi/(1-|\xi|)$ and therefore $\operatorname{prox}_\phi = (\operatorname{Id}+\phi')^{-1}\colon \mu \mapsto \mu/(1+|\mu|)$. ☐

Example 2.16 The inverse hyperbolic sine activation function arcsinh is the proximity operator of $\phi = \cosh - |\cdot|^2/2$.

Proof. Let $\xi \in \mathbb{R}$. Then $\xi + \phi'(\xi) = \sinh\xi$ and therefore $\operatorname{prox}_\phi = (\operatorname{Id}+\phi')^{-1} = \operatorname{arcsinh}$. ☐

Example 2.17 The logarithmic activation function [10]

$$\varrho\colon \mathbb{R} \to \mathbb{R}\colon \xi \mapsto \operatorname{sign}(\xi)\ln\big(1+|\xi|\big) \tag{2.15}$$

is the proximity operator of

$$\phi\colon \mathbb{R} \to \left]-\infty,+\infty\right]\colon \xi \mapsto e^{|\xi|} - |\xi| - 1 - \frac{\xi^2}{2}. \tag{2.16}$$

Proof. We have $\phi'\colon \xi \mapsto \operatorname{sign}(\xi)(e^{|\xi|}-1) - \xi$. Hence $(\operatorname{Id}+\phi')\colon \xi \mapsto \operatorname{sign}(\xi)(e^{|\xi|}-1)$ and, in turn, $\operatorname{prox}_\phi = (\operatorname{Id}+\phi')^{-1}\colon \xi \mapsto \operatorname{sign}(\xi)\ln(1+|\xi|)$. ☐

The class of activation functions $\mathcal{A}(\mathbb{R})$ has interesting stability properties.

Proposition 2.18 *The following hold:*

(i) *Let $\alpha \in \left]0,+\infty\right[$ and $\beta \in \left]0,+\infty\right[$ be such that $\alpha\beta \leqslant 1$, and let $\varrho \in \mathcal{A}(\mathbb{R})$. Then $\alpha\varrho(\beta\cdot) \in \mathcal{A}(\mathbb{R})$.*

(ii) *Let $(\varrho_i)_{i\in I}$ be a finite family in $\mathcal{A}(\mathbb{R})$ and let $(\omega_i)_{i\in I}$ be real numbers in $\left]0,1\right]$ such that $\sum_{i\in I}\omega_i = 1$. Then $\sum_{i\in I}\omega_i\varrho_i \in \mathcal{A}(\mathbb{R})$.*

(iii) *Let $\varrho_1 \in \mathcal{A}(\mathbb{R})$ and $\varrho_2 \in \mathcal{A}(\mathbb{R})$. Then $\varrho_1 \circ \varrho_2 \in \mathcal{A}(\mathbb{R})$.*

(iv) *Let $\varrho \in \mathcal{A}(\mathbb{R})$. Then $\operatorname{Id} - \varrho \in \mathcal{A}(\mathbb{R})$.*

(v) *Let $\varrho_1 \in \mathcal{A}(\mathbb{R})$ and $\varrho_2 \in \mathcal{A}(\mathbb{R})$. Then $(\varrho_1 - \varrho_2 + \operatorname{Id})/2 \in \mathcal{A}(\mathbb{R})$.*

(vi) *Let $\varrho_1 \in \mathcal{A}(\mathbb{R})$ and $\varrho_2 \in \mathcal{A}(\mathbb{R})$. Then $\varrho_1 \circ (2\varrho_2 - \operatorname{Id}) + \operatorname{Id} - \varrho_2 \in \mathcal{A}(\mathbb{R})$.*

Proof. (i)–(iii): This follows at once from Definition 2.2.

(iv)–(v): The fact that the resulting operators are proximity operators is established in [21, Section 3.3]. The fact that they are proximity operators of a function $\phi \in \Gamma_0(\mathcal{H})$ that is minimal at 0 is equivalent to the fact that $\operatorname{prox}_\phi 0 = 0$ Lemma 2.1(i). This identity is easily seen to hold in each instance.

(vi): Set $\varrho = \varrho_1 \circ (2\varrho_2 - \operatorname{Id}) + \operatorname{Id} - \varrho_2$. Then ϱ is firmly nonexpansive [8, Proposition 4.31(ii)]. It is therefore increasing and nonexpansive. Finally, $\varrho(0) = 0$. ☐

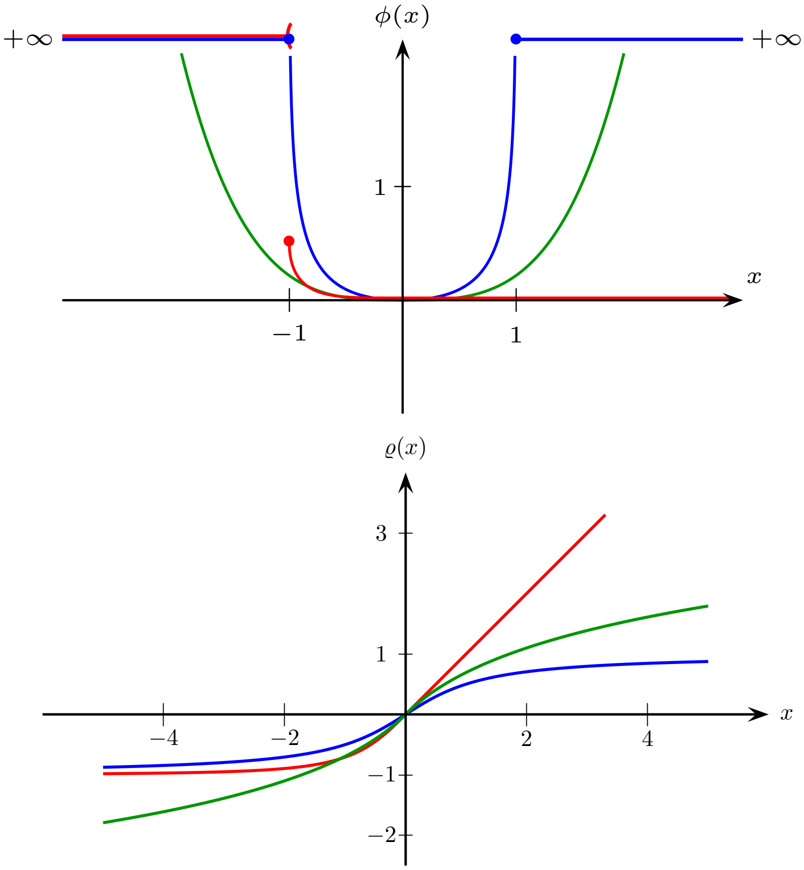

Figure 1: The function ϕ (top) and the corresponding proximal activation function (bottom) ϱ in Proposition 2.3. Example 2.10 is in red, Example 2.11 is in blue, Example 2.17 is in green.

Remark 2.19 Using Proposition 2.18, the above examples can be combined to obtain additional activation functions. For instance, it follows from Example 2.5 and Proposition 2.18(iv) that the soft thresholder

$$\varrho\colon \mathbb{R} \to \mathbb{R} \colon \xi \mapsto \begin{cases} \xi - 1, & \text{if } \ \xi > 1; \\ 0, & \text{if } \ -1 \leqslant \xi \leqslant 1; \\ \xi + 1, & \text{if } \ \xi < -1 \end{cases} \tag{2.17}$$

belongs to $\mathcal{A}(\mathbb{R})$. It was proposed as an activation function in [48].

2.2 Activation operators

In Section 2.1, we have described activation functions which model neuronal activity in terms of a scalar function. In this section, we extend this notion to more general activation operators.

Definition 2.20 Let $\mathcal{H}$ be a real Hilbert space and let $R\colon \mathcal{H} \to \mathcal{H}$. Then R belongs to the class $\mathcal{A}(\mathcal{H})$ if there exists a function $\varphi \in \Gamma_0(\mathcal{H})$ which is minimal at the zero vector and such that $R = \operatorname{prox}_\varphi$.

Property (ii) below shows that activation operators in $\mathcal{A}(\mathcal{H})$ have strong stability properties. On the other hand, the boundedness property (iv) is important in neural network-based functional approximation [26, 29].

Proposition 2.21 *Let $\mathcal{H}$ be a real Hilbert space and let $R \in \mathcal{A}(\mathcal{H})$. Then the following hold:*

(i) $R0=0$.

(ii) *Let x and y be in $\mathcal{H}$. Then* $\|Rx-Ry\|^2 \leqslant \|x-y\|^2 - \|x-y-Rx+Ry\|^2$.

(iii) *Let $x \in \mathcal{H}$. Then* $\|Rx\| \leqslant \|x\|$.

(iv) *Let $\varphi \in \Gamma_0(\mathcal{H})$ be such that $R = \operatorname{prox}_\varphi$. Then $\operatorname{ran} R$ is bounded if and only if $\operatorname{dom}\varphi$ is bounded.*

Proof. (i): This follows from Lemma 2.1(i).

(ii): This follows from the firm nonexpansiveness of proximity operators [8, Proposition 12.28].

(iii): Set $y=0$ in (ii) and use (i).

(iv): We have $\operatorname{ran} R = \operatorname{ran}(\mathrm{Id}+\partial\varphi)^{-1} = \operatorname{dom}(\mathrm{Id}+\partial\varphi) = \operatorname{dom}\partial\varphi$. On the other hand, $\operatorname{dom}\partial\varphi$ is a dense subset of $\operatorname{dom}\varphi$ [8, Corollary 16.39]. ☐

Proposition 2.22 *Let $\mathcal{H}$ and $\mathcal{G}$ be real Hilbert spaces. Then the following hold:*

(i) *Let $L \in \mathcal{B}(\mathcal{H},\mathcal{G})$ be such that $\|L\| \leqslant 1$ and let $R \in \mathcal{A}(\mathcal{H})$. Then $L^* \circ R \circ L \in \mathcal{A}(\mathcal{H})$.*

(ii) *Let $(R_i)_{i\in I}$ be a finite family in $\mathcal{A}(\mathcal{H})$ and let $(\omega_i)_{i\in I}$ be real numbers in $]0,1]$ such that $\sum_{i\in I}\omega_i = 1$. Then $\sum_{i\in I}\omega_i R_i \in \mathcal{A}(\mathcal{H})$.*

(iii) *Let $R \in \mathcal{A}(\mathcal{H})$. Then $\mathrm{Id}-R \in \mathcal{A}(\mathcal{H})$.*

(iv) *Let $R_1 \in \mathcal{A}(\mathcal{H})$ and $R_2 \in \mathcal{A}(\mathcal{H})$. Then $(R_1 - R_2 + \mathrm{Id})/2 \in \mathcal{A}(\mathcal{H})$.*

Proof. The fact that the resulting operators are proximity operators is established in [21, Section 3.3]. In addition, 0 is clearly a fixed point of the resulting operators. In view of Lemma 2.1(i), the proof is complete. ☐

Example 2.23 The softmax activation operator [15] is

$$R\colon \mathbb{R}^N \to \mathbb{R}^N \colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto \left(\exp(\xi_k) \Bigg/ \sum_{j=1}^{N}\exp(\xi_j)\right)_{1\leqslant k\leqslant N} - u, \tag{2.18}$$

where $u=(1,\ldots,1)/N \in \mathbb{R}^N$. We have $R=\operatorname{prox}_\varphi$, where $\varphi = \psi(\cdot+u) + \langle\cdot \mid u\rangle$ and

$$\begin{aligned}\psi\colon \mathbb{R}^N &\to \left]-\infty,+\infty\right]\\ (\xi_k)_{1\leqslant k\leqslant N} &\mapsto \begin{cases}\displaystyle\sum_{k=1}^{N}\left(\xi_k\ln\xi_k - \frac{\xi_k^2}{2}\right), & \text{if } (\xi_k)_{1\leqslant i\leqslant N}\in[0,1]^N \text{ and } \displaystyle\sum_{k=1}^{N}\xi_k=1;\\ +\infty, & \text{otherwise,}\end{cases}\end{aligned} \tag{2.19}$$

with the convention $0\ln 0 = 0$.

Proof. Set

$$\begin{aligned}g\colon \mathbb{R}^N &\to \left]-\infty,+\infty\right]\\ (\xi_k)_{1\leqslant k\leqslant N} &\mapsto \begin{cases}\displaystyle\sum_{k=1}^{N}\xi_k\ln\xi_k, & \text{if } (\xi_k)_{1\leqslant k\leqslant N}\in[0,1]^N \text{ and } \displaystyle\sum_{k=1}^{N}\xi_k=1;\\ +\infty, & \text{otherwise.}\end{cases}\end{aligned} \tag{2.20}$$

Then $\psi = g - \|\cdot\|^2/2$ and [40, Section 16] asserts that

$$g^*\colon \mathbb{R}^N \to \mathbb{R}\colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto \ln\left(\sum_{k=1}^{N} \exp(\xi_k)\right). \tag{2.21}$$

Since $\nabla g^* = R + u$, according to Lemma 2.1(ii), $R = \operatorname{prox}_\psi - u$. We complete the proof by invoking the shift properties of proximity operators [8, Proposition 24.8(iii)]. ☐

Separable activation operators supply another important instance of activation operators.

Proposition 2.24 *Let $\mathcal{H}$ be a separable real Hilbert space, let $(e_k)_{k\in\mathbb{K}\subset\mathbb{N}}$ be an orthonormal basis of $\mathcal{H}$, and let $(\phi_k)_{k\in\mathbb{K}}$ be a family of functions in $\Gamma_0(\mathbb{R})$ such that $(\forall k\in\mathbb{K})\ \phi_k \geqslant \phi_k(0) = 0$. Define*

$$R\colon \mathcal{H} \to \mathcal{H}\colon x \mapsto \sum_{k\in\mathbb{K}} \big(\operatorname{prox}_{\phi_k}\langle x \mid e_k\rangle\big) e_k. \tag{2.22}$$

Then $R \in \mathcal{A}(\mathcal{H})$.

Proof. The fact that R is the proximity operator of the $\Gamma_0(\mathcal{H})$ function $\varphi\colon x \mapsto \sum_{k\in\mathbb{K}} \phi_k(\langle x \mid e_k\rangle)$ is established in [23, Example 2.19]. In addition, it is clear that φ is minimal at 0. ☐

3 Compositions of firmly nonexpansive and affine operators

Our analysis will revolve around the following property for a family of linear operators $(W_i)_{1\leqslant i\leqslant m+1}$.

Condition 3.1 Let $m \geqslant 0$ be an integer, let $(\mathcal{H}_i)_{0\leqslant i\leqslant m}$ be real Hilbert spaces, set $\mathcal{H}_{m+1} = \mathcal{H}_0$, and let $\alpha \in [1/2, 1]$. For every $i \in \{1,\ldots,m+1\}$, let $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$ and set

$$L_i\colon \mathcal{H}_0 \times \cdots \times \mathcal{H}_{i-1} \to \mathcal{H}_i\colon (x_k)_{0\leqslant k\leqslant i-1} \mapsto \sum_{k=0}^{i-1} (W_i \circ \cdots \circ W_{k+1}) x_k. \tag{3.1}$$

It is required that, for every $\boldsymbol{x} = (x_i)_{0\leqslant i\leqslant m} \in \mathcal{H}_0 \times \cdots \times \mathcal{H}_m$ such that

$$(\forall i \in \{0,\ldots,m\}) \quad \|x_i\| \leqslant \begin{cases} 1, & \text{if } i = 0;\\ \|L_i(x_0,\ldots,x_{i-1})\|, & \text{if } i \geqslant 1, \end{cases} \tag{3.2}$$

there holds

$$\|L_{m+1}\boldsymbol{x} - 2^{m+1}(1-\alpha)x_0\| + \|L_{m+1}\boldsymbol{x}\| \leqslant 2^{m+1}\alpha\|x_0\|. \tag{3.3}$$

Remark 3.2 In Condition 3.1, we take $\alpha \geqslant 1/2$ because, if $\boldsymbol{x} = (x_i)_{0\leqslant i\leqslant m} \in (\mathcal{H}_0\smallsetminus\{0\})\times\mathcal{H}_1\times\cdots\times\mathcal{H}_m$ satisfies (3.3), then $2^{m+1}(1-\alpha)\|x_0\| \leqslant \|L_{m+1}\boldsymbol{x} - 2^{m+1}(1-\alpha)x_0\| + \|L_{m+1}\boldsymbol{x}\| \leqslant 2^{m+1}\alpha\|x_0\|$.

We establish some preliminary results before providing properties that imply Condition 3.1.

Lemma 3.3 *Let $m \geqslant 1$ be an integer, let $(\mathcal{H}_i)_{0\leqslant i\leqslant m}$ be real Hilbert spaces, and set $\theta_0 = 1$. For every $i \in \{1,\ldots,m\}$, let $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$ and set*

$$\begin{aligned}\theta_i = \|W_i \circ \cdots \circ W_1\|\\ + \sum_{k=1}^{i-1} \sum_{1\leqslant j_1<\ldots<j_k\leqslant i-1} \|W_i \circ \cdots \circ W_{j_k+1}\|\,\|W_{j_k} \circ \cdots \circ W_{j_{k-1}+1}\| \cdots \|W_{j_1} \circ \cdots \circ W_1\|.\end{aligned} \tag{3.4}$$

Let $(x_i)_{0\leqslant i\leqslant m} \in \mathcal{H}_0 \times \cdots \times \mathcal{H}_m$ be such that (3.2) is satisfied. Then the following hold:

(i) $(\forall i \in \{1,\ldots,m\})\ \theta_i = \sum_{k=0}^{i-1} \theta_k \|W_i \circ \cdots \circ W_{k+1}\|$.

(ii) $(\forall i \in \{1,\ldots,m\})\ \|x_i\| \leqslant \theta_i \|x_0\|$.

Proof. (i): This follows recursively from (3.4).

(ii): For every $i \in \{1,\ldots,m\}$, let L_i be as in (3.1). We proceed by induction on m. We first observe that the inequality is satisfied if $m = 1$ since $\|x_1\| \leqslant \|L_1 x_0\| = \|W_1 x_0\| \leqslant \|W_1\|\,\|x_0\| = \theta_1 \|x_0\|$. Now assume that $m \geqslant 2$ and that the inequalities hold for $(x_1,\ldots,x_{m-1})$. Then, since (i) yields

$$\theta_m = \|W_m \circ \cdots \circ W_1\| + \sum_{k=1}^{m-1} \theta_k \|W_m \circ \cdots \circ W_{k+1}\|, \tag{3.5}$$

we obtain

$$\begin{aligned}
\|x_m\| \leqslant \|L_m(x_0,\ldots,x_{m-1})\| &= \Bigg\| \sum_{k=0}^{m-1} (W_m \circ \cdots \circ W_{k+1}) x_k \Bigg\| \\
&\leqslant \sum_{k=0}^{m-1} \|W_m \circ \cdots \circ W_{k+1}\|\,\|x_k\| \\
&\leqslant \Bigg(\|W_m \circ \cdots \circ W_1\| + \sum_{k=1}^{m-1} \theta_k \|W_m \circ \cdots \circ W_{k+1}\| \Bigg) \|x_0\| \\
&= \theta_m \|x_0\|,
\end{aligned} \tag{3.6}$$

which conclude the proof. □

Lemma 3.4 *Let $\mathcal{H}$ be a real Hilbert space, and let x and y be in $\mathcal{H}$. Then*

$$\|x\|\,\|y\| - \langle x \mid y \rangle \leqslant (\|x\| + \|y\| - \|x+y\|)(\|x\| + \|y\|). \tag{3.7}$$

Proof. Since $\|x+y\|^2 - 2\|x+y\|(\|x\|+\|y\|) + (\|x\|+\|y\|)^2 \geqslant 0$, we have

$$\begin{aligned}
&\|x\|^2 + \|y\|^2 + \langle x \mid y \rangle + \|x\|\,\|y\| \\
&\quad = \|x\|^2 + \|y\|^2 + \frac{\|x+y\|^2 - \|x\|^2 - \|y\|^2}{2} + \frac{(\|x\|+\|y\|)^2 - \|x\|^2 - \|y\|^2}{2} \\
&\quad = \frac{\|x+y\|^2 + (\|x\|+\|y\|)^2}{2} \\
&\quad \geqslant \|x+y\|(\|x\|+\|y\|),
\end{aligned} \tag{3.8}$$

as claimed. □

Notation 3.5 Let $m \geqslant 0$ be an integer, and let $(\mathcal{H}_i)_{0 \leqslant i \leqslant m}$ be real Hilbert spaces. Let $\boldsymbol{\mathcal{X}}$ be the standard vector space $\mathcal{H}_0 \times \cdots \times \mathcal{H}_m$ equipped with the norm $\|\cdot\|_{\boldsymbol{\mathcal{X}}} \colon \boldsymbol{x} = (x_i)_{0 \leqslant i \leqslant m} \mapsto \max_{0 \leqslant i \leqslant m} \|x_i\|$ and let $\boldsymbol{\mathcal{Y}}$ be the standard vector space $\mathcal{H}_0 \times \mathcal{H}_0$ equipped with the norm $\|\cdot\|_{\boldsymbol{\mathcal{Y}}} \colon \boldsymbol{y} = (y_1, y_2) \mapsto \|y_1\| + \|y_2\|$. Henceforth, the norm of $\boldsymbol{M} \in \mathcal{B}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ is denoted by $\|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}}$.

Proposition 3.6 *Let $m \geqslant 0$ be an integer, let $(\mathcal{H}_i)_{0 \leqslant i \leqslant m}$ be nonzero real Hilbert spaces, set $\mathcal{H}_{m+1} = \mathcal{H}_0$, and use Notation 3.5. For every $i \in \{1,\ldots,m+1\}$, let $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$. Further, let $\alpha \in [1/2, 1]$, let*

$\theta_0 = 1$, let $(\theta_i)_{1\leqslant i\leqslant m+1}$ be as in (3.4), and set

$$\begin{cases} W = W_{m+1}\circ\cdots\circ W_1 & (3.9a)\\ \mu = \inf\limits_{x\in\mathcal{H}_0,\,\|x\|=1} \langle Wx \mid x\rangle & (3.9b)\\ M\colon \boldsymbol{\mathcal{X}}\to\mathcal{H}_0\colon \boldsymbol{x}\mapsto \sum\limits_{i=0}^{m}\theta_i(W_{m+1}\circ\cdots\circ W_{i+1})x_i & (3.9c)\\ \boldsymbol{M}\colon \boldsymbol{\mathcal{X}}\to\boldsymbol{\mathcal{Y}}\colon \boldsymbol{x}\mapsto \dfrac{1}{2^{m+1}\alpha}\big(M\boldsymbol{x}-2^{m+1}(1-\alpha)x_0, M\boldsymbol{x}\big). & (3.9d)\end{cases}$$

Suppose that one of the following holds:

(i) *There exists $i\in\{1,\ldots,m+1\}$ such that $W_i = 0$.*

(ii) $\|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}}\leqslant 1$.

(iii) $\|W-2^{m+1}(1-\alpha)\,\mathrm{Id}\,\|-\|W\|+2\theta_{m+1}\leqslant 2^{m+1}\alpha$.

(iv) *$\alpha\neq 1$, for every $i\in\{1,\ldots,m+1\}$ $W_i\neq 0$, and there exists $\eta\in[0,\alpha/((1-\alpha)\theta_{m+1})]$ such that*

$$\begin{cases}\theta_{m+1}\leqslant 2^{m+1}\alpha\\ \alpha\theta_{m+1}+(1-\alpha)(\|\,\mathrm{Id}-\eta W\|-\eta\|W\|)(\theta_{m+1}-\|W\|)\leqslant 2^m(2\alpha-1)+(1-\alpha)\mu.\end{cases} \tag{3.10}$$

Then $(W_i)_{1\leqslant i\leqslant m+1}$ satisfies Condition 3.1.

Proof. We use the operators $(L_i)_{1\leqslant i\leqslant m+1}$ introduced in Condition 3.1. Per Notation 3.5 and (3.9d),

$$\sup_{\substack{\boldsymbol{y}\in\boldsymbol{\mathcal{X}}\\ \max\limits_{0\leqslant i\leqslant m}\|y_i\|\leqslant 1}} \frac{\|M\boldsymbol{y}-2^{m+1}(1-\alpha)y_0\|+\|M\boldsymbol{y}\|}{2^{m+1}\alpha} = \sup_{\substack{\boldsymbol{y}\in\boldsymbol{\mathcal{X}}\\ \|\boldsymbol{y}\|_{\boldsymbol{\mathcal{X}}}\leqslant 1}} \|\boldsymbol{M}\boldsymbol{y}\|_{\boldsymbol{\mathcal{Y}}} = \|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}} \tag{3.11}$$

and therefore

$$(\forall \boldsymbol{y}\in\boldsymbol{\mathcal{X}})\quad \max_{0\leqslant i\leqslant m}\|y_i\|\leqslant 1 \quad\Rightarrow\quad \|M\boldsymbol{y}-2^{m+1}(1-\alpha)y_0\|+\|M\boldsymbol{y}\|\leqslant 2^{m+1}\alpha\|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}}. \tag{3.12}$$

Now let $\boldsymbol{x}\in\boldsymbol{\mathcal{X}}$ be such that

$$(\forall i\in\{0,\ldots,m\})\quad \|x_i\|\leqslant\begin{cases}1, & \text{if } i=0;\\ \|L_i(x_0,\ldots,x_{i-1})\|, & \text{if } i\geqslant 1.\end{cases} \tag{3.13}$$

(i): We assume that $m\geqslant 1$. For every $k\in\{i,\ldots,m\}$, it follows from (3.4) that $\theta_k = 0$ and in turn from Lemma 3.3(ii) and (3.13) that $x_k = 0$. Therefore,

$$L_{m+1}\boldsymbol{x} = \sum_{k=0}^{m}(W_{m+1}\circ\cdots\circ W_{k+1})x_k = \sum_{k=0}^{i-1}(W_{m+1}\circ\cdots\circ W_{k+1})x_k = 0, \tag{3.14}$$

and (3.3) clearly holds.

(ii): In view of (i), we assume that, if $m\geqslant 1$, $(\forall i\in\{1,\ldots,m\})$ $W_i\neq 0$. We then derive from (3.4) that $(\forall i\in\{1,\ldots,m\})$ $\theta_i\geqslant\prod_{k=1}^{i}\|W_k\|>0$. If $x_0 = 0$, (3.3) trivially follows from Lemma 3.3(ii), we therefore assume otherwise. Now set

$$(\forall i\in\{0,\ldots,m\})\quad y_i = \frac{x_i}{\theta_i\|x_0\|}. \tag{3.15}$$

According to Lemma 3.3(ii), $(\forall i \in \{0,\dots,m\})\ \|y_i\| \leqslant 1$. On the other hand, it follows from (3.9c), (3.15), and (3.1) that $M\boldsymbol{y} = L_{m+1}\boldsymbol{x}/\|x_0\|$. Altogether, we deduce from (3.12) that (3.3) holds.

(iii)⇒(ii): Take $\boldsymbol{y} \in \boldsymbol{\mathcal{X}}$ such that $\|\boldsymbol{y}\|_{\boldsymbol{\mathcal{X}}} \leqslant 1$. Then it follows from (3.9c) and Lemma 3.3(i) that

$$\begin{aligned}
&\|M\boldsymbol{y} - 2^{m+1}(1-\alpha)y_0\| + \|M\boldsymbol{y}\| \\
&\quad\leqslant \|W - 2^{m+1}(1-\alpha)\,\mathrm{Id}\,\|\,\|y_0\| + \|W\|\,\|y_0\| + 2\sum_{i=1}^{m}\theta_i\|W_{m+1}\circ\cdots\circ W_{i+1}\|\,\|y_i\| \\
&\quad\leqslant \|W - 2^{m+1}(1-\alpha)\,\mathrm{Id}\,\| - \|W\| + 2\theta_{m+1} \\
&\quad\leqslant 2^{m+1}\alpha.
\end{aligned}\tag{3.16}$$

In turn, (3.11) yields $\|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}} \leqslant 1$.

(iv)⇒(ii): Let $\boldsymbol{y} = (y_0,\dots,y_m) \in \boldsymbol{\mathcal{X}}$ be such that $\|y_0\| = \cdots = \|y_m\| = 1$, and set

$$u = \begin{cases} \displaystyle\sum_{i=1}^{m}\theta_i(W_{m+1}\circ\cdots\circ W_{i+1})y_i, & \text{if } \ m \neq 0; \\ 0, & \text{if } \ m = 0. \end{cases}\tag{3.17}$$

The assumptions and (3.9b) imply that

$$\begin{cases} \eta\theta_{m+1} \leqslant \alpha/(1-\alpha) \\ \theta_{m+1} \leqslant 2^{m+1}\alpha \\ \alpha\theta_{m+1} + (1-\alpha)(\|\,\mathrm{Id} - \eta W\| - \eta\|W\|)(\theta_{m+1} - \|W\|) \\ \qquad\qquad\qquad\qquad \leqslant 2^m(2\alpha-1) + (1-\alpha)\langle Wy_0 \mid y_0\rangle. \end{cases}\tag{3.18}$$

On the other hand,

$$\begin{aligned}
&\alpha\|Wy_0 + u\| - (1-\alpha)\langle y_0 \mid u\rangle \\
&\quad= \alpha\|Wy_0 + u\| - (1-\alpha)\langle \eta Wy_0 + (\mathrm{Id} - \eta W)y_0 \mid u\rangle \\
&\quad\leqslant \alpha\|Wy_0 + u\| - \eta(1-\alpha)\langle Wy_0 \mid u\rangle + (1-\alpha)\|(\mathrm{Id} - \eta W)y_0\|\,\|u\|.
\end{aligned}\tag{3.19}$$

Since, by Lemma 3.3(i) and (3.18),

$$\eta\sum_{i=0}^{m}\theta_i\|W_{m+1}\circ\cdots\circ W_{i+1}\| = \eta\theta_{m+1} \leqslant \frac{\alpha}{1-\alpha},\tag{3.20}$$

we deduce from (3.17) that

$$\eta(1-\alpha)(\|Wy_0\| + \|u\|) \leqslant \alpha.\tag{3.21}$$

However, by Lemma 3.4,

$$\|Wy_0\|\,\|u\| - \langle Wy_0 \mid u\rangle \leqslant (\|Wy_0\| + \|u\| - \|Wy_0 + u\|)(\|Wy_0\| + \|u\|).\tag{3.22}$$

In view of (3.21), this yields

$$\eta(1-\alpha)\big(\|Wy_0\|\,\|u\| - \langle Wy_0 \mid u\rangle\big) \leqslant \alpha(\|Wy_0\| + \|u\| - \|Wy_0 + u\|),\tag{3.23}$$

that is,

$$\alpha\|Wy_0 + u\| - \eta(1-\alpha)\langle Wy_0 \mid u\rangle \leqslant \alpha(\|Wy_0\| + \|u\|) - \eta(1-\alpha)\|Wy_0\|\,\|u\|.\tag{3.24}$$

Therefore, since (3.21) implies that $\alpha-\eta(1-\alpha)\|u\|\geqslant 0$, it results from (3.19) that

$$\begin{aligned}
&\alpha\|Wy_0+u\|-(1-\alpha)\langle y_0\mid u\rangle\\
&\qquad\leqslant \alpha(\|Wy_0\|+\|u\|)-\eta(1-\alpha)\|Wy_0\|\,\|u\|+(1-\alpha)\|(\mathrm{Id}-\eta W)y_0\|\,\|u\|\\
&\qquad= \alpha\|u\|+(\alpha-\eta(1-\alpha)\|u\|)\|Wy_0\|+(1-\alpha)\|(\mathrm{Id}-\eta W)y_0\|\,\|u\|\\
&\qquad\leqslant \alpha\|u\|+(\alpha-\eta(1-\alpha)\|u\|)\|W\|+(1-\alpha)\|(\mathrm{Id}-\eta W)y_0\|\,\|u\|\\
&\qquad= \alpha\|W\|+\big(\alpha-\eta(1-\alpha)\|W\|\big)\|u\|+(1-\alpha)\|\mathrm{Id}-\eta W\|\,\|u\|.
\end{aligned}\tag{3.25}$$

However, since (3.20) implies that $\alpha-\eta(1-\alpha)\|W\|\geqslant 0$, while (3.17) implies that $\|u\|\leqslant\theta_{m+1}-\|W\|$, we derive from (3.25) that

$$\begin{aligned}
&\alpha\|Wy_0+u\|-(1-\alpha)\langle y_0\mid u\rangle\\
&\qquad\leqslant \alpha\|W\|+\big(\alpha-\eta(1-\alpha)\|W\|\big)\big(\theta_{m+1}-\|W\|\big)+(1-\alpha)\|\mathrm{Id}-\eta W\|\,(\theta_{m+1}-\|W\|).
\end{aligned}\tag{3.26}$$

We also have

$$\|Wy_0+u\|\leqslant\|W\|+\|u\|\leqslant\theta_{m+1}.\tag{3.27}$$

Hence, using (3.26), (3.27), (3.9c), (3.9a), and (3.9d) we obtain

$$\begin{aligned}
(3.18)\quad&\Rightarrow\quad \begin{cases}\|Wy_0+u\|\leqslant 2^{m+1}\alpha\\ \alpha\|Wy_0+u\|-(1-\alpha)\langle y_0\mid Wy_0+u\rangle\leqslant 2^m(2\alpha-1)\end{cases}\\
&\Leftrightarrow\quad \begin{cases}\|M\boldsymbol{y}\|\leqslant 2^{m+1}\alpha\\ \alpha\|M\boldsymbol{y}\|-(1-\alpha)\langle y_0\mid M\boldsymbol{y}\rangle\leqslant 2^m\big(\alpha^2-(1-\alpha)^2\big)\end{cases}\\
&\Leftrightarrow\quad \begin{cases}\|M\boldsymbol{y}\|\leqslant 2^{m+1}\alpha\\ \|M\boldsymbol{y}-2^{m+1}(1-\alpha)y_0\|^2\leqslant\big(2^{m+1}\alpha-\|M\boldsymbol{y}\|\big)^2\end{cases}\\
&\Leftrightarrow\quad \|M\boldsymbol{y}-2^{m+1}(1-\alpha)y_0\|+\|M\boldsymbol{y}\|\leqslant 2^{m+1}\alpha\\
&\Leftrightarrow\quad \|\boldsymbol{M}\boldsymbol{y}\|_{\boldsymbol{\mathcal{Y}}}\leqslant 1.
\end{aligned}\tag{3.28}$$

Now set $\boldsymbol{C}=\big\{\boldsymbol{y}\in\boldsymbol{\mathcal{X}}\;\big|\;\|y_0\|=\cdots=\|y_m\|=1\big\}$. Then, in view of (3.11), (3.28), and [8, Proposition 11.1(ii)], we conclude that $\|\boldsymbol{M}\|_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}}=\sup_{\boldsymbol{y}\in\mathrm{conv}\,\boldsymbol{C}}\|\boldsymbol{M}\boldsymbol{y}\|_{\boldsymbol{\mathcal{Y}}}=\sup_{\boldsymbol{y}\in\boldsymbol{C}}\|\boldsymbol{M}\boldsymbol{y}\|_{\boldsymbol{\mathcal{Y}}}\leqslant 1$. ☐

The next result establishes a link between deep neural network structures and the operators introduced in (3.1).

Lemma 3.7 *Let $m\geqslant 1$ be an integer and let $(\mathcal{H}_i)_{0\leqslant i\leqslant m+1}$ be nonzero real Hilbert spaces. For every $i\in\{1,\ldots,m+1\}$, let $W_i\in\mathcal{B}(\mathcal{H}_{i-1},\mathcal{H}_i)$ and let L_i be as in (3.1). Further, for every $i\in\{1,\ldots,m\}$, let $P_i\colon\mathcal{H}_i\to\mathcal{H}_i$ be firmly nonexpansive. Set*

$$T_m=W_{m+1}\circ P_m\circ W_m\circ\cdots\circ P_1\circ W_1,\tag{3.29}$$

let x and y be distinct points in $\mathcal{H}_0$, and set $v_0=(x-y)/\|x-y\|$. Then there exists $(v_1,\ldots,v_m)\in\mathcal{H}_1\times\cdots\times\mathcal{H}_m$ such that

$$\begin{cases}(\forall i\in\{1,\ldots,m\})\quad \|v_i\|\leqslant\|L_i(v_0,\ldots,v_{i-1})\|\\ \dfrac{2^m(T_mx-T_my)}{\|x-y\|}=L_{m+1}(v_0,\ldots,v_m).\end{cases}\tag{3.30}$$

Proof. For every $i \in \{1,\dots,m\}$, since P_i is firmly nonexpansive, there exists a nonexpansive operator $Q_i\colon \mathcal{H}_i \to \mathcal{H}_i$ such that

$$P_i = \frac{\operatorname{Id} + Q_i}{2}. \tag{3.31}$$

We proceed by induction on m. Suppose that $m = 1$ and set

$$v_1 = \frac{Q_1(W_1x) - Q_1(W_1y)}{\|x-y\|}, \tag{3.32}$$

which implies that $\|v_1\| \leqslant \|W_1(x-y)\|/\|x-y\| = \|L_1v_0\|$. Then

$$\begin{aligned} 2(T_1x - T_1y) &= (W_2 \circ W_1)(x-y) + (W_2 \circ Q_1 \circ W_1)x - (W_2 \circ Q_1 \circ W_1)y \\ &= \|x-y\|\big((W_2 \circ W_1)v_0 + W_2v_1\big). \end{aligned} \tag{3.33}$$

Thus, (3.30) holds for $m = 1$. Next, we assume that $m > 1$ and that there exists $(v_1,\dots,v_{m-1}) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_{m-1}$ such that

$$\begin{cases} (\forall i \in \{1,\dots,m-1\}) \quad \|v_i\| \leqslant \|L_i(v_0,\dots,v_{i-1})\| \\ \dfrac{2^{m-1}\big(T_{m-1}x - T_{m-1}y\big)}{\|x-y\|} = L_m(v_0,\dots,v_{m-1}), \end{cases} \tag{3.34}$$

and we set

$$v_m = \frac{2^{m-1}\big((Q_m \circ T_{m-1})x - (Q_m \circ T_{m-1})y\big)}{\|x-y\|}. \tag{3.35}$$

Then (3.29), (3.31), and (3.34) yield

$$\begin{aligned} &T_mx - T_my \\ &\quad= \frac{(W_{m+1} \circ T_{m-1})x - (W_{m+1} \circ T_{m-1})y}{2} + \frac{(W_{m+1} \circ Q_m \circ T_{m-1})x - (W_{m+1} \circ Q_m \circ T_{m-1})y}{2} \\ &\quad= \frac{\|x-y\|}{2^m}\big((W_{m+1} \circ L_m)(v_0,\dots,v_{m-1}) + W_{m+1}v_m\big) \\ &\quad= \frac{\|x-y\|}{2^m} L_{m+1}(v_0,\dots,v_m). \end{aligned} \tag{3.36}$$

In addition, it follows from (3.34) and (3.35) that

$$\|v_m\| \leqslant \frac{2^{m-1}\|T_{m-1}x - T_{m-1}y\|}{\|x-y\|} = \|L_m(v_0,\dots,v_{m-1})\|, \tag{3.37}$$

which completes the proof. $\square$

We now establish connections between Condition 3.1 for linear operators and the concept of averagedness for composite nonlinear operators.

Theorem 3.8 *Let $m \geqslant 1$ be an integer, let $(\mathcal{H}_i)_{0 \leqslant i \leqslant m-1}$ be nonzero real Hilbert spaces, set $\mathcal{H}_m = \mathcal{H}_0$, and let $\alpha \in [1/2, 1]$. For every $i \in \{1,\dots,m\}$, let $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$ and let $P_i\colon \mathcal{H}_i \to \mathcal{H}_i$ be firmly nonexpansive. Suppose that $(W_i)_{1 \leqslant i \leqslant m}$ satisfies Condition 3.1. Then $P_m \circ W_m \circ \cdots \circ P_1 \circ W_1$ is α-averaged.*

Proof. Set $T = P_m \circ W_m \circ \cdots \circ P_1 \circ W_1$. We must show that

$$Q = \left(1 - \frac{1}{\alpha}\right) \operatorname{Id} + \frac{1}{\alpha} T \tag{3.38}$$

is nonexpansive. By assumption, for every $i \in \{1, \ldots, m\}$, there exists a nonexpansive operator $Q_i \colon \mathcal{H}_i \to \mathcal{H}_i$ such that (3.31) holds. Let $(L_i)_{1 \leqslant i \leqslant m}$ be as in (3.1) and let x and y be distinct points in $\mathcal{H}_0$. According to Lemma 3.7, there exists $\boldsymbol{v} = (v_0, \ldots, v_{m-1}) \in \mathcal{H}_0 \times \cdots \times \mathcal{H}_{m-1}$ such that

$$\begin{cases} v_0 = \dfrac{x - y}{\|x - y\|} \\ (\forall i \in \{1, \ldots, m-1\}) \quad \|v_i\| \leqslant \|L_i(v_0, \ldots, v_{i-1})\| \\ \dfrac{2^{m-1}\big((W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y\big)}{\|x - y\|} = L_m \boldsymbol{v}. \end{cases} \tag{3.39}$$

Condition 3.1 imposes that

$$\|L_m \boldsymbol{v} - 2^m (1 - \alpha) v_0\| + \|L_m \boldsymbol{v}\| \leqslant 2^m \alpha \|v_0\| = 2^m \alpha, \tag{3.40}$$

which is equivalent to

$$\begin{aligned} &\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y - 2(1 - \alpha)(x - y)\| \\ &\qquad + \|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y\| \leqslant 2\alpha \|x - y\|. \end{aligned} \tag{3.41}$$

In turn, we derive from (3.38) and (3.31) that

$$\begin{aligned} &\|Qx - Qy\| \\ &\leqslant \frac{1}{\alpha} \left\| \left(\frac{\operatorname{Id} + Q_m}{2} \circ W_m \circ \cdots \circ P_1 \circ W_1\right)x - \left(\frac{\operatorname{Id} + Q_m}{2} \circ W_m \circ \cdots \circ P_1 \circ W_1\right)y - (1 - \alpha)(x - y) \right\| \\ &\leqslant \frac{1}{2\alpha} \Big(\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y - 2(1 - \alpha)(x - y)\| \\ &\quad + \|(Q_m \circ W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (Q_m \circ W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y\| \Big) \\ &\leqslant \frac{1}{2\alpha} \Big(\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y - 2(1 - \alpha)(x - y)\| \\ &\quad + \|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \cdots \circ P_1 \circ W_1)y\| \Big) \\ &\leqslant \|x - y\|, \end{aligned} \tag{3.42}$$

which establishes the nonexpansiveness of Q. ☐

Example 3.9 Consider Theorem 3.8 with $m = 2$. In view of Proposition 3.6(iii), $P_2 \circ W_2 \circ P_1 \circ W_1$ is α-averaged if $\|W_2 \circ W_1 - 4(1 - \alpha) \operatorname{Id}\| + \|W_2 \circ W_1\| + 2\|W_2\| \, \|W_1\| \leqslant 4\alpha$. In particular, if $\alpha = 1$, this condition is obviously less restrictive than requiring that W_1 and W_2 be nonexpansive.

4 A variational inequality model

In this section, we first investigate an autonomous version of Model 1.1.

Model 4.1 This is the special case of Model 1.1 in which, for every $i \in \{1,\ldots,m\}$, there exist $R_i \in \mathcal{A}(\mathcal{H}_i)$, say $R_i = \operatorname{prox}_{\varphi_i}$ for some $\varphi_i \in \Gamma_0(\mathcal{H}_i)$ with $\varphi_i(0) = \inf \varphi_i(\mathcal{H}_i)$, $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$, and $b_i \in \mathcal{H}_i$ such that $(\forall n \in \mathbb{N})$ $R_{i,n} = R_i$, $W_{i,n} = W_i$, $b_{i,n} = b_i$. We set

$$(\forall i \in \{1,\ldots,m\}) \quad T_i \colon \mathcal{H}_{i-1} \to \mathcal{H}_i \colon x \mapsto R_i(W_i x + b_i) \tag{4.1}$$

and

$$\begin{cases} F = \operatorname{Fix}(T_m \circ \cdots \circ T_1) \\ \boldsymbol{\mathcal{H}} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_{m-1} \oplus \mathcal{H}_m \\ \vec{\boldsymbol{\mathcal{H}}} = \mathcal{H}_m \oplus \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_{m-1} \\ \boldsymbol{S} \colon \boldsymbol{\mathcal{H}} \to \vec{\boldsymbol{\mathcal{H}}} \colon (x_1,\ldots,x_{m-1},x_m) \mapsto (x_m,x_1,\ldots,x_{m-1}) \\ \boldsymbol{W} \colon \vec{\boldsymbol{\mathcal{H}}} \to \boldsymbol{\mathcal{H}} \colon (x_m,x_1,\ldots,x_{m-1}) \mapsto (W_1x_m, W_2x_1,\ldots,W_mx_{m-1}) \\ \boldsymbol{\varphi} \colon \boldsymbol{\mathcal{H}} \to \left]-\infty,+\infty\right] \colon \boldsymbol{x} \mapsto \sum_{i=1}^m \varphi_i(x_i) \\ \boldsymbol{\psi} \colon \boldsymbol{\mathcal{H}} \to \left]-\infty,+\infty\right] \colon \boldsymbol{x} \mapsto \sum_{i=1}^m \big(\varphi_i(x_i) - \langle x_i \mid b_i\rangle\big) \\ \boldsymbol{F} = \big\{\boldsymbol{x} \in \boldsymbol{\mathcal{H}} \;\big|\; x_1 = T_1x_m,\ x_2 = T_2x_1,\ldots,\ x_m = T_mx_{m-1}\big\}, \end{cases} \tag{4.2}$$

where $\boldsymbol{x} = (x_1,\ldots,x_m)$ denotes a generic element in $\boldsymbol{\mathcal{H}}$.

4.1 Static analysis

We start with a property of the compositions of the operators $(T_i)_{1\leqslant i\leqslant m}$ of (4.1).

Proposition 4.2 *Consider the setting of Model 4.1, let i and j be integers such that $1 \leqslant j \leqslant i \leqslant m$, and let $x \in \mathcal{H}_{j-1}$. Then*

$$\|(T_i \circ \cdots \circ T_j)x\| \leqslant \|x\| \prod_{k=j}^{i} \|W_k\| + \sum_{q=j}^{i} \left(\|b_q\| \prod_{k=q+1}^{i} \|W_k\| \right). \tag{4.3}$$

Proof. In view of (4.1), the property is satisfied when $i = j$. We now assume that $i > j$. Since $R_i \in \mathcal{A}(\mathcal{H}_i)$, Proposition 2.21(i) yields

$$\begin{aligned} \|(T_i \circ \cdots \circ T_j)x\| &= \|R_i(W_i(T_{i-1} \circ \cdots \circ T_j)x + b_i)\| \\ &= \|R_i(W_i(T_{i-1} \circ \cdots \circ T_j)x + b_i) - R_i0\| \\ &\leqslant \|W_i(T_{i-1} \circ \cdots \circ T_j)x + b_i\| \\ &\leqslant \|W_i\|\,\|(T_{i-1} \circ \cdots \circ T_j)x\| + \|b_i\|. \end{aligned} \tag{4.4}$$

We thus obtain (4.3) recursively. ☐

Next, we establish a connection between Model 4.1 and a variational inequality.

Proposition 4.3 *In the setting of Model 4.1, consider the variational inequality problem*

$$\text{find } \overline{x}_1 \in \mathcal{H}_1,\ldots,\overline{x}_m \in \mathcal{H}_m \text{ such that } \begin{cases} b_1 \in \overline{x}_1 - W_1\overline{x}_m + \partial\varphi_1(\overline{x}_1) \\ b_2 \in \overline{x}_2 - W_2\overline{x}_1 + \partial\varphi_2(\overline{x}_2) \\ \quad\vdots \\ b_m \in \overline{x}_m - W_m\overline{x}_{m-1} + \partial\varphi_m(\overline{x}_m). \end{cases} \tag{4.5}$$

Then the following hold:

(i) *The set of solutions to* (4.5) *is* $\boldsymbol{F}$.

(ii) $\boldsymbol{F} = \operatorname{zer}(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\psi}) = \operatorname{Fix}(\operatorname{prox}_{\boldsymbol{\psi}} \circ \boldsymbol{W} \circ \boldsymbol{S})$.

(iii) $\boldsymbol{F} = \big\{(T_1\overline{x}_m, (T_2 \circ T_1)\overline{x}_m, \ldots, (T_{m-1} \circ \cdots \circ T_1)\overline{x}_m, \overline{x}_m) \;\big|\; \overline{x}_m \in F\big\}$.

(iv) *Suppose that* $(W_i)_{1\leqslant i\leqslant m}$ *satisfies Condition* 3.1 *for some* $\alpha \in [1/2, 1]$. *Then* F *is closed and convex.*

(v) *Suppose that* $(W_i)_{1\leqslant i\leqslant m}$ *satisfies Condition* 3.1 *for some* $\alpha \in [1/2, 1]$ *and that one of the following holds:*

(a) $\operatorname{ran}(T_m \circ \cdots \circ T_1)$ *is bounded.*

(b) *There exists* $j \in \{1, \ldots, m\}$ *such that* $\operatorname{dom}\varphi_j$ *is bounded.*

Then F *and* $\boldsymbol{F}$ *are nonempty.*

(vi) *Suppose that* $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ *is monotone. Then* $\boldsymbol{F}$ *is closed and convex. In addition,* F *and* $\boldsymbol{F}$ *are nonempty if any of the following holds:*

(a) $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi}$ *is surjective.*

(b) $\partial\boldsymbol{\varphi} - \boldsymbol{W} \circ \boldsymbol{S}$ *is maximally monotone.*

(c) $\max_{1\leqslant i\leqslant m}\|W_i\| \leqslant 1$, $\boldsymbol{S}^* - \boldsymbol{W}$ *has closed range, and* $\ker(\boldsymbol{S} - \boldsymbol{W}^*) = \{\mathbf{0}\}$.

(d) $\max_{1\leqslant i\leqslant m}\|W_i\| \leqslant 1$ *and, for every* $i \in \{1, \ldots, m\}$, $\operatorname{dom}\varphi_i^* = \mathcal{H}_i$.

(e) *For every* $i \in \{1, \ldots, m\}$, $\operatorname{dom}\varphi_i = \mathcal{H}$ *and* $\operatorname{dom}\varphi_i^* = \mathcal{H}_i$.

(f) $\boldsymbol{S}^* - \boldsymbol{W}$ *has closed range,* $\ker(\boldsymbol{S} - \boldsymbol{W}^*) = \{\mathbf{0}\}$, *and, for every* $i \in \{1, \ldots, m\}$, $\operatorname{dom}\varphi_i = \mathcal{H}_i$.

(g) *For every* $i \in \{1, \ldots, m\}$, $\operatorname{dom}\varphi_i$ *is bounded.*

Proof. We first observe that $\boldsymbol{S} \in \mathcal{B}(\boldsymbol{\mathcal{H}}, \overrightarrow{\boldsymbol{\mathcal{H}}})$, $\boldsymbol{W} \in \mathcal{B}(\overrightarrow{\boldsymbol{\mathcal{H}}}, \boldsymbol{\mathcal{H}})$, $\boldsymbol{\varphi} \in \Gamma_0(\boldsymbol{\mathcal{H}})$, and $\boldsymbol{\psi} \in \Gamma_0(\boldsymbol{\mathcal{H}})$.

(i): Let $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$. Then

$$\boldsymbol{x} \text{ solves (4.5)} \quad\Leftrightarrow\quad \begin{cases} W_1x_m + b_1 \in x_1 + \partial\varphi_1(x_1) \\ W_2x_1 + b_2 \in x_2 + \partial\varphi_2(x_2) \\ \quad\vdots \\ W_mx_{m-1} + b_m \in x_m + \partial\varphi_m(x_m). \end{cases} \tag{4.6}$$

$$\Leftrightarrow\quad \begin{cases} x_1 = \operatorname{prox}_{\varphi_1}(W_1x_m + b_1) = T_1x_m \\ x_2 = \operatorname{prox}_{\varphi_2}(W_2x_1 + b_2) = T_2x_1 \\ \quad\vdots \\ x_m = \operatorname{prox}_{\varphi_m}(W_mx_{m-1} + b_m) = T_mx_{m-1}. \end{cases} \tag{4.7}$$

(ii): Let $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$. Using (4.2), we obtain

$$\boldsymbol{x} \text{ solves (4.5)} \quad\Leftrightarrow\quad \mathbf{0} \in \boldsymbol{x} - \boldsymbol{W}(\boldsymbol{S}\boldsymbol{x}) + \partial\boldsymbol{\psi}(\boldsymbol{x}) \quad\Leftrightarrow\quad \boldsymbol{x} = \operatorname{prox}_{\boldsymbol{\psi}}\big(\boldsymbol{W}(\boldsymbol{S}\boldsymbol{x})\big). \tag{4.8}$$

(iii): Clear from the definitions of F and $\boldsymbol{F}$.

(iv): Define m firmly nonexpansive operators by $(\forall i \in \{1, \ldots, m\})\; P_i\colon \mathcal{H}_i \to \mathcal{H}_i\colon y \mapsto R_i(y + b_i)$. Then it follows from (4.1) and Theorem 3.8 applied to $(P_i)_{1\leqslant i\leqslant m}$ that $T_m \circ \cdots \circ T_1$ is nonexpansive. In turn, we derive from [8, Corollary 4.24] that its fixed point set F is closed and convex.

(v): Thanks to (iii), it is enough to show that $F \neq \varnothing$. Set $T = T_m \circ \cdots \circ T_1$ and recall that it is nonexpansive by virtue of Theorem 3.8.

(a): Let C be a closed ball such that $\operatorname{ran} T \subset C$ and set $S = T|_C$. Then $S\colon C \to C$ is nonexpansive and therefore [8, Proposition 4.29] asserts that $\operatorname{Fix} T = \operatorname{Fix} S \neq \varnothing$.

(b)⇒(a): We have $\operatorname{ran} T_j \subset \operatorname{ran} R_j = \operatorname{ran} \operatorname{prox}_{\varphi_j} = \operatorname{dom}(\operatorname{Id} + \partial\varphi_j) = \operatorname{dom} \partial\varphi_j \subset \operatorname{dom} \varphi_j$. Hence $\operatorname{ran} T_j$ is bounded and Proposition 4.2 (with $i = m$) implies that

$$\operatorname{ran} T \subset \begin{cases} \operatorname{ran} T_m, & \text{if } \ j = m; \\ (T_m \circ \cdots \circ T_{j+1})(\operatorname{ran} T_j), & \text{if } \ 1 \leqslant j \leqslant m-1 \end{cases} \tag{4.9}$$

is likewise.

(vi): Set $\boldsymbol{A} = \mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\psi}$. Since $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ is monotone and continuous, it is maximally monotone [8, Corollary 20.28], with $\boldsymbol{\mathcal{H}}$ as its domain. Since $\partial\boldsymbol{\psi}$ is also maximally monotone [8, Theorem 20.25], $\boldsymbol{A}$ is likewise [8, Corollary 25.5(i)] and hence $\boldsymbol{F} = \operatorname{zer} \boldsymbol{A}$ is closed and convex [8, Proposition 23.39]. Next, we note that, in view of (iii), $F \neq \varnothing \Leftrightarrow \boldsymbol{F} \neq \varnothing$.

(a): The hypothesis implies that $(b_i)_{1 \leqslant i \leqslant m} \in \operatorname{ran}(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi})$ and therefore that (4.5) has a solution, i.e., $\boldsymbol{F} \neq \varnothing$.

(b)⇒(a): The claim follows from Minty's theorem [8, Theorem 21.1].

(c)⇒(a): We have $\|\boldsymbol{W} \circ \boldsymbol{S}\| = \|\boldsymbol{W}\| = \max_{1 \leqslant i \leqslant m} \|W_i\| \leqslant 1$. Therefore, $-\boldsymbol{W} \circ \boldsymbol{S}$ is nonexpansive, which implies that $(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S})/2$ is firmly nonexpansive [8, Corollary 4.5], that is $(\forall \boldsymbol{x} \in \boldsymbol{\mathcal{H}})$ $\langle \boldsymbol{x} - \boldsymbol{W}(\boldsymbol{S}\boldsymbol{x}) \mid \boldsymbol{x} \rangle \geqslant \|\boldsymbol{x} - \boldsymbol{W}(\boldsymbol{S}\boldsymbol{x})\|^2/2$. Consequently, $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ is 3^* monotone [8, Proposition 25.16], while $\partial\boldsymbol{\varphi}$ is also 3^* monotone [8, Example 25.13]. Finally, since $\boldsymbol{S}$ is unitary,

$$\operatorname{ran}\big(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}\big) = \operatorname{ran}\big(\boldsymbol{S}^* - \boldsymbol{W}\big) = \overline{\operatorname{ran}}\,\big(\boldsymbol{S} - \boldsymbol{W}^*\big)^* = \Big(\ker\big(\boldsymbol{S} - \boldsymbol{W}^*\big)\Big)^{\perp} = \boldsymbol{\mathcal{H}}, \tag{4.10}$$

which shows that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ is surjective. Altogether, since [8, Corollary 25.5(i)] implies that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi}$ is maximally monotone, it follows from [8, Corollary 25.27(i)] that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi}$ is surjective.

(d)⇒(a): We have $\operatorname{dom} \boldsymbol{\varphi}^* = \boldsymbol{\mathcal{H}}$. Hence since $\operatorname{int} \operatorname{dom} \boldsymbol{\varphi}^* \subset \operatorname{dom} \partial\boldsymbol{\varphi}^*$ [8, Proposition 16.27], we have $\operatorname{ran} \partial\boldsymbol{\varphi} = \operatorname{dom} (\partial\boldsymbol{\varphi})^{-1} = \operatorname{dom} \partial\boldsymbol{\varphi}^* = \boldsymbol{\mathcal{H}}$. Hence, $\partial\boldsymbol{\varphi}$ is surjective. We conclude using the same arguments as in (c): $\partial\boldsymbol{\varphi}$ and $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ are both 3^* monotone and their sum is maximally monotone, which allows us to invoke [8, Corollary 25.27(i)].

(e)⇒(a): As seen in (d), $\partial\boldsymbol{\varphi}$ is surjective. We have $\boldsymbol{\mathcal{H}} = \operatorname{int} \operatorname{dom} \boldsymbol{\varphi} \subset \operatorname{dom} \partial\boldsymbol{\varphi}$ [8, Proposition 16.27]. Consequently, $\boldsymbol{\mathcal{H}} = \operatorname{dom}(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}) \subset \operatorname{dom} \partial\boldsymbol{\varphi}$. Altogether, since $\partial\boldsymbol{\varphi}$ is 3^* monotone, it follows from [8, Corollary 25.27(ii)] that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi}$ is surjective.

(f)⇒(a): As seen in (c), $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ is surjective and $\partial\boldsymbol{\varphi}$ is 3^* monotone. In addition, $\operatorname{dom}(\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}) \subset \operatorname{dom} \partial\boldsymbol{\varphi}$ since $\boldsymbol{\mathcal{H}} = \operatorname{int} \operatorname{dom} \boldsymbol{\varphi} \subset \operatorname{dom} \partial\boldsymbol{\varphi}$ [8, Proposition 16.27]. Altogether, it follows from [8, Corollary 25.27(ii)] that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S} + \partial\boldsymbol{\varphi}$ is surjective.

(g): Here $\operatorname{dom} \boldsymbol{A} = \operatorname{dom} \partial\boldsymbol{\varphi} \subset \operatorname{dom} \boldsymbol{\varphi} = \times_{i=1}^{m} \operatorname{dom} \varphi_i$ is bounded. Hence, $\boldsymbol{F} = \operatorname{zer} \boldsymbol{A} \neq \varnothing$ [8, Proposition 23.36(iii)]. ☐

Remark 4.4 In Proposition 4.3(vi), it is required that $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ be monotone, or equivalently, that its self-adjoint part $\mathbf{Id} - (\boldsymbol{W} \circ \boldsymbol{S} + \boldsymbol{S}^* \circ \boldsymbol{W}^*)/2$ be positive. In a finite-dimensional setting, this just means that the eigenvalues of the matrix $\boldsymbol{W}\boldsymbol{S} + \boldsymbol{S}^*\boldsymbol{W}^*$ are in $]{-\infty}, 2]$.

Remark 4.5 Let $\overline{\boldsymbol{x}} \in \boldsymbol{\mathcal{H}}$ be a solution to the variational inequality (4.5). A natural question is whether $\overline{\boldsymbol{x}}$ solves a minimization problem. In general the answer is negative. For instance, for $m \geqslant 3$ layers, even if the Hilbert spaces $(\mathcal{H}_i)_{1 \leqslant i \leqslant m}$ are identical, $\boldsymbol{W} = \mathbf{Id}$, the vectors $(b_i)_{1 \leqslant i \leqslant m}$ are zero, and the functions $(\varphi_i)_{1 \leqslant i \leqslant m}$ are indicator functions of closed convex sets $(C_i)_{1 \leqslant i \leqslant m}$, the solutions to

(4.5) do not minimize any function $\boldsymbol{\Phi}\colon \boldsymbol{\mathcal{H}} \to \mathbb{R}$ [4]. A rather restrictive scenario in which the answer is positive is when $\mathbf{Id} - \boldsymbol{W} \circ \boldsymbol{S}$ is monotone and $\boldsymbol{W} \circ \boldsymbol{S}$ is self-adjoint. Then $\overline{\boldsymbol{x}}$ is a minimizer of $\boldsymbol{\Phi}\colon \boldsymbol{x} \mapsto (1/2)\langle \boldsymbol{x} - \boldsymbol{W}(\boldsymbol{S}\boldsymbol{x}) \mid \boldsymbol{x}\rangle + \boldsymbol{\psi}(\boldsymbol{x})$.

Example 4.6 In Model 4.1, suppose that, for every $i \in \{1,\ldots,m\}$, $\mathcal{H}_i = \mathbb{R}^{N_i}$ for some strictly positive integer N_i. In addition, assume that, for every $i \in \{1,\ldots,m\}$, R_i is a separable activation operator with respect to the canonical basis of $\mathbb{R}^{N_i}$ (see Proposition 2.24), and that it employs the ReLU activation functions of Example 2.6. For every $i \in \{1,\ldots,m\}$, let $x_i = (\xi_{i,k})_{1\leqslant k\leqslant N_i} \in \mathbb{R}^{N_i}$ and set $b_i = (\beta_{i,k})_{1\leqslant k\leqslant N_i}$. Then it follows from Proposition 4.3(i) that $(x_1,\ldots,x_m) \in \boldsymbol{F}$ if and only if, for every $i \in \{1,\ldots,m\}$, $x_i \in [0,+\infty[^{N_i}$ and

$$\begin{cases}(\forall k \in \{1,\ldots,N_1\}) \quad [W_1 x_m]_k + \beta_{1,k} - \xi_{1,k} \in \mathcal{I}(\xi_{1,k})\\ (\forall k \in \{1,\ldots,N_2\}) \quad [W_2 x_1]_k + \beta_{2,k} - \xi_{2,k} \in \mathcal{I}(\xi_{2,k})\\ \quad\vdots\\ (\forall k \in \{1,\ldots,N_{m-1}\}) \quad [W_{m-1} x_{m-2}]_k + \beta_{m-1,k} - \xi_{m-1,k} \in \mathcal{I}(\xi_{m-1,k})\\ (\forall k \in \{1,\ldots,N_m\}) \quad [W_m x_{m-1}]_k + \beta_{m,k} - \xi_{m,k} \in \mathcal{I}(\xi_{m,k})\end{cases} \tag{4.11}$$

where, given $x \in \mathcal{H}_{i-1}$, $[W_i x]_k$ is the kth component of $W_i x$ and

$$(\forall \xi \in [0,+\infty[) \quad \mathcal{I}(\xi) = \begin{cases}\{0\}, & \text{if } \xi \in \left]0,+\infty\right[;\\ \left]-\infty,0\right], & \text{if } \xi = 0.\end{cases} \tag{4.12}$$

Altogether, we conclude that $\boldsymbol{F}$ is a closed convex polyhedron.

4.2 Asymptotic analysis

Next, we investigate the asymptotic behavior of (1.2) in the context of Model 4.1.

Theorem 4.7 *In the setting of Model 4.1, set $T = T_m \circ \cdots \circ T_1$, let $\alpha \in [1/2,1]$, and suppose that the following hold:*

(a) $F \neq \varnothing$.

(b) $(W_i)_{1\leqslant i\leqslant m}$ *satisfies Condition 3.1 with parameter* α.

(c) *One of the following is satisfied:*

  (i) $\lambda_n \equiv 1/\alpha = 1$ *and* $Tx_n - x_n \to 0$.

  (ii) $(\lambda_n)_{n\in\mathbb{N}}$ *lies in* $\left]0,1/\alpha\right[$ *and* $\sum_{n\in\mathbb{N}} \lambda_n(1-\alpha\lambda_n) = +\infty$.

Then $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{x}_m \in F$ and $(T_1\overline{x}_m, (T_2\circ T_1)\overline{x}_m, \ldots, (T_{m-1}\circ\cdots\circ T_1)\overline{x}_m, \overline{x}_m)$ solves (4.5). Now suppose that, in addition, any of the following holds:

(iii) *For every $i \in \{1,\ldots,m-1\}$, R_i is weakly sequentially continuous.*

(iv) *For every $i \in \{1,\ldots,m-1\}$, R_i is a separable activation operator in the sense of Proposition 2.24.*

(v) *For every $i \in \{1,\ldots,m-1\}$, $\mathcal{H}_i$ is finite-dimensional.*

(vi) *For some $\varepsilon \in \left]0,1/2\right[$, $(\lambda_n)_{n\in\mathbb{N}}$ lies in $[\varepsilon,(1-\varepsilon)(\varepsilon+1/\alpha)]$ and, for every $i \in \{1,\ldots,m\}$, $\mathcal{H}_i = \mathcal{H}$ and there exists $\beta_i \in \left]0,1\right[$ such that $\|W_i - 2(1-\beta_i)\,\mathrm{Id}\,\| + \|W_i\| \leqslant 2\beta_i$.*

Then, for every $i \in \{1,\ldots,m-1\}$, $(x_{i,n})_{n\in\mathbb{N}}$ converges weakly to $\overline{x}_i = (T_i\circ\cdots\circ T_1)\overline{x}_m$ and $(\overline{x}_1,\ldots,\overline{x}_m)$ solves (4.5).

Proof. We first derive from (1.2) and Model 4.1 that

$$(\forall n\in\mathbb{N})\quad x_{n+1} = x_n + \lambda_n(Tx_n - x_n). \tag{4.13}$$

Now set $(\forall i\in\{1,\ldots,m\})$ $P_i\colon \mathcal{H}_i\to\mathcal{H}_i\colon y\mapsto R_i(y+b_i)$. Then (4.1) yields $T = P_m\circ W_m\circ\cdots\circ P_1\circ W_1$ and, since the operators $(R_i)_{1\leqslant i\leqslant m}$ are firmly nonexpansive, the operators $(P_i)_{1\leqslant i\leqslant m}$ are likewise. Hence, it follows from (b), Theorem 3.8, and (4.2) that

$$T \text{ is } \alpha\text{-averaged and } \operatorname{Fix} T = F. \tag{4.14}$$

(i): In view of (4.14), T is nonexpansive and hence we derive from [8, Theorem 5.14(i)] that $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in F. The second assertion follows from Proposition 4.3(iii).

(ii): In view of (4.13) and (4.14), [8, Theorem 5.15(iii) and Proposition 5.16(iii)] imply that $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in F, and we conclude by invoking Proposition 4.3(iii).

We now prove the convergence of the individual sequences under each assumption.

(iii): We have already established that $x_n \rightharpoonup \overline{x}_m$. Since W_1 is weakly continuous as a bounded linear operator, so is T_1 in (4.1). Hence, (1.2) implies that $x_{1,n} = T_1x_n \rightharpoonup T_1\overline{x}_m = \overline{x}_1$. Likewise, we obtain successively $x_{2,n} = T_2x_{1,n} \rightharpoonup T_2\overline{x}_1 = \overline{x}_2$, $x_{3,n} = T_3x_{2,n} \rightharpoonup T_3\overline{x}_2 = \overline{x}_3,\ldots$, $x_{m,n} = T_mx_{m-1,n} \rightharpoonup T_m\overline{x}_{m-1} = \overline{x}_m$.

(iv)⇒(iii): See [8, Proposition 24.12(iii)].

(v)⇒(iii): A proximity operator is nonexpansive and therefore continuous, hence weakly continuous in a finite-dimensional setting.

(vi): As shown above, $x_n \rightharpoonup \overline{x}_m \in F$. It follows from Proposition 3.6(iii) and Theorem 3.8 (applied with $m=1$) that, for every $i\in\{1,\ldots,m\}$, T_i is β_i-averaged. Hence, upon applying [24, Theorem 3.5(ii)] with α as an averaging constant of T, we infer that

$$\begin{cases}(\operatorname{Id}-T_1)x_n - (\operatorname{Id}-T_1)\overline{x}_m \to 0\\ (\operatorname{Id}-T_2)(T_1x_n) - (\operatorname{Id}-T_2)(T_1\overline{x}_m)\to 0\\ \qquad\vdots\\ (\operatorname{Id}-T_m)((T_{m-1}\circ\cdots\circ T_1)x_n) - (\operatorname{Id}-T_m)((T_{m-1}\circ\cdots\circ T_1)\overline{x}_m)\to 0.\end{cases} \tag{4.15}$$

Thus, $x_{1,n}-x_n = T_1x_n - x_n \to T_1\overline{x}_m - \overline{x}_m$, which implies that $x_{1,n} = (x_{1,n}-x_n)+x_n \rightharpoonup (T_1\overline{x}_m - \overline{x}_m)+\overline{x}_m = T_1\overline{x}_m$. However, since $x_{2,n}-x_{1,n} = (T_2\circ T_1)x_n - T_1x_n \to (T_2\circ T_1)\overline{x}_m - T_1\overline{x}_m$, we obtain $x_{2,n} \rightharpoonup (T_2\circ T_1)\overline{x}_m$. Continuing this telescoping process yields the claim. ☐

The next result covers the case when the variational inequality problem (4.5) has no solution.

Proposition 4.8 *In the setting of Model 4.1, suppose that $(W_i)_{1\leqslant i\leqslant m}$ satisfies Condition 3.1 with $\alpha\in[1/2,1]$, that $(\lambda_n)_{n\in\mathbb{N}}$ lies in $[\varepsilon,(1/\alpha)-\varepsilon]$, for some $\varepsilon\in\,]0,1/2[$, and that $F=\varnothing$. Then $\|x_n\|\to+\infty$.*

Proof. We derive from (4.13) and (4.14) that, for every $n\in\mathbb{N}$, $x_{n+1} = x_n + \mu_n(Qx_n - x_n)$, where $Q = (1-1/\alpha)\operatorname{Id} + (1/\alpha)T$ is nonexpansive and such that $\operatorname{Fix} Q = F$, and $\mu_n = \alpha\lambda_n \in\,]0,1[$. Hence the claims follows from [8, Proposition 4.29] and [12, Corollary 9(b)]. ☐

Remark 4.9 When assumptions (a)–(c) in Theorem 4.7 are satisfied, the neural network described in Model 1.1 is robust to perturbations of its input. Indeed, since T is α-averaged in (4.13), we can write the updating rule as $x_{n+1} = Q_nx_n$, where Q_n is nonexpansive. In turn, if x_0 and $\widetilde{x}_0$ are two inputs in $\mathcal{H}_0$, for a given $n\in\mathbb{N}$, the resulting outputs x_n and $\widetilde{x}_n$ are such that $\|x_n-\widetilde{x}_n\|\leqslant\|x_0-\widetilde{x}_0\|$.

Remark 4.10 In connection with Theorem 4.7 and Remark 4.5, let us underline that in general the weak limit $\overline{x}_m$ of $(x_n)_{n\in\mathbb{N}}$ does not solve a minimization problem. A very special case in which it does is the following. Suppose that $m = 2$, $\mathcal{H}_1 = \mathcal{H}$, $\|W_1\| \leqslant 1$, and $W_2 = W_1^*$. Set $\psi_1 = \varphi_1 - \langle \cdot \mid b_1 \rangle$ and $\psi_2 = \varphi_2 - \langle \cdot \mid b_2 \rangle$, and let $\overline{x}_2 \in F$, i.e., $\overline{x}_2 = (\operatorname{prox}_{\psi_2} \circ W_1^* \circ \operatorname{prox}_{\psi_1} \circ W_1)\overline{x}_2$. It follows from [21, Remark 3.10(iv)] that there exists a function $\vartheta \in \Gamma_0(\mathcal{H})$ such that $W_1^* \circ \operatorname{prox}_{\psi_1} \circ W_1 = \operatorname{prox}_{\vartheta}$. Thus, $\overline{x}_2$ is a fixed point of the backward-backward operator $\operatorname{prox}_{\psi_2} \circ \operatorname{prox}_{\vartheta}$. It then follows from [20, Remark 6.13] that $\overline{x}_2$ is a minimizer of ${}^1\vartheta + \psi_2$, where ${}^1\vartheta\colon x \mapsto \inf_{y\in\mathcal{H}}(\vartheta(y) + \|x-y\|^2/2)$ is the Moreau envelope of ϑ.

Remark 4.11 To model closely existing deep neural networks, we have chosen the activation operators in Definition 2.20 and Model 4.1 to be proximity operators. However, as is clear from the results of Section 3 and in particular the central Theorem 3.8, an activation operator $R_i\colon \mathcal{H}_i \to \mathcal{H}_i$ could more generally be a firmly nonexpansive operator that admits 0 as a fixed point. By [8, Corollary 23.9], this means that R_i is the resolvent of some maximally monotone operator such $A_i\colon \mathcal{H}_i \to 2^{\mathcal{H}_i}$ (i.e., $R_i = (\operatorname{Id} + A_i)^{-1}$) such that $0 \in A_i 0$. In this context, the variational inequality (4.5) assumes the more general form of a system of monotone inclusions, namely,

$$\text{find } \overline{x}_1 \in \mathcal{H}_1, \ldots, \overline{x}_m \in \mathcal{H}_m \text{ such that } \begin{cases} b_1 \in \overline{x}_1 - W_1\overline{x}_m + A_1\overline{x}_1 \\ b_2 \in \overline{x}_2 - W_2\overline{x}_1 + A_2\overline{x}_2 \\ \quad\vdots \\ b_m \in \overline{x}_m - W_m\overline{x}_{m-1} + A_m\overline{x}_m. \end{cases} \tag{4.16}$$

5 Analysis of nonperiodic networks

We analyze the deep neural network described in Model 1.1 in the following scenario.

Assumption 5.1 In the setting of Model 1.1, there exist sequences $(\omega_n)_{n\in\mathbb{N}} \in \ell_+^1$, $(\rho_n)_{n\in\mathbb{N}} \in \ell_+^1$, $(\eta_n)_{n\in\mathbb{N}} \in \ell_+^1$, and $(\nu_n)_{n\in\mathbb{N}} \in \ell_+^1$ for which the following hold for every $i \in \{1, \ldots, m\}$:

(i) There exists $W_i \in \mathcal{B}(\mathcal{H}_{i-1}, \mathcal{H}_i)$ such that $(\forall n \in \mathbb{N})\ \|W_{i,n} - W_i\| \leqslant \omega_n$.

(ii) There exists $R_i \in \mathcal{A}(\mathcal{H}_i)$ such that $(\forall n \in \mathbb{N})(\forall x \in \mathcal{H}_i)\ \|R_{i,n}x - R_i x\| \leqslant \rho_n\|x\| + \eta_n$.

(iii) There exists $b_i \in \mathcal{H}_i$ such that $(\forall n \in \mathbb{N})\ \|b_{i,n} - b_i\| \leqslant \nu_n$.

In addition, we set

$$(\forall i \in \{1, \ldots, m\}) \quad T_i\colon \mathcal{H}_{i-1} \to \mathcal{H}_i\colon x \mapsto R_i(W_i x + b_i). \tag{5.1}$$

Proposition 5.2 *In the setting of Model 1.1, suppose that Assumption 5.1 is satisfied, let $i \in \{1, \ldots, m\}$, and set*

$$(\forall n \in \mathbb{N}) \quad \chi_{i,n} = \rho_n\|W_{i,n}\| + \omega_n \quad \text{and} \quad \zeta_{i,n} = \rho_n\|b_{i,n}\| + \eta_n + \nu_n. \tag{5.2}$$

Then $(\chi_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$, $(\zeta_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$, and $(\forall n \in \mathbb{N})(\forall x \in \mathcal{H}_{i-1})\ \|T_{i,n}x - T_i x\| \leqslant \chi_{i,n}\|x\| + \zeta_{i,n}$.

Proof. According to Assumptions 5.1(i) and 5.1(iii), $\sup_{n\in\mathbb{N}}\|W_{i,n}\| < +\infty$ and $\sup_{n\in\mathbb{N}}\|b_{i,n}\| < +\infty$. It then follows from (5.2) that $(\chi_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$ and $(\zeta_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$. Hence, we deduce from (1.1), (5.1), the nonexpansiveness of R_i, and Assumption 5.1 that

$$
\begin{aligned}
(\forall n \in \mathbb{N})(\forall x \in \mathcal{H}_{i-1}) \quad & \|T_{i,n}x - T_i x\| \\
& \leqslant \|R_{i,n}(W_{i,n}x + b_{i,n}) - R_i(W_{i,n}x + b_{i,n})\| + \|R_i(W_{i,n}x + b_{i,n}) - R_i(W_i x + b_i)\| \\
& \leqslant \rho_n \|W_{i,n}x + b_{i,n}\| + \eta_n + \|W_{i,n}x + b_{i,n} - W_i x - b_i\| \\
& \leqslant \rho_n (\|W_{i,n}\|\,\|x\| + \|b_{i,n}\|) + \eta_n + \|W_{i,n} - W_i\|\,\|x\| + \|b_{i,n} - b_i\| \\
& \leqslant \rho_n (\|W_{i,n}\|\,\|x\| + \|b_{i,n}\|) + \eta_n + \omega_n\,\|x\| + \nu_n \\
& = \chi_{i,n}\|x\| + \zeta_{i,n},
\end{aligned}
\tag{5.3}
$$

as claimed. ☐

Proposition 5.3 *In the setting of Model 1.1, suppose that Assumption 5.1 is satisfied. Then, for every $i \in \{1, \ldots, m\}$, there exist $(\tau_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$ and $(\theta_{i,n})_{n\in\mathbb{N}} \in \ell_+^1$ such that*

$$
(\forall n \in \mathbb{N})(\forall x \in \mathcal{H}) \quad \|(T_{i,n} \circ \cdots \circ T_{1,n})x - (T_i \circ \cdots \circ T_1)x\| \leqslant \tau_{i,n}\|x\| + \theta_{i,n}. \tag{5.4}
$$

Proof. For every $i \in \{1, \ldots, m\}$, define $(\chi_{i,n})_{n\in\mathbb{N}}$ and $(\zeta_{i,n})_{n\in\mathbb{N}}$ as in (5.2), According to Proposition 5.2, (5.4) is satisfied for $i = 1$ by setting $(\forall n \in \mathbb{N})$ $\tau_{1,n} = \chi_{1,n}$ and $\theta_{1,n} = \zeta_{1,n}$. Next, let us assume that (5.4) holds for $i \in \{1, \ldots, m-1\}$ and set

$$
(\forall n \in \mathbb{N}) \quad
\begin{cases}
\tau_{i+1,n} = (\|W_{i+1}\| + \chi_{i+1,n})\tau_{i,n} + \chi_{i+1,n} \displaystyle\prod_{k=1}^{i} \|W_k\| \\
\theta_{i+1,n} = (\|W_{i+1}\| + \chi_{i+1,n})\theta_{i,n} + \chi_{i+1,n} \displaystyle\sum_{j=1}^{i} \left(\|b_j\| \prod_{k=j+1}^{i} \|W_k\| \right) + \zeta_{i+1,n}.
\end{cases}
\tag{5.5}
$$

Then the sequences $(\tau_{i+1,n})_{n\in\mathbb{N}}$ and $(\theta_{i+1,n})_{n\in\mathbb{N}}$ belong to ℓ_+^1. Now let $n \in \mathbb{N}$ and $x \in \mathcal{H}$. Upon invoking Proposition 5.2, the nonexpansiveness of R_{i+1}, and Proposition 4.2, we obtain

$$
\begin{aligned}
& \|(T_{i+1,n} \circ \cdots \circ T_{1,n})x - (T_{i+1} \circ \cdots \circ T_1)x\| \\
& \quad \leqslant \|(T_{i+1,n} \circ T_{i,n} \circ \cdots \circ T_{1,n})x - (T_{i+1} \circ T_{i,n} \circ \cdots \circ T_{1,n})x\| \\
& \qquad + \|(T_{i+1} \circ T_{i,n} \circ \cdots \circ T_{1,n})x - (T_{i+1} \circ T_i \circ \cdots \circ T_1)x\| \\
& \quad \leqslant \chi_{i+1,n}\|(T_{i,n} \circ \cdots \circ T_{1,n})x\| + \zeta_{i+1,n} + \|(T_{i+1} \circ T_{i,n} \circ \cdots \circ T_{1,n})x - (T_{i+1} \circ T_i \circ \cdots \circ T_1)x\| \\
& \quad \leqslant \chi_{i+1,n}(\|(T_{i,n} \circ \cdots \circ T_{1,n})x - (T_i \circ \cdots \circ T_1)x\| + \|(T_i \circ \cdots \circ T_1)x\|) + \zeta_{i+1,n} \\
& \qquad + \big\|R_{i+1}\big((W_{i+1} \circ T_{i,n} \circ \cdots T_{1,n})x + b_{i+1}\big) - R_{i+1}\big((W_{i+1} \circ T_i \circ \cdots \circ T_1)x + b_{i+1}\big)\big\| \\
& \quad \leqslant (\|W_{i+1}\| + \chi_{i+1,n})\|(T_{i,n} \circ \cdots \circ T_{1,n})x - (T_i \circ \cdots \circ T_1)x\| + \chi_{i+1,n}\|(T_i \circ \cdots \circ T_1)x\| + \zeta_{i+1,n} \\
& \quad \leqslant (\|W_{i+1}\| + \chi_{i+1,n})(\tau_{i,n}\|x\| + \theta_{i,n}) + \chi_{i+1,n} \left(\|x\| \prod_{k=1}^{i} \|W_k\| + \sum_{j=1}^{i} \left(\|b_j\| \prod_{k=j+1}^{i} \|W_k\| \right) \right) + \zeta_{i+1,n} \\
& \quad = \tau_{i+1,n}\|x\| + \theta_{i+1,n},
\end{aligned}
\tag{5.6}
$$

which proves the result by induction. ☐

We can now present the main result of this section on the asymptotic behavior of Model 1.1. The proof of this result relies on Theorem 4.7, which it extends.

Theorem 5.4 *Consider the setting of Model 1.1 and let $\alpha \in [1/2, 1]$. Suppose that Assumption 5.1 is satisfied as well as the following:*

(a) *$F = \operatorname{Fix} T \neq \varnothing$, where $T = T_m \circ \cdots \circ T_1$.*

(b) *$(W_i)_{1\leqslant i\leqslant m}$ satisfies Condition 3.1 with parameter α.*

(c) *One of the following is satisfied:*

  (i) *$\lambda_n \equiv \alpha = 1$ and $Tx_n - x_n \to 0$.*

  (ii) *$(\lambda_n)_{n\in\mathbb{N}}$ lies in $]0, 1/\alpha[$ and $\sum_{n\in\mathbb{N}} \lambda_n(1-\alpha\lambda_n) = +\infty$.*

Then $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{x}_m \in F$ and $(T_1\overline{x}_m, (T_2\circ T_1)\overline{x}_m, \ldots, (T_{m-1}\circ\cdots\circ T_1)\overline{x}_m, \overline{x}_m)$ solves (4.5). Now suppose that, in addition, any of the following holds:

(iii) *For every $i \in \{1, \ldots, m-1\}$, R_i is weakly sequentially continuous.*

(iv) *For every $i \in \{1, \ldots, m-1\}$, R_i is a separable activation function in the sense of Proposition 2.24.*

(v) *For every $i \in \{1, \ldots, m-1\}$, $\mathcal{H}_i$ is finite-dimensional.*

(vi) *For some $\varepsilon \in]0, 1/2[$, $(\lambda_n)_{n\in\mathbb{N}}$ lies in $[\varepsilon, (1-\varepsilon)(\varepsilon + 1/\alpha)]$ and, for every $i \in \{1, \ldots, m\}$, $\mathcal{H}_i = \mathcal{H}$ and there exists $\beta_i \in]0, 1[$ such that $\|W_i - 2(1-\beta_i)\operatorname{Id}\| + \|W_i\| \leqslant 2\beta_i$.*

Then, for every $i \in \{1, \ldots, m-1\}$, $(x_{i,n})_{n\in\mathbb{N}}$ converges weakly to $\overline{x}_i = (T_i\circ\cdots\circ T_1)\overline{x}_m$ and $(\overline{x}_1, \ldots, \overline{x}_m)$ solves (4.5).

Proof. Let $(y_n)_{n\in\mathbb{N}}$ be the sequence defined by $y_0 = x_0$ and

$$
\begin{array}{l}
\text{for } n = 0, 1, \ldots \\
\left\lfloor
\begin{array}{ll}
y_{1,n} & = T_1 y_n \\
y_{2,n} & = T_2 y_{1,n} \\
& \vdots \\
y_{m,n} & = T_m y_{m-1,n} \\
y_{n+1} & = y_n + \lambda_n(y_{m,n} - y_n).
\end{array}
\right.
\end{array}
\tag{5.7}
$$

For every $n \in \mathbb{N}$, set $S_n = T_{m,n} \circ \cdots \circ T_{1,n}$. We derive from (1.2) and (5.7) that

$$
\begin{aligned}
(\forall n \in \mathbb{N}) \quad \|x_{n+1} - y_{n+1}\| &= \|x_n + \lambda_n(S_n x_n - x_n) - y_n - \lambda_n(Ty_n - y_n)\| \\
&\leqslant \lambda_n\|S_n x_n - Tx_n\| + \|x_n - y_n + \lambda_n(Tx_n - Ty_n - x_n + y_n)\|.
\end{aligned}
\tag{5.8}
$$

At the same time, by Proposition 5.3, there exist $(\tau_{m,n})_{n\in\mathbb{N}} \in \ell^1_+$ and $(\theta_{m,n})_{n\in\mathbb{N}} \in \ell^1_+$ such that

$$
(\forall n \in \mathbb{N}) \quad \|S_n x_n - Tx_n\| \leqslant \tau_{m,n}\|x_n\| + \theta_{m,n} \tag{5.9}
$$
$$
\leqslant \tau_{m,n}(\|x_n - y_n\| + \|y_n\|) + \theta_{m,n}. \tag{5.10}
$$

On the other hand, by Theorem 3.8, Assumption 5.1(ii), and (b), T is α-averaged. Hence, there exists a nonexpansive operator $Q\colon \mathcal{H} \to \mathcal{H}$ such that $T = (1-\alpha)\operatorname{Id} + \alpha Q$. Since (c) implies that $(\lambda_n)_{n\in\mathbb{N}}$ lies in $]0, 1/\alpha]$, we deduce that

$$
\begin{aligned}
(\forall n \in \mathbb{N}) \quad \|x_n - y_n + \lambda_n(Tx_n - Ty_n - x_n + y_n)\| &= \|(1-\alpha\lambda_n)(x_n - y_n) + \alpha\lambda_n(Qx_n - Qy_n)\| \\
&\leqslant (1-\alpha\lambda_n)\|x_n - y_n\| + \alpha\lambda_n\|Qx_n - Qy_n\| \\
&\leqslant \|x_n - y_n\|.
\end{aligned}
\tag{5.11}
$$

Altogether (5.8), (5.10), and (5.11) yield

$$
(\forall n \in \mathbb{N}) \quad \|x_{n+1} - y_{n+1}\| \leqslant \Big(1 + \frac{\tau_{m,n}}{\alpha}\Big)\|x_n - y_n\| + \frac{1}{\alpha}\big(\tau_{m,n}\|y_n\| + \theta_{m,n}\big). \tag{5.12}
$$

However, Theorem 4.7 guarantees that $\delta = \sup_{n\in\mathbb{N}} \|y_n\| < +\infty$ and therefore that

$$(\forall n \in \mathbb{N}) \quad \|x_{n+1} - y_{n+1}\| \leqslant \Big(1 + \frac{\tau_{m,n}}{\alpha}\Big)\|x_n - y_n\| + \frac{1}{\alpha}\big(\tau_{m,n}\delta + \theta_{m,n}\big). \tag{5.13}$$

Since $(\tau_{m,n})_{n\in\mathbb{N}}$ and $(\tau_{m,n}\delta + \theta_{m,n})_{n\in\mathbb{N}}$ are in ℓ^1_+, there exists $\nu \in [0, +\infty[$ such that $\|x_n - y_n\| \to \nu$ [8, Lemma 5.31]. Consequently, $\delta' = \sup_{n\in\mathbb{N}} \|x_n\| \leqslant \delta + \sup_{n\in\mathbb{N}} \|x_n - y_n\| < +\infty$. Now, set

$$(\forall n \in \mathbb{N}) \quad e_n = \frac{1}{\alpha}(S_n x_n - T x_n). \tag{5.14}$$

Then it follows from (5.9) that

$$\sum_{n\in\mathbb{N}} \|e_n\| \leqslant \frac{1}{\alpha} \sum_{n\in\mathbb{N}} \big(\tau_{m,n}\|x_n\| + \theta_{m,n}\big) \leqslant \frac{\delta'}{\alpha} \sum_{n\in\mathbb{N}} \tau_{m,n} + \frac{1}{\alpha} \sum_{n\in\mathbb{N}} \theta_{m,n} < +\infty. \tag{5.15}$$

In view of (1.2), we have

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \mu_n(Qx_n + e_n - x_n), \quad \text{where} \quad \mu_n = \alpha\lambda_n \in \left]0, 1\right[. \tag{5.16}$$

(i): The weak convergence of $(x_n)_{n\in\mathbb{N}}$ to a point $\overline{x}_m \in \operatorname{Fix} Q = F$ follows from (5.16) and [8, Theorem 5.33(iv)] by arguing as in the proof of [8, Theorem 5.14(i)].

(ii): It follows from (5.16) that $\sum_{n\in\mathbb{N}} \mu_n(1 - \mu_n) = +\infty$. Hence [8, Proposition 5.34(iii)] implies that $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{x}_m \in \operatorname{Fix} Q = F$.

In (i)–(ii) above, Proposition 4.3(iii) ensures that $(T_1\overline{x}_m, (T_2 \circ T_1)\overline{x}_m, \ldots, (T_{m-1} \circ \cdots \circ T_1)\overline{x}_m, \overline{x}_m)$ solves (4.5).

(iii)–(v): If one of these assumptions holds, by proceeding as in the proof of Theorem 4.7(iii)–(v), we obtain that, for every $i \in \{1, \ldots, m-1\}$, $(T_i \circ \cdots \circ T_1)x_n \rightharpoonup \overline{x}_i = (T_i \circ \cdots \circ T_1)\overline{x}_m$ and that, furthermore, $(\overline{x}_1, \ldots, \overline{x}_m)$ solves (4.5). However, Proposition 5.3 asserts that, for every $i \in \{1, \ldots, m-1\}$, there exist $(\tau_{i,n})_{n\in\mathbb{N}} \in \ell^1_+$ and $(\theta_{i,n})_{n\in\mathbb{N}} \in \ell^1_+$ such that, for every $n \in \mathbb{N}$,

$$\|x_{i,n} - (T_i \circ \cdots \circ T_1)x_n\| = \|(T_{i,n} \circ \cdots \circ T_{1,n})x_n - (T_i \circ \cdots \circ T_1)x_n\| \leqslant \tau_{i,n}\|x_n\| + \theta_{i,n}. \tag{5.17}$$

Since $(x_n)_{n\in\mathbb{N}}$ is bounded, $x_{i,n} - (T_i \circ \cdots \circ T_1)x_n \to 0$ and therefore $x_{i,n} \rightharpoonup \overline{x}_i$.

(vi): For every $i \in \{1, \ldots, m\}$, set

$$(\forall n \in \mathbb{N}) \quad e_{i,n} = (T_{i,n} \circ T_{i-1,n} \circ \cdots \circ T_{1,n})x_n - (T_i \circ T_{i-1,n} \circ \cdots \circ T_{1,n})x_n, \tag{5.18}$$

and let $(\chi_{i,n})_{n\in\mathbb{N}}$ and $(\zeta_{i,n})_{n\in\mathbb{N}}$ be defined as in (5.2). By Propositions 4.2, 5.2, and 5.3, we have

$$(\forall n \in \mathbb{N}) \quad \|e_{1,n}\| \leqslant \chi_{1,n}\|x_n\| + \zeta_{1,n} \tag{5.19}$$

and

$$\begin{aligned}
&(\forall i \in \{2, \ldots, m\})(\exists\, (\tau_{i-1,n})_{n\in\mathbb{N}} \in \ell^1_+)(\exists\, (\theta_{i-1,n})_{n\in\mathbb{N}} \in \ell^1_+)(\forall n \in \mathbb{N}) \\
&\quad \|e_{i,n}\| \leqslant \chi_{i,n}\|(T_{i-1,n} \circ \cdots \circ T_{1,n})x_n\| + \zeta_{i,n} \\
&\quad\quad \leqslant \chi_{i,n}\big(\|(T_{i-1,n} \circ \cdots \circ T_{1,n})x_n - (T_{i-1} \circ \cdots \circ T_1)x_n\| + \|(T_{i-1} \circ \cdots \circ T_1)x_n\|\big) + \zeta_{i,n} \\
&\quad\quad \leqslant \chi_{i,n}\left(\tau_{i-1,n}\|x_n\| + \theta_{i-1,n} + \|x_n\| \prod_{k=1}^{i-1} \|W_k\| + \sum_{j=1}^{i-1} \|b_j\| \left(\prod_{k=j+1}^{i-1} \|W_k\|\right)\right) + \zeta_{i,n}.
\end{aligned} \tag{5.20}$$

Thus, since $(x_n)_{n\in\mathbb{N}}$ is bounded,

$$(\forall i \in \{1,\ldots,m\}) \quad (\|e_{i,n}\|)_{n\in\mathbb{N}} \in \ell_+^1. \tag{5.21}$$

In addition, by (5.18) and (1.2),

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \lambda_n\big(T_m(T_{m-1}(\cdots T_2(T_1x_n + e_{1,n}) + e_{2,n}\cdots) + e_{m-1,n}) + e_{m,n} - x_n\big). \tag{5.22}$$

Thus, since Proposition 3.6(iii) and Theorem 3.8 imply that the operators $(T_i)_{1\leqslant i\leqslant m}$ are averaged, the proof can be completed as that of Theorem 4.7(vi) since [24, Theorem 3.5(ii)] asserts that (4.15) remains valid under (5.21). ☐

References

[1] F. J. Aragón Artacho and R. Campoy, A new projection method for finding the closest point in the intersection of convex sets, *Comput. Optim. Appl.,* vol. 69, pp. 99–132, 2018.

[2] H. Attouch, J. Peypouquet, and P. Redont, Backward-forward algorithms for structured monotone inclusions in Hilbert spaces, *J. Math. Anal. Appl.,* vol. 457, pp. 1095–1117, 2018.

[3] J.-B. Baillon, R. E. Bruck, and S. Reich, On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces, *Houston J. Math.,* vol. 4, pp. 1–9, 1978.

[4] J.-B. Baillon, P. L. Combettes, and R. Cominetti, There is no variational characterization of the cycles in the method of periodic projections, *J. Funct. Anal.,* vol. 262, pp. 400–408, 2012.

[5] C. Bargetz, S. Reich, and R. Zalas, Convergence properties of dynamic string-averaging projection methods in the presence of perturbations, *Numer. Algorithms,* vol. 77, pp. 185–209, 2018.

[6] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory,* vol. 39, pp. 930–941, 1993.

[7] H. H. Bauschke and J. M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Rev.,* vol. 38, pp. 367–426, 1996.

[8] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces,* 2nd ed. Springer, New York, 2017.

[9] H. H. Bauschke, D. Noll, and H. M. Phan, Linear and strong convergence of algorithms involving averaged nonexpansive operators, *J. Math. Anal. Appl.,* vol. 421, pp. 1–20, 2015.

[10] J. Bilski, The backpropagation learning with logarithmic transfer function, *Proc. 5th Conf. Neural Netw. Soft Comput.,* pp. 71–76, 2000.

[11] J. M. Borwein, G. Li, and M. K. Tam, Convergence rate analysis for averaged fixed point iterations in common fixed point problems, *SIAM J. Optim.,* vol. 27, pp. 1–33, 2017.

[12] J. Borwein, S. Reich, and I. Shafrir, Krasnoselski-Mann iterations in normed spaces, *Canad. Math. Bull.,* vol. 35, pp. 21–28, 1992.

[13] R. I. Boţ and E. R. Csetnek, A dynamical system associated with the fixed points set of a nonexpansive operator, *J. Dynam. Differential Equations,* vol. 29, pp. 155–168, 2017.

[14] M. Bravo and R. Cominetti, Sharp convergence rates for averaged nonexpansive maps, *Israel J. Math.,* vol. 227, pp. 163–188, 2018.

[15] J. S. Bridle, Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, In: *Neurocomputing, NATO ASI Series, Series F,* vol. 68, pp. 227–236. Springer, Berlin, 1990.

[16] B. Carlile, G. Delamarter, P. Kinney, A. Marti, and B. Whitney, Improving deep learning by inverse square root linear units (ISRLUs), 2017. https://arxiv.org/abs/1710.09967

[17] A. Cegielski, *Iterative Methods for Fixed Point Problems in Hilbert Spaces,* Lecture Notes in Mathematics, vol. 2057. Springer, Heidelberg, 2012.

[18] Y. Censor and R. Mansour, New Douglas–Rachford algorithmic structures and their convergence analyses, *SIAM J. Optim.,* vol. 26, pp. 474–487, 2016.

[19] P. L. Combettes, Construction d'un point fixe commun à une famille de contractions fermes, C. R. Acad. Sci. Paris Sér. I Math., vol. 320, pp. 1385–1390, 1995.

[20] P. L. Combettes, Solving monotone inclusions via compositions of nonexpansive averaged operators, *Optimization,* vol. 53, pp. 475–504, 2004.

[21] P. L. Combettes, Monotone operator theory in convex optimization, *Math. Programming,* vol. B170, pp. 177–206, 2018.

[22] P. L. Combettes and J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, *SIAM J. Optim.,* vol. 18, pp. 1351–1376, 2007.

[23] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.,* vol. 4, pp. 1168–1200, 2005.

[24] P. L. Combettes and I. Yamada, Compositions and convex combinations of averaged nonexpansive operators, *J. Math. Anal. Appl.,* vol. 425, pp. 55–70, 2015.

[25] L. Condat, A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms, *J. Optim. Theory Appl.,* vol. 158, pp. 460–479, 2013.

[26] G. Cybenko, Approximation by superposition of sigmoidal functions, *Math. Control Signals Systems,* vol. 2, pp. 303–314, 1989.

[27] J. Eckstein and D. P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Math. Program.,* vol. 55, pp. 293–318, 1992.

[28] D. L. Elliot, A better activation function for artificial neural networks, *Institute for Systems Research, University of Maryland,* Tech. Rep. 93-8, 1993.

[29] K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks, *Neural Netw.,* vol. 2, pp. 183–192, 1989.

[30] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, *Proc. 14th Int. Conf. Artificial Intell. Stat.,* pp. 315–323, 2011.

[31] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd ed. Pearson Education, Singapore, 1998.

[32] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proc. Int. Conf. Comput. Vision,* pp. 1026–1034, 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *Proc. IEEE Conf. Comput. Vision Pattern Recogn.,* pp. 770-778, 2016.

[34] Y. A. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature,* vol. 521, pp. 436–444, 2015.

[35] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, Efficient backprop, *Lect. Notes Comput. Sci.,* vol. 1524, pp. 9–50, 1998.

[36] B. Martinet, Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox, *C. R. Acad. Sci. Paris,* vol. A274, pp. 163–165, 1972.

[37] W. S. McCulloch and W. H. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.,* vol. 5, pp. 115–133, 1943.

[38] W. M. Moursi, The forward-backward algorithm and the normal problem, *J. Optim. Theory Appl.,* vol. 176, pp. 605–624, 2018.

[39] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, *Proc. 27st Int. Conf. Machine Learn.,* pp. 807–814, 2010.

[40] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[41] R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control Optim.,* vol. 14, pp. 877–898, 1976.

[42] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Rev.,* vol. 65, pp. 386–408, 1958.

[43] E. K. Ryu, R. Hannah, and W. Yin, Scaled relative graph: Nonexpansive operators via 2D Euclidean geometry. https://arxiv.org/abs/1902.09788

[44] R. K. Srivastava, K. Greff, and J. Schmidhuber, Training very deep networks, *Proc. Neural Inform. Process. Syst. Conf.,* vol. 28, pp. 2377–2385, 2015.

[45] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, Deep dictionary learning, *IEEE Access*, vol. 4, pp. 10096–10109, 2016.

[46] P. Tseng, On the convergence of products of firmly nonexpansive mappings, *SIAM J. Optim.,* vol. 2, pp. 425–434, 1992.

[47] M. Yamagishi and I. Yamada, Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization, *Inverse Problems,* vol. 33, art. 044003, 35 pp., 2017.

[48] X.-P. Zhang, Thresholding neural network for adaptive noise reduction, *IEEE Trans. Neural Netw.,* vol. 12, pp. 567–584, 2001.