Regime Change Hypothesis: Foundations for Decoupled Dynamics in Neural Network Training

Cristian Pérez-Corral*
Universitat Politècnica de València
Valencia, Spain
cpercor@upv.es

Alberto Fernández-Hernández
Universitat Politècnica de València
Valencia, Spain
a.fernandez@upv.es

Jose I. Mestre
Universitat Politècnica de València
Valencia, Spain
jmiravet@uji.es

Manuel F. Dolz
Universitat Jaume I
Castelló de la Plana, Spain
dolzm@uji.es

Jose Duato
Openchip & Software Technologies
Barcelona, Spain
jose.duato@openchip.com

Enrique S. Quintana-Ortí
Universitat Politècnica de València
Valencia, Spain
quintana@disca.upv.es

***Abstract*—Despite the empirical success of Deep Neural Networks (DNNs), their internal training dynamics remain difficult to characterize. In ReLU-based models, the activation pattern induced by a given input determines the piecewise-linear region in which the network behaves affinely. Motivated by this geometry, we investigate whether training exhibits a two-timescale behavior: an early stage with substantial changes in activation patterns and a later stage where parameter updates predominantly refine the model within largely stable activation regimes. We first prove a local stability property: outside measure-zero sets of parameters and inputs, sufficiently small parameter perturbations preserve the activation pattern of a fixed input, implying locally affine behavior within activation regions. We then empirically track per-iteration changes in parameters and activation patterns across fully-connected and convolutional architectures, as well as Transformer-based models, where activation patterns are recorded in the ReLU feed-forward (MLP/FFN) submodules, using fixed validation subsets. Across the evaluated settings, activation-pattern changes decay 3 times earlier than parameter-update magnitudes, showing that late-stage training often proceeds within relatively stable activation regimes. These findings provide a concrete, architecture-agnostic instrument for monitoring training dynamics and motivate further study of decoupled optimization strategies for piecewise-linear networks.**

***Index Terms*—Activation patterns, ReLU neural networks, Learning dynamics, Regime change hypothesis, Piecewise-linear geometry, Convergence analysis.**

I. Introduction

Deep Learning (DL) has achieved remarkable success over the past years, establishing itself as the state-of-the-art paradigm across a wide range of domains, including Computer Vision (CV) and Natural Language Processing (NLP), among many others [1]. This progress builds upon decades of theoretical and algorithmic development, from early connectionist models [2], through recurrent architectures [3], to convolutional networks for visual recognition [4] and to the recent dominance of Transformer-based architectures [5].

Multilayer Perceptrons (MLPs) have played a key role in this progress and constitute a fundamental building block of most deep learning architectures. They are known as universal approximators of continuous nonlinear functions [6], which is achieved through the composition of linear and nonlinear transformations. When activation functions are piecewise, the network's expressivity stems from the combinatorial arrangement of activations, which partitions the input space into distinct regions, each exhibiting different functional behavior [7].

Building on recent insights, we hypothesize that knowledge within Rectified Linear Unit (ReLU) MLPs does not evolve uniformly throughout training. Hartmann et al. [8] analyzed how activation-pattern statistics (e.g., similarity measures) evolve over optimization and how they relate to the network's training dynamics. Along related lines, Duato et al. [9] argued that it can be useful to distinguish between two complementary aspects of learning in ReLU MLP: *structural knowledge*, associated with the activation masks that determine the partition of the input space into piecewise-affine regions, and *quantitative knowledge*, associated with the parameter values that specify the affine mapping within each region. Taken together, these observations motivate the working hypothesis that learning dynamics may admit a decoupled, two-stage interpretation, where activation regimes stabilize earlier while parameters continue to refine within those regimes.

In this work, we provide a local stability result for ReLU activation patterns under small parameter perturbations and complement it with an empirical analysis across architectures and datasets. While local stability does not imply global convergence under Stochastic Gradient Descent (SGD) or Adam, our results motivate a two-stage interpretation of training dynamics in which activation regimes stabilize earlier than parameters. The contributions of this paper are threefold:

1) **Local stability of activation patterns in ReLU MLPs**. We formalize a measure-theoretic statement showing that, for almost all parameter configurations and inputs, sufficiently small parameter perturbations preserve the activation pattern of a fixed input. This can be extended straightforwardly to piecewise-linear operations on CNNs.

*Corresponding author

2) **A practical instrumentation of "activation regime" dynamics during training**. We introduce a simple per-iteration tracking protocol that compares parameter-update magnitudes against activation-pattern changes, enabling side-by-side convergence profiling across models.
3) **Empirical evidence across architectures under controlled training**. We report a multi-architecture study on MLPs, Convolutional Neural Networks (CNNs) and Transformer-based models with ReLU non-linearities showing that activation-pattern changes typically decay earlier than parameter-update magnitudes under the evaluated settings, thus motivating a two-stage regime-change viewpoint of training.

The remainder of the paper is organized as follows. Section II reviews related work on network structure, activation patterns, and interpretability. Section III introduces the definitions and theoretical results underlying our approach. Section IV presents the experimental evaluation across different configurations. Finally, Section V concludes the paper and discusses future research directions.

II. Related Work

The expressive power of ReLU-based DNNs has been a central topic of study over the past decade [7], [10]. One of the main approaches to understanding this connects the success of learning with how the piecewise-linear transformations induced by ReLU activations partition the input space into distinct linear regions [10]. The number and geometry of these regions are bounded by the network's architecture and provide insight into how data points are clustered and separated in the learned representation space. Although the partition is implicitly governed by activation sign patterns, comparatively fewer works study their dynamics directly. Xu et al. [11] introduced a novel idea of convergence in neural networks, implying that, once the activations are fixed, the neural networks converge as an infinite product of matrices. Hartmann et al. [8] measured activation statistics such as entropy and similarity in ReLU-based DNNs, offering one of the first systematic analyses of activation pattern convergence. Fernández-Hernández et al. [12] introduced a metric that quantifies how activation patterns across different samples diverge during training, whose behavior correlates with overfitting and underfitting, thus providing a principled way to determine the optimal weight decay value for a given network and dataset. Montavon et al. [13] proposed an interpretability framework based on studying which features maximize neuron activations, enabling a more fine-grained understanding of feature specialization. Activation patterns also play a role in interpretability, due to their relation to the input space partition. Chu et al. [14] proposed an interpretability method based on the tessellation of the input space into polyhedra, relying on the linearity on each of those for fully interpretable methods. Xu et al. [15] study how subtle changes in activation can change the region of interest, providing a sensitivity analysis valuable in explainability research.

These studies highlight that activation patterns encode essential information about how neural networks transform and structure their input space, yet the question of how and when these patterns stabilize during training remains open (a gap this work addresses). Duato et al. [9] were one of the earliest to propose that activations and parameters can be trained in separate regimes, motivating the viewpoint studied here.

III. Theoretical foundations

This section provides the theoretical background necessary to support the results presented in this paper. We briefly include the definition of MLP used in this work, to unify notation and establish a consistent foundation for the concepts introduced in this section.

Definition 1. A **Multilayer Perceptron (MLP)** with ReLU activations and L hidden layers is a mapping $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$ defined as follows. Let $h_0(x) = x$ and, for $l = 1, \ldots, L$, define the *pre-activations*

$$z_l(x) = W_{l-1}h_{l-1}(x) + b_{l-1} \in \mathbb{R}^{n_l}, \tag{1}$$

and the *activations* $h_l(x) = \sigma(z_l(x))$ with $\sigma(t) = \max(t, 0)$ applied elementwise. The output layer is affine:

$$f(x) = W_L h_L(x) + b_L. \tag{2}$$

To explicitly indicate the dependence of the mapping $f(x)$ on the network parameters $w = (W_0, \ldots, W_L, b_0, \ldots, b_L)$, we use the notation $f(x)$ or $f(x; w)$ interchangeably, making this dependence explicit whenever needed.

A pivotal definition in this mathematical framework is that of activation pattern, as it will form the basis for the subsequent construction.

Definition 2. Let $f(\,\cdot\,; w)$ be an MLP as in Definition 1. For a fixed input $x \in \mathbb{R}^{n_0}$, the neuron (l, i) is **active** if $z_l(x; w)_i > 0$ and **inactive** otherwise. The **activation pattern** is $\mathrm{act}(x; w) = (\mathrm{act}_1(x; w), \ldots, \mathrm{act}_L(x; w))$, where $\mathrm{act}_l(x; w) = \mathbf{1}\{z_l(x; w) > 0\} \in \{0, 1\}^{n_l}$.

The activation pattern $\mathrm{act}(x; w)$ determines the polyhedral region of the input space $\mathbb{R}^{n_0}$ that contains x. Moreover, $\mathrm{act}(x; w)$ is not unique to the single sample x: it is shared by all points x' lying in that same polyhedral region, so every point in the region exhibits the same activation pattern.

Each of these regions corresponds to a convex polyhedron defined by the intersection of half-spaces determined by the ReLU activation states [10]. This geometric structure underlies the expressive power of ReLU-based DNNs [7]. While the existence and characterization of these polyhedral regions are well established in the literature, comparatively less attention has been given to understanding the sensitivity and stability of their boundaries; specifically, how small perturbations in parameters or inputs can lead to transitions across adjacent activation regions.

Here, we formalize this perspective and develop it as an analytical framework to support our subsequent hypotheses on activation pattern dynamics and training stability. Thus, the following proposition establishes that activation patterns in

ReLU networks exhibit a form of local stability. Although it is well known that ReLU networks are continuous piecewise-affine and restricted to an affine map in each linear region, we also provide an affine-region corollary to fix notation and make explicit the measure-zero exclusions, which will be assumed in our experiments. To the best of our knowledge, this statement is not commonly spelled out in the measure-theoretic form we use here (with explicit null sets in both parameter and input space); we therefore provide proofs for completeness.

Proposition 1 (Local stability of activation patterns)**.** *Let $f(\cdot\,; w)$ be a ReLU-based MLP on $\mathbb{R}^{n_0}$ with parameters $w \in \mathbb{R}^p$. Then there exists a set $B \subset \mathbb{R}^p$ of measure zero with the property that every $w_0 \in \mathbb{R}^p \setminus B$ admits a measure-zero set $Z \subset \mathbb{R}^{n_0}$ for which the following holds: for all $x \in \mathbb{R}^{n_0} \setminus Z$ one can find $\varepsilon > 0$ such that whenever $w \in \mathbb{R}^p$ satisfies $\|w - w_0\| < \varepsilon$, the activation pattern of x is identical under w_0 and w.*

Proof. Let us assume that $w_0 \in \mathbb{R}^p$ satisfies the following condition: for every neuron (l, i), with $l \in \{1, \dots, L\}$ and $i \in \{1, \dots, n_l\}$, the corresponding pre-activation function is nonzero on every non-empty open subset of $\mathbb{R}^{n_0}$. In other words, if

$$B = \big\{ w_0 \in \mathbb{R}^p \,\big|\, \exists\,(l,i) \text{ and } \exists \mathcal{U} \subseteq \mathbb{R}^{n_0} \text{ open and non-empty s.t. } z_l(x; w_0)_i = 0,\ \forall x \in \mathcal{U} \big\} \tag{3}$$

then let $w_0 \in \mathbb{R}^p \setminus B$. It turns out that B is a null subset of $\mathbb{R}^p$, although we defer its proof until the end for the sake of clarity.

For a fixed w_0, the function $x \mapsto z_l(x; w_0)_i$ is continuous and piecewise linear in x. Hence, there exists a finite collection of closed polyhedra $C_1, \dots, C_M \subseteq \mathbb{R}^{n_0}$, each defined by affine inequalities ($\leq$ or $\geq$), with nonempty and pairwise disjoint interiors, such that $C_1 \cup \ldots \cup C_M = \mathbb{R}^{n_0}$, and for every $x \in C_m$ one has $z_l(x; w_0)_i = a_{m,l,i} \cdot x + b_{m,l,i}$, for some $a_{m,l,i} \in \mathbb{R}^{n_0}$ and $b_{m,l,i} \in \mathbb{R}$ depending on w_0. For each neuron (l, i), let the zero level set of its pre-activation be defined as $Z_{l,i} = \{x \in \mathbb{R}^{n_0} \mid z_l(x; w_0)_i = 0\}$. For every $m \in \{1, \dots, M\}$, it follows that

$$Z_{l,i} \cap C_m = \{x \in C_m \mid a_{m,l,i} \cdot x + b_{m,l,i} = 0\}. \tag{4}$$

Taking this into account, three cases can be distinguished according to the coefficients of the above equation:

1) If $a_{m,l,i} \neq 0$, the set $Z_{l,i} \cap C_m$ is contained in an $(n_0 - 1)$-dimensional affine subspace and therefore has measure zero.
2) If $a_{m,l,i} = 0$ and $b_{m,l,i} \neq 0$, the intersection $Z_{l,i} \cap C_m$ is empty.
3) If $a_{m,l,i} = 0$ and $b_{m,l,i} = 0$, the equality $Z_{l,i} \cap C_m = C_m$ holds, implying that $\forall x \in \mathrm{int}(C_m)$, $z_l(x; w_0)_i = 0$, with $\mathrm{int}(C_m)$ an open nonempty subset of $\mathbb{R}^{n_0}$, which contradicts the initial assumption.

Since $C_1 \cup \ldots \cup C_M = \mathbb{R}^{n_0}$, it follows that $Z_{l,i} = \bigcup_{m=1}^{M} (Z_{l,i} \cap C_m)$ is a finite union of $(n_0 - 1)$-dimensional polyhedra and is therefore null in $\mathbb{R}^{n_0}$. Consequently, the set $Z = \bigcup_{l=1}^{L} \bigcup_{i=1}^{n_l} Z_{l,i}$ is also null.

Let a sample $x \in \mathbb{R}^{n_0} \setminus Z$ be fixed. For each neuron (l, i), the value $z_l(x; w_0)_i$ is nonzero, and hence its sign is either positive or negative. As the mapping $\mathbb{R}^p \to \mathbb{R}$ given by $w \longmapsto z_l(x; w)_i$ is continuous at w_0, it follows that there exists $\varepsilon_{l,i} > 0$ such that, for all $w \in \mathbb{R}^p$ satisfying $\|w - w_0\| < \varepsilon_{l,i}$, the same inequality $z_l(x; w)_i > 0$ (or < 0) holds. As a result, the pre-activation of each neuron at input x preserves its sign in a neighborhood of w_0, and consequently, the corresponding activation state of the neuron (l, i) for the sample x remains unchanged. Let $\varepsilon := \min\{\varepsilon_{l,i} \mid (l, i)\} > 0$. Then, for every $w \in \mathbb{R}^p$ satisfying $\|w - w_0\| < \varepsilon$, the activation pattern of the sample x remains identical under parameters w and w_0. This completes the first part of the proof.

It remains to show that

$$B = \big\{ w_0 \in \mathbb{R}^p \,\big|\, \exists\,(l,i) \text{ and } \exists \mathcal{U} \subseteq \mathbb{R}^{n_0} \text{ open and non-empty s.t. } z_l(x; w_0)_i = 0,\ \forall x \in \mathcal{U} \big\} \tag{5}$$

is a null subset of $\mathbb{R}^p$. Let us define

$$B_{l,i} = \big\{ w_0 \in \mathbb{R}^p \,\big|\, \exists \mathcal{U} \subseteq \mathbb{R}^{n_0} \text{ open and non-empty s.t. } z_l(x; w_0)_i = 0,\ \forall x \in \mathcal{U} \big\} \tag{6}$$

Clearly $B = \bigcup_l \bigcup_i B_{l,i}$, so it suffices to show that each $B_{l,i}$ is null in $\mathbb{R}^p$. Decompose the parameters as $w_0 = (u, \theta)$, where $\theta \in \mathbb{R}^d$ collects the weights and bias of neuron (l, i), and $u \in \mathbb{R}^{p-d}$ corresponds to the remaining parameters. Since the outputs of layer $l - 1$ depend only on u, they can be written as $\phi(x, u)$. Define $\hat{\phi}(x, u) = (\phi(x, u), 1)$, noting that the last component corresponds to the bias term. Then, $z_l(x; w_0)_i = \theta \cdot \hat{\phi}(x, u)$. Assume now that $w_0 \in B_{l,i}$. Then there exists an open nonempty set $\mathcal{U} \subseteq \mathbb{R}^{n_0}$ such that

$$0 = z_l(x; w_0)_i = \theta \cdot \hat{\phi}(x, u), \qquad \forall x \in \mathcal{U}. \tag{7}$$

Hence, θ is orthogonal to the vector space $V(u) = \mathrm{span}\{\hat{\phi}(x, u) \mid x \in \mathcal{U}\}$. Since $\hat{\phi}(x, u) \neq 0$ for all $x \in \mathcal{U}$, it follows that $\dim(V(u)) \geq 1$ in $\mathbb{R}^d$, and thus $\theta \in H(u) := V(u)^{\perp}$. In fact, $w_0 \in B_{l,i}$ if and only if $\theta \in H(u)$. As $\dim(H(u)) \leq d - 1$, the set $\{\theta \mid w_0 \in B_{l,i}\}$ has measure zero. By Fubini's theorem,

$$\mu(B_{l,i}) = \int_{\mathbb{R}^{p-d}} \mu(\{\theta \mid w_0 \in B_{l,i}\})\, du = 0, \tag{8}$$

and therefore $B_{l,i}$ is null in $\mathbb{R}^p$, as claimed. □

Remark 1. The measure-zero exclusions ($w \notin B$, $x \notin Z$) are essential: for fixed w, activation changes occur only on a finite union of codimension-one polyhedral facets in x (hence measure zero), so patterns are locally stable for almost every input. Degenerate parameters (e.g., $w = 0$ in $\mathrm{ReLU}(w \cdot x)$) make pre-activations vanish on regions with interior, so any perturbation flips the pattern; such singular w form a measure-zero set.

Corollary 1. Let $f(\cdot\,; w)$ be a ReLU-based MLP on $\mathbb{R}^{n_0}$ with parameters $w \in \mathbb{R}^p$. There exists a measure-zero set $F \subset \mathbb{R}^{n_0}$

such that, for every $x \in \mathbb{R}^{n_0} \setminus F$, one can find $\varepsilon > 0$, a matrix $K \in \mathbb{R}^{n_{L+1} \times n_0}$, and a vector $c \in \mathbb{R}^{n_{L+1}}$ satisfying

$$f(x'; w) = Kx' + c, \quad \forall x' \in \mathbb{R}^{n_0} \text{ s.t. } \|x' - x\| < \varepsilon. \quad (9)$$

Proof. Following the proof of Proposition 1, fix $w \in \mathbb{R}^p$. There exists a finite collection of closed polyhedra $C_1, \ldots, C_M \subseteq \mathbb{R}^{n_0}$ with nonempty and pairwise disjoint interiors such that $\bigcup_i C_i = \mathbb{R}^{n_0}$, where for every $x \in C_m$ one has $f(x; w)_j = a_{m,j} \cdot x + b_{m,j}$, with $a_{m,j}, b_{m,j}$ dependent on w. Let $F = \partial C_1 \cup \cdots \cup \partial C_M$, which clearly has measure zero in $\mathbb{R}^{n_0}$. Now, take $x \in \mathbb{R}^{n_0} \setminus F$ such that $x \in C_m$ for some m, and therefore $x \in \operatorname{int}(C_m)$. Therefore, there exists $\varepsilon > 0$ such that $\forall x'$ satisfying $\|x' - x\| < \varepsilon$ lies in $\operatorname{int}(C_m)$, thus each neuron is described in x' by the same affine expression $f(x'; w)_i = a_{m,i} \cdot x' + b_{m,i}$. Let $\operatorname{act}(x; w) = (\operatorname{act}_1(x; w), \ldots, \operatorname{act}_L(x; w))$ denote the activation pattern of x and define $D_l = \operatorname{diag}(\operatorname{act}_l(x; w)) \in \mathbb{R}^{n_l \times n_l}$ for $l = 1, \ldots, L$. Since $x \in \operatorname{int}(C_m)$, there exists $\varepsilon > 0$ such that every x' with $\|x' - x\| < \varepsilon$ lies in $\operatorname{int}(C_m)$, therefore $\operatorname{act}(x; w) = \operatorname{act}(x'; w)$ and thus induces the same diagonal masks $D_1, \ldots, D_L$. Therefore, on this neighborhood the network reduces to an affine map $f(x'; w) = Kx' + c$, where

$$K = W_L D_L W_{L-1} D_{L-1} \cdots W_1 D_1 W_0 \quad (10)$$

and

$$c = b_L + \sum_{k=0}^{L-1} \left(W_L D_L W_{L-1} D_{L-1} \cdots W_{k+1} D_{k+1} \right) b_k, \quad (11)$$

and the result follows. □

Remark 2. The above results extend to convolutional ReLU networks. After a standard linear unfolding of input patches (e.g., `im2col`), a convolution is represented by multiplication with a structured matrix (Toeplitz/block-Toeplitz), so the activation boundaries remain defined by linear equations and have measure zero. Thus Proposition 1 and Corollary 1 carry over to standard CNN layers.

These results support a local "freezing" intuition: for almost every input, activation patterns are stable under sufficiently small parameter perturbations. We refer to the empirical manifestation of this effect as the **regime change hypothesis**: activation regimes tend to stabilize earlier than parameters, even while parameters continue to adapt within the stabilized regimes (under the evaluated settings). From this point, a *purely ReLU-based* network (e.g., MLPs and CNNs composed of piecewise-linear operations) can be effectively approximated, in a neighborhood of the data, by a collection of affine operators corresponding to fixed activation regions, with the set of ReLU gates frozen. For Transformers, we track activation masks only in the ReLU FFN/MLP submodules; attention (softmax) lies outside this piecewise-linear analysis. We emphasize that our theoretical results establish a local stability property of activation patterns under sufficiently small parameter perturbations. While this does not by itself imply a global two-stage training law, it provides a rigorous mechanism that is consistent with, and partially explains, the empirical regime-change behavior observed across architectures in Section IV, where we quantify this separation by tracking activation-pattern changes and parameter-update magnitudes throughout training.

IV. Experiments

Once the main theoretical definitions and results have been introduced, this section aims to empirically evaluate the claim that, in general, the set of activation patterns stabilizes before the set of network parameters.

To this end, our goal is not to achieve state-of-the-art performance, but rather to assess the behavior of standard training processes. We therefore do not include certain common techniques, such as learning rate scheduling or label smoothing among others, that primarily enhance validation accuracy. Instead, we use simple, schedule-free baselines, but we sweep optimizers/hyperparameters to avoid drawing conclusions from clearly undertrained runs; we report the best-validation configuration per pair. Three categories of experiments are performed: *(i)* MLP experiments, evaluating the hypothesis on tabular, sequential, and image datasets; *(ii)* CNN experiments, analyzing similar behavior only on image datasets; and *(iii)* Transformer experiments, examining whether the hypothesis holds when MLPs are embedded within larger architectures. In the Transformer setting, activation patterns are recorded only for the ReLU feed-forward (MLP/FFN) blocks. Every dataset follows the standard train-val split.

TABLE I: MLPs used in the experiments with the configuration of each model specifying the input size (IS), number of classes (NC), hidden dimensions (HD), batch normalization (BN), and dropout (D).

	IS	NC	HD	BN	D
MLP - Adult	104	2	(256, 128, 64)	Yes	Yes
MLP - Breast Cancer	30	2	(64, 32)	Yes	Yes
MLP - HAR	561	6	(512, 256)	Yes	No
MLP - MNIST	784	10	(256, 128)	Yes	Yes

Within each category, multiple experiments were conducted with different configurations, as summarized below. Each network was trained from scratch, initialized with default settings. Early stopping was used in all experiments, with a patience of a third of the total number of epochs and a minimum improvement threshold of 1% for MLPs and CNNs, and 0% for Transformers. The training configurations performed are the following.

- **MLPs and LeNet5:** trained for up to 30 epochs using SGD and Adam optimizers, with learning rate $\in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay $\in \{10^{-3}, 10^{-4}\}$. The MLP configurations are provided in Table I.
- **AlexNet, ResNet34, and EfficientNetB0:** trained for up to 200 epochs using SGD (with and without Nesterov momentum), Adam, and AdamW, with learning rate $\in \{10^{-2}, 3{\cdot}10^{-2}, 6{\cdot}10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay $\in \{10^{-3}, 5 \cdot 10^{-3}, 10^{-4}\}$.

- **Transformers:** trained for up to 50 epochs using SGD, Adam, and AdamW. In most cases, AdamW achieved superior results, as is standard for Transformer architectures. The hyperparameters used were learning rate $\in \{5 \cdot 10^{-4}, 3 \cdot 10^{-4}, 10^{-4}, 2 \cdot 10^{-5}\}$ and weight decay $\in \{5 \cdot 10^{-2}, 10^{-2}\}$.

All experiments were conducted using Python 3.10 and PyTorch 2.6. For all networks, the ReLU activation function was employed, either as specified in the original architecture or by replacing the default activation with ReLU as in EfficientNetB0. For Transformer-based models, the replacement is applied to the feed-forward (MLP/FFN) non-linearities as specified.

Experiments were performed on a workstation equipped with an NVIDIA A100 GPU (40 GB VRAM), a 16-core AMD EPYC 7282 processor, and 128 GB of system memory. To ensure reproducibility, random seeds were fixed across all relevant libraries.

Table II summarizes the results of our experiments, reporting for each model–dataset pair the optimizer, weight decay (WD), and learning rate (LR) that achieved the best validation score (BVS), together with the epoch at which early stopping was triggered. Training and validation scores correspond to accuracy for all models except GPT-2, for which they reflect perplexity. To quantify *parameter convergence*, let $w_i \in \mathbb{R}^p$ denote the vector of all trainable parameters after processing the i-th training batch. We measure the (normalized) average absolute parameter update between consecutive iterations as

$$\Delta w_i \;=\; \frac{1}{p}\,\|w_i - w_{i-1}\|_1 \;=\; \frac{1}{p}\sum_{j=1}^{p}\big|(w_i)_j - (w_{i-1})_j\big|, \quad (12)$$

which in practice is computed layerwise by taking elementwise absolute differences of each parameter tensor and averaging them. To quantify *activation-pattern convergence*, for each layer $l \in \{1, \ldots, L\}$ we use the binary activation mask $\mathrm{act}_l(x; w) \in \{0,1\}^{n_l}$ from Definition 2 and evaluate it on a fixed subset S of validation samples of size $|S| = 0.2 \times$ (batch size), fixed through the whole training process. The per-batch Hamming-type change is computed as

$$\begin{aligned} \Delta a_i &= \frac{1}{L}\sum_{l=1}^{L} \Delta a_i^{(l)}, \\ \Delta a_i^{(l)} &= \frac{1}{|S|\,n_l}\sum_{x\in S}\big\|\mathrm{act}_l(x; w_i) - \mathrm{act}_l(x; w_{i-1})\big\|_1, \end{aligned} \quad (13)$$

i.e., the average fraction of bits that flip between consecutive batches (equivalently, the normalized Hamming distance since the masks are binary). These activation masks are computed in eval mode with dropout disabled and BN using running stats. This yields two trajectories, $\{\Delta w_i\}_{i=1}^{T}$ and $\{\Delta a_i\}_{i=1}^{T}$, with T denoting the number of batches. To reduce noise and highlight long-term trends, we smooth each per-batch trajectory using a simple moving average (SMA) implemented as a 1D convolution with a uniform kernel. The smoothing window is set to 30% of the number of batches per epoch. Since parameter- and activation-change curves live on different numerical scales, we apply a robust min-max normalization based on percentiles 0.5^{th} and 99.5^{th}. This normalization enables within-run trend comparison across scales, so AUC and Δ are interpreted as relative cumulative-change proxies rather than direct stabilization times. Finally, because the batch index is uniformly spaced, we summarize the overall amount of change by the mean value of the normalized curve, $\mathrm{AUC}(x) = \frac{1}{T}\sum_{i=1}^{T}\hat{x}_i$ (equivalently, a discretized area under the curve up to a constant factor). We report $\mathrm{AUC}_W = \mathrm{AUC}(\Delta w)$ for parameters and $\mathrm{AUC}_P = \mathrm{AUC}(\Delta a)$ for activation patterns, and define the relative speedup as $\rho = \mathrm{AUC}_W/\mathrm{AUC}_P$; in this setting, smaller AUC indicates faster stabilization, as it reflects lower cumulative change across training. To support reproducibility, all the implementations can be found in our repository[1].

As shown in Table II, a consistent trend emerges: for most models, activations exhibit lower cumulative change and typically reach a low-change regime earlier (see threshold analysis/curves) compared with parameters, with an average acceleration of $3.85\times$ and a positive and negative standard deviation of 6.43 and 2.03, respectively. The only exception is the MLP trained on MNIST, for which activations exhibit slightly slower convergence.

This behavior can be interpreted by examining the curves shown in Figure 1, which present the convergence trajectories for MLP, CNNs, and Transformers experiments across the different combinations of model and dataset. For the MLP trained on MNIST, the delayed activation convergence is attributable to substantial fluctuations in activation differences during the initial training batches. In all other experiments, the acceleration is pronounced, typically representing more than a threefold improvement, which highlights the earlier stabilization of activation patterns relative to parameters. Overall, the results across all networks are positive. With the exception of MNIST, the MLPs exhibit a clear earlier convergence of activation patterns relative to the parameters. This is also observed in the CNNs, where this behavior is most evident for EfficientNet-B0. The Transformers follow the same trend, implying that the same behavior trend appears even if the monitored network is embedded in a larger system.

V. Conclusion and Future Work

A. Conclusions

In this work, we have presented both theoretical and empirical results that advance the understanding of convergence dynamics in neural networks, with a particular focus on the relationship between parameter and activation convergence. Concretely, we established that activation patterns remain stable under sufficiently small perturbations of the network parameters. Building on this foundation, we proposed the *regime change hypothesis*, which claims that, during training, activation-pattern changes typically stabilize earlier than parameter-update magnitudes (under the evaluated settings). Also, we have

[1] https://github.com/cperezln/IJCNN_2026-RegimeChange

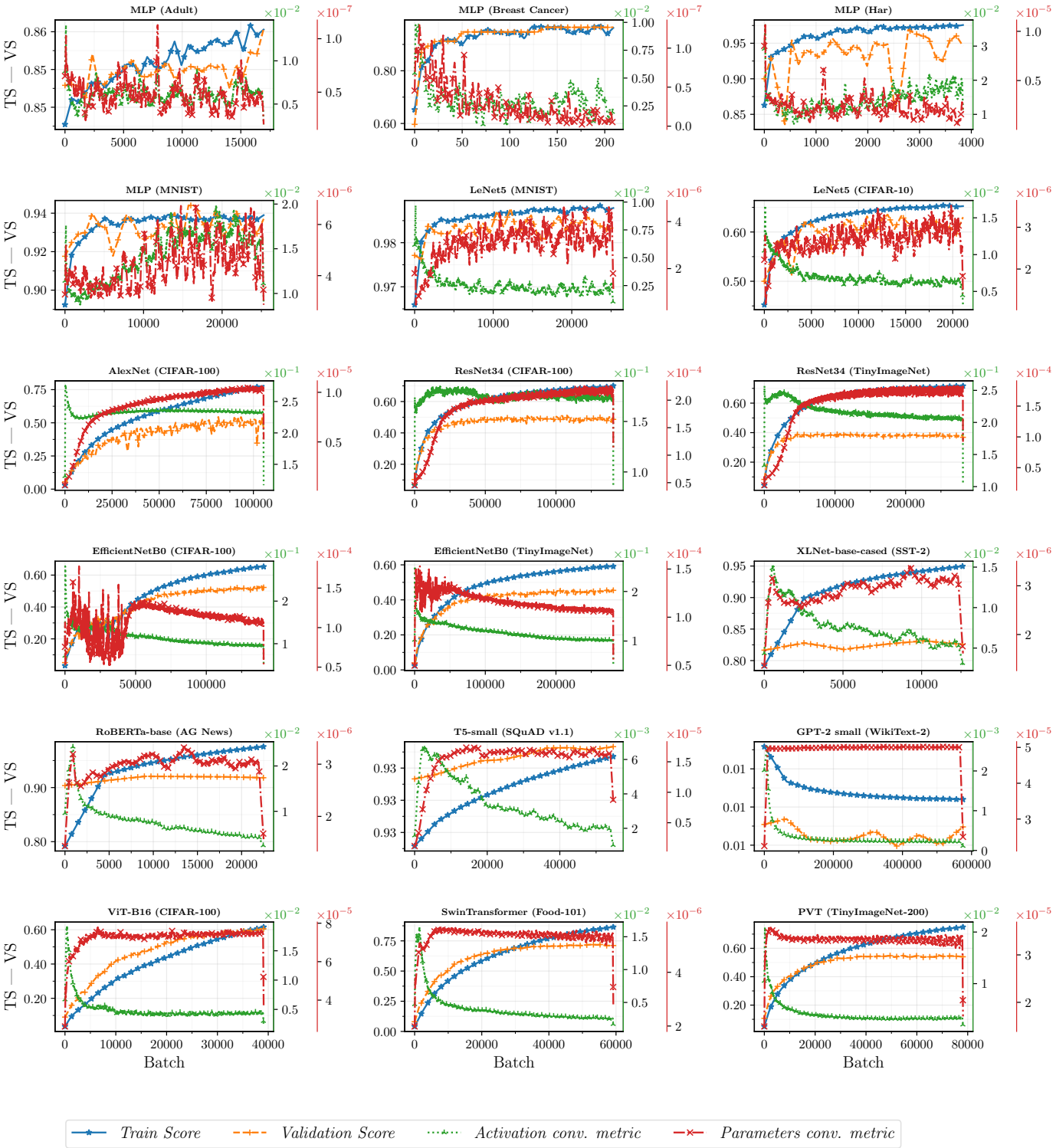

Fig. 1: Model curves showing the training and validation scores (scale at y-left axis), as well as activation and parameter convergences (scale at y-right axis), for all model–dataset pairs. The train/validation score (TS/VS) corresponds to train/validation accuracy for all models except GPT-2, for which it reflects perplexity. Activation and parameters convergence are measured as described before.

TABLE II: Comparisons between activation patterns convergence and parameter convergence for different models and datasets. Ep. refers to the early stopping epoch. BVS refers to the best validation score of each model – dataset. ρ refers to the convergence speedup factor between AUC_W and AUC_P.

Model	Dataset	Optim.	WD	LR	BVS	Ep.	AUC_W	AUC_P	ρ
MLP	Adult	Adam	10^{-4}	10^{-2}	0.86	15	0.32	0.30	**1.07**×
MLP	Breast Cancer	SGD	10^{-4}	10^{-2}	0.97	18	0.32	0.26	**1.23**×
MLP	HAR	SGD	10^{-4}	10^{-2}	0.97	15	0.20	0.19	**1.05** ×
MLP	MNIST	SGD	10^{-4}	10^{-2}	0.94	15	0.29	0.41	0.71 ×
LeNet5	MNIST	SGD	10^{-4}	10^{-2}	0.99	25	0.63	0.20	**3.15**×
LeNet5	CIFAR-10	Adam	10^{-4}	10^{-3}	0.64	29	0.67	0.20	**3.35**×
AlexNet	CIFAR-100	SGD	10^{-3}	10^{-3}	0.50	146	0.80	0.22	**3.6**×
ResNet34	CIFAR-100	Adam	10^{-4}	10^{-2}	0.51	101	0.74	0.66	**1.12**×
ResNet34	TinyImageNet-200	SGD[a]	$5 \cdot 10^{-3}$	10^{-3}	0.39	64	0.63	0.49	**1.3**×
EfficientNetB0	CIFAR-100	AdamW	10^{-3}	$2 \cdot 10^{-2}$	0.51	159	0.58	0.18	**3.2**×
EfficientNetB0	TinyImageNet-200	AdamW	10^{-3}	$2 \cdot 10^{-2}$	0.44	113	0.42	0.21	**2**×
XLNet	SST-2	AdamW	10^{-2}	$2 \cdot 10^{-5}$	0.83	4	0.75	0.44	**1.7**×
RoBERTa	AG News	AdamW	10^{-2}	$2 \cdot 10^{-5}$	0.92	5	0.77	0.22	**3.5**×
T5	SQuAD	AdamW	10^{-2}	10^{-4}	0.93	6	0.87	0.38	**2.3**×
GPT-2	WikiText-2	AdamW	10^{-2}	$3 \cdot 10^{-4}$	24[b]	6	0.97	0.08	**12**×
ViT-B16	CIFAR-100	AdamW	$5 \cdot 10^{-2}$	10^{-4}	0.61	50	0.87	0.08	**11**×
Swin	Food-101	AdamW	$5 \cdot 10^{-2}$	10^{-4}	0.72	50	0.86	0.12	**7.2** ×
PyramidViT (PVT)	TinyImageNet-200	AdamW	$5 \cdot 10^{-2}$	$5 \cdot 10^{-4}$	0.55	50	0.69	0.07	**9.9**×

[a] Stochastic gradient descent (SGD) with Nesterov momentum set to 0.9.
[b] GPT-2 reports perplexity as V.S., not accuracy.

provided extensive experimental validation of this hypothesis across a diverse set of architectures, including MLPs, CNNs, and Transformers, consistently supporting our claims. Our theoretical results formalize a local stability property: under well-defined conditions, network parameters may continue to evolve while activation patterns only change residually. This observation naturally motivates the regime change hypothesis, which is not formally proven yet the experimental evidence demonstrates that, in the vast majority of training scenarios, activation patterns stabilize significantly earlier than the parameters.

B. Future Work

The identification of two distinct phases of training (an initial phase in which activation patterns stabilize, followed by a phase of parameter refinement with residual changes in the activations) opens several promising directions for future research. This conceptual separation suggests that training could be decomposed into two simpler sub-processes, rather than treated as a single, monolithic optimization task, potentially accelerating convergence and reducing computational costs. In this regard, the regime change hypothesis suggests concrete training strategies, including two-phase optimization, updates constrained to the local affine regime once activations stabilize, and sensitivity-aware rules guided by the observed gap between activation-pattern and parameter convergence. This is a promising line of research that remains largely underexplored, with few approximations in the state-of-the-art. It will be interesting to extend the proposed analysis beyond ReLU-based networks. In particular, it remains open how analogous "regime-change" behavior manifests in architectures with smooth nonlinearities (e.g., GELU or SiLU), and in standard Transformer models where such activations are the default choice. As ongoing work, we are working on defining comparable activation-regime metrics in these settings and whether similar two-timescale dynamics can be observed; a comprehensive study of non-ReLU networks is therefore left as ongoing work and future research. Moreover, these results carry important implications for interpretability. Once activation patterns are fixed, a purely ReLU-based network can be expressed as an explicit collection of affine transformations over activation regions, each of which is potentially more interpretable and analyzable. In summary, the results presented here establish a theoretical and empirical foundation for viewing neural network training as a two-phase process. By explicitly treating training in this manner, future work may develop more efficient, interpretable, and distributed training strategies, further advancing our understanding of how deep networks learn and generalize.

ACKNOWLEDGMENTS

This research was funded by the projects PID2023-146569NB-C21 and PID2023-146569NB-C22 supported by MICIU/AEI/10.13039/501100011033 and ERDF/UE. Cristian Pérez-Corral received support from the *Conselleria de Educación, Cultura, Universidades y Empleo* (reference CIACIF/2024/412) through the European Social Fund Plus 2021–2027 (FSE+) program of the *Comunitat Valenciana*. Alberto Fernández-Hernández was supported by the predoctoral grant PREP2023-001826 supported by MICIU/AEI/10.13039/501100011033 and ESF+. Jose I. Mestre was supported by the predoctoral grant ACIF/2021/281 of the *Generalitat Valenciana*. Manuel F. Dolz was supported by the Plan Gen–T grant CIDEXG/2022/13 of the *Generalitat Valenciana*

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[2] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, 1958.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, 1989.

[7] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *International Conference on Machine Learning (ICML)*, 2017.

[8] D. Hartmann, D. Franzen, and S. Brodehl, "Studying the evolution of neural activation patterns during training of feed-forward relu networks," *Frontiers in Artificial Intelligence*, 2021.

[9] J. Duato, J. I. Mestre, M. F. Dolz, E. S. Quintana-Orti, and J. Cano, "Decoupling structural and quantitative knowledge in relu-based deep neural networks," in *Proceedings of the 5th Workshop on Machine Learning and Systems*, ser. EuroMLSys '25, 2025.

[10] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[11] Y. Xu and H. Zhang, "Convergence of deep relu networks," *Neurocomputing*, 2024.

[12] A. Fernández-Hernández, J. Mestre, M. F. Dolz, J. Duato, and E. S. Quintana-Orti, "Oui need to talk about weight decay: A new perspective on overfitting detection," 04 2025.

[13] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, 2018.

[14] L. Chu, X. Hu, J. Hu, L. Wang, and J. Pei, "Exact and consistent interpretation for piecewise linear neural networks: A closed form solution," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[15] S. Xu, J. Vaughan, J. Chen, A. Zhang, and A. Sudjianto, "Traversing the local polytopes of relu neural networks: A unified approach for network verification," 2022.