Notes on Differential Geometry in Neural Networks

Manifolds, Activations, and Lost Geometry: How Pointwise Nonlinearities Break the Map

Taha Bouhsine

azetta.ai

20 February 2026

Abstract

A neural network layer defines a map $F : \mathbb{R}^n \to \mathbb{R}^m$, and its restriction to a data manifold M determines the geometry of the learned representation. The Jacobian $\mathbf{J}_F$ controls how local distances, angles, and volumes are transported from M to representation space, so the question of what each architectural choice does to geometry becomes a question about $\mathbf{J}_F$. We analyse pointwise activation functions in this framework. Every such activation factors the Jacobian as $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$, with $\mathbf{D}_\phi$ a diagonal modulation of the linear part. Under the activations in common use, $\mathbf{D}_\phi$ admits only a narrow set of structural options: zeros (ReLU), saturating tails (sigmoid, tanh), or smooth strictly-positive entries (GELU, softplus, leaky ReLU). Each option translates into a distinct geometric pathology—exact rank collapse, ill-conditioning, metric warping, non-injectivity—and we make the translation precise in each case. We then identify a topological sufficient condition under which the layer preserves the manifold structure (strict monotonicity of ϕ together with full column rank of $\mathbf{W}$) and contrast it with the activations that fail the condition. The analysis points to a recurring design tension we call the *expressivity–geometry tradeoff*: sharper selective nonlinearities separate classes better and degrade geometric fidelity in lockstep, and no purely pointwise nonlinearity escapes the tradeoff.

1 Introduction

The manifold hypothesis [1]—that high-dimensional data concentrates near a much lower-dimensional submanifold $M \subset \mathbb{R}^D$—has become a foundational working assumption in deep learning. Under the hypothesis, the network's task is not to fit a function on all of $\mathbb{R}^D$ but to learn a *map from the data manifold into a representation space*: $M \to N \subset \mathbb{R}^m$, where the downstream task becomes simple.

The quality of this map is governed by its Jacobian. A well-behaved Jacobian preserves local geometry: nearby points stay nearby, angles between tangent vectors are approximately maintained, and local volumes do not collapse. A degenerate Jacobian destroys local geometry, and the loss is irreversible—the next layer cannot recover what this one collapsed.

Not all geometric distortion is harmful. Collapsing nuisance directions—directions that carry no task-relevant information—is in fact useful: it reduces the effective dimensionality of the representation to what matters. The point of the analysis is therefore not that activations should preserve all geometry, but that one should know which geometric structure is being preserved and which is being thrown away. The analysis below characterises *what* activations do to geometry; whether a particular distortion helps or hurts is a task-level judgement.

We focus on the specific mechanism by which geometry is degraded or destroyed in practice: **pointwise activation functions**. Every standard activation—ReLU, sigmoid, tanh, GELU, softplus—acts coordinatewise on the pre-activation vector $\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}$, and this single fact forces the Jacobian to factor as $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$, with $\mathbf{D}_\phi$ a diagonal matrix of activation derivatives.

The diagonal factor is therefore the activation's entire geometric contribution. When entries of $\mathbf{D}_\phi$ vanish (as in ReLU), $\mathbf{J}_F$ can lose rank exactly and collapse directions of the data manifold onto a lower-dimensional subset of representation space. When entries of $\mathbf{D}_\phi$ approach zero without vanishing (as in sigmoid or tanh), the Jacobian remains full-rank in the strict sense but its condition number grows, local volumes shrink toward zero, and the effective dimension of the representation degrades.

The paper is organised as follows. Section 2 fixes notation and recalls the manifold hypothesis. Section 3 states the factorisation $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$ and records its consequences for rank, conditioning, and the pullback metric. Section 4 analyses each standard activation in this framework. Section 5 gives a sufficient condition—strict monotonicity of ϕ together with full column rank of $\mathbf{W}$—under which the layer restricted to a compact data manifold is a homeomorphism, and treats the pullback metric in the same setting. Section 7 records a high-dimensional scaling result for ReLU. Section 6 is the discussion section. We frame the *expressivity–geometry tradeoff* there as a conjectural design principle rather than as a theorem, together with the deeper Riemannian reading of the layer as a representation; that material is offered as motivation and consequence rather than as further mathematical content.

2 Preliminaries: Manifolds and Maps

2.1 The Manifold Hypothesis

Definition 1 (Data manifold)**.** *A* data manifold *is a compact, connected, smooth submanifold $M \hookrightarrow \mathbb{R}^D$ of intrinsic dimension $d \ll D$, such that the data distribution p has its support concentrated on or near M:*

$$\operatorname{supp}(p) \subseteq M_\varepsilon := \{\mathbf{x} \in \mathbb{R}^D : \operatorname{dist}(\mathbf{x}, M) < \varepsilon\}$$

for some small $\varepsilon > 0$.

The manifold hypothesis asserts that natural data (images, text, speech) typically lies on such an M, even though each datum is represented as a vector in $\mathbb{R}^D$ with D potentially in the millions. The intrinsic dimension d is the geometric content the network must preserve; the ambient dimension D is incidental.

2.2 Neural Network Layers as Smooth Maps

A single layer of a neural network, including its activation function, defines a map:

$$F : \mathbb{R}^n \to \mathbb{R}^m, \qquad F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $\phi : \mathbb{R} \to \mathbb{R}$ is applied coordinatewise: $\phi(\mathbf{z})_i = \phi(z_i)$.

When restricted to the data manifold M, F becomes a map $F|_M : M \to N$, where $N = F(M)$ is the *representation image*. When $F|_M$ is a smooth immersion (in particular, when $\mathbf{J}_F$ has full column rank everywhere on M), N inherits a smooth manifold structure; otherwise, N may be a lower-dimensional or singular subset of $\mathbb{R}^m$.

Definition 2 (Pullback metric)**.** *Let (N, g_N) be a Riemannian manifold and $F : M \to N$ a smooth map. The* pullback metric *on M is the tensor:*

$$(F^* g_N)(\mathbf{u}, \mathbf{v}) = g_N\big(\mathbf{J}_F \cdot \mathbf{u},\ \mathbf{J}_F \cdot \mathbf{v}\big) \qquad \text{for } \mathbf{u}, \mathbf{v} \in T_{\mathbf{x}} M, \tag{2}$$

or in matrix form: $g_M = \mathbf{J}_F^\top g_N \mathbf{J}_F$.

The pullback metric encodes how F stretches, compresses, and rotates tangent vectors. If g_N is the standard Euclidean metric ($g_N = \mathbf{I}$), then $g_M = \mathbf{J}_F^\top \mathbf{J}_F$, and local distances in data space are governed entirely by the singular values of $\mathbf{J}_F$.

3 The Jacobian of a Layer with Activations

3.1 Diagonal Structure

Remark 1 (Scope: ambient analysis). The data manifold M has intrinsic dimension d, and the geometrically correct object is the restricted differential $d(F|_M)_{\mathbf{x}} : T_{\mathbf{x}}M \to T_{F(\mathbf{x})}\mathbb{R}^m$. However, since $T_{\mathbf{x}}M \subset \mathbb{R}^n$, the restricted differential is the composition of the ambient Jacobian $\mathbf{J}_F(\mathbf{x})$ with the inclusion of $T_{\mathbf{x}}M$. For clarity, we analyze the ambient Jacobian $\mathbf{J}_F \in \mathbb{R}^{m\times n}$ throughout; the ambient Jacobian constrains what the restricted map $F|_M$ can preserve, but the actual behavior on M depends on how the tangent space $T_{\mathbf{x}}M$ sits relative to the singular directions and kernel of $\mathbf{J}_F(\mathbf{x})$. If $\mathbf{J}_F$ has full column rank, then $d(F|_M)_{\mathbf{x}}$ is injective on $T_{\mathbf{x}}M$ for any d-dimensional M. If $\mathbf{J}_F$ is rank-deficient, the restriction to $T_{\mathbf{x}}M$ may or may not lose rank depending on the alignment of $T_{\mathbf{x}}M$ with the kernel.

The structural fact that drives everything below is that pointwise activations contribute a *diagonal* factor to the Jacobian, and only a diagonal factor.

Proposition 3 (Layer Jacobian). *For the layer $F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x}+\mathbf{b})$ with ϕ applied coordinatewise, the Jacobian at $\mathbf{x}$ is:*

$$\mathbf{J}_F(\mathbf{x}) = \mathbf{D}_\phi(\mathbf{z}) \cdot \mathbf{W}, \qquad \mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}, \tag{3}$$

where $\mathbf{D}_\phi(\mathbf{z}) = \operatorname{diag}\big(\phi'(z_1), \phi'(z_2), \ldots, \phi'(z_m)\big)$ is the diagonal matrix of activation derivatives.

Proof. By the chain rule, $\frac{\partial F_i}{\partial x_j} = \phi'(z_i) \cdot \mathbf{W}_{ij}$, which in matrix form is $\mathbf{J}_F = \mathbf{D}_\phi \cdot \mathbf{W}$. □

Remark. The Jacobian $\mathbf{J}_F = \mathbf{D}_\phi \cdot \mathbf{W}$ factors the layer's geometric action into two parts:

(a) $\mathbf{W}$: a linear map that rotates and scales tangent vectors (the "learned geometry").

(b) $\mathbf{D}_\phi$: a diagonal rescaling that selectively amplifies or suppresses each coordinate (the "activation geometry").

The activation does not create new geometric structure—it can only *modulate* what the linear part provides.

3.2 Rank, Singular Values, and Volume

Definition 4 (Effective rank). *The* effective rank *of $\mathbf{J}_F$ at a point $\mathbf{x}$ is the number of singular values $\sigma_i(\mathbf{J}_F(\mathbf{x}))$ that exceed a threshold $\varepsilon > 0$. For pointwise activations:*

$$\operatorname{rank}_\varepsilon(\mathbf{J}_F(\mathbf{x})) = \operatorname{rank}_\varepsilon(\mathbf{D}_\phi(\mathbf{z}) \cdot \mathbf{W}) \le \min\big(\operatorname{rank}(\mathbf{W}),\ |\{i : |\phi'(z_i)| > \varepsilon\}|\big). \tag{4}$$

Proposition 5 (Volume element). *Let $F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$ with $\mathbf{W} \in \mathbb{R}^{m\times n}$. The local n-dimensional volume scaling factor of the ambient map (measuring how F scales infinitesimal n-dimensional parallelepipeds) is:*

$$\operatorname{vol}_n(\mathbf{J}_F) = \sqrt{\det(\mathbf{J}_F^\top \mathbf{J}_F)} = \sqrt{\det(\mathbf{W}^\top \mathbf{D}_\phi^2 \mathbf{W})}. \tag{5}$$

Here $\mathbf{J}_F^\top \mathbf{J}_F \in \mathbb{R}^{n\times n}$, so the Gram determinant is well-defined regardless of whether $m = n$ or $m > n$.

In the special case $m = n$, this simplifies to $|\det(\mathbf{D}_\phi)| \cdot |\det(\mathbf{W})| = \big(\prod_{i=1}^m |\phi'(z_i)|\big)|\det(\mathbf{W})|$. For general $m \ge n$: if any $\phi'(z_i) = 0$, the corresponding row of $\mathbf{J}_F = \mathbf{D}_\phi\mathbf{W}$ is zeroed, reducing the row space. When sufficiently many rows are zeroed that $\operatorname{rank}(\mathbf{J}_F) < n$, the Gram determinant vanishes: the map squashes a neighborhood of $\mathbf{x}$ to a set of dimension less than n. For saturating

activations where $\phi'(z_i) \to 0$ without reaching zero, the Gram determinant approaches (but does not reach) zero—the volume shrinks but does not exactly collapse.

When M has intrinsic dimension $d < n$, the relevant quantity is the d-dimensional volume element on $T_{\mathbf{x}}M$, which depends on the restriction of $\mathbf{J}_F$ to $T_{\mathbf{x}}M$ rather than on $\mathbf{J}_F$ alone (see Remark 1).

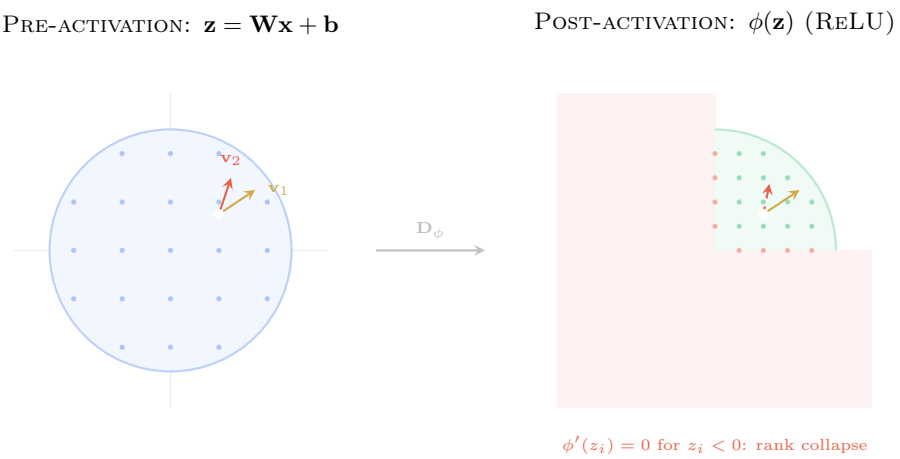

Figure 1: Schematic: a 2D disk before and after ReLU. **Left:** The pre-activation space with two tangent vectors $\mathbf{v}_1, \mathbf{v}_2$ at a point. **Right:** After ReLU, the negative quadrants collapse onto the axes. The red shaded region maps to zero; the green quarter-disk survives. Tangent vector directions into the collapsed region are lost.

4 Activation Functions as Geometric Operators

We now analyze each standard activation through the lens of the Jacobian framework. For each activation ϕ, we characterize: (1) the derivative ϕ' and its range, (2) whether exact rank collapse or ill-conditioning occurs (and where), and (3) the metric distortion.

4.1 Linear (Identity)

Definition 6. *$\phi(x) = x$, so $\phi'(x) = 1$ everywhere.*

Proposition 7. *With the identity activation, $\mathbf{D}_\phi = \mathbf{I}$ and $\mathbf{J}_F = \mathbf{W}$. The layer is a pure linear map: it preserves rank, invertibility is determined solely by $\mathbf{W}$, and the pullback metric is $g_M = \mathbf{W}^\top \mathbf{W}$. No saturation, no collapse.*

This is the baseline: no expressivity beyond linearity, but perfect geometric fidelity.

4.2 ReLU

Definition 8. *$\phi(x) = \max(0, x)$, so $\phi'(x) = \mathbf{1}_{x>0}$ (with $\phi'(0)$ undefined or set to 0).*

Theorem 9 (ReLU rank collapse)**.** *Let $F(\mathbf{x}) = \mathrm{ReLU}(\mathbf{W}\mathbf{x}+\mathbf{b})$ and let $S(\mathbf{x}) = \{i : (\mathbf{W}\mathbf{x}+\mathbf{b})_i > 0\}$ be the set of active neurons. Then:*

$$\mathrm{rank}(\mathbf{J}_F(\mathbf{x})) = \mathrm{rank}(\mathbf{W}_{S(\mathbf{x}),:}), \tag{6}$$

where $\mathbf{W}_{S(\mathbf{x}),:}$ is the submatrix of $\mathbf{W}$ consisting of rows indexed by $S(\mathbf{x})$.

Proof. $\mathbf{D}_\phi = \mathrm{diag}(\mathbf{1}_{z_i>0})$ zeros out entire rows of $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$ for each $i \notin S(\mathbf{x})$. The remaining rows are exactly $\mathbf{W}_{S(\mathbf{x}),:}$. □

Example. ReLU creates *dead zones*: for any $\mathbf{x}$ such that $z_i \leq 0$, the i-th output coordinate is identically zero, and the corresponding row of $\mathbf{J}_F$ vanishes. The map is piecewise linear with different ranks in different regions.

- **Non-invertibility:** ReLU is coordinatewise many-to-one: any two pre-activations that differ only in coordinates where both are negative produce the same output, since both are mapped to zero in those coordinates.
- **Volume collapse:** $\mathrm{vol}_n(\mathbf{J}_F) = 0$ whenever $\mathrm{rank}(\mathbf{W}_{S(\mathbf{x}),:}) < n$, i.e., when zeroing inactive rows drops $\mathbf{J}_F$ below full column rank.
- **Information loss:** The map loses sensitivity to input perturbations lying in directions annihilated by the surviving row submatrix $\mathbf{W}_{S(\mathbf{x}),:}$.

4.3 Leaky ReLU

Definition 10. *$\phi(x) = x$ for $x \geq 0$, $\phi(x) = \alpha x$ for $x < 0$, with $\alpha \in (0,1)$ (typically $\alpha = 0.01$). So $\phi'(x) = 1$ for $x > 0$ and $\phi'(x) = \alpha$ for $x < 0$.*

Proposition 11 (Leaky ReLU preserves rank)**.** *Since $\phi'(x) \in \{\alpha, 1\}$ with $\alpha > 0$, we have $\mathbf{D}_\phi = \mathrm{diag}(d_1, \ldots, d_m)$ with every $d_i \geq \alpha > 0$. Therefore:*

$$\mathrm{rank}(\mathbf{J}_F) = \mathrm{rank}(\mathbf{W}).$$

The Jacobian never loses rank due to the activation. However, the condition number is amplified:

$$\kappa(\mathbf{J}_F) \leq \frac{1}{\alpha} \cdot \kappa(\mathbf{W}), \tag{7}$$

so for $\alpha = 0.01$, the conditioning can worsen by a factor of 100. To see this (assuming $\mathbf{W}$ has full column rank), note that all singular values of $\mathbf{D}_\phi$ lie in $[\alpha, 1]$, so $\sigma_{\min}(\mathbf{D}_\phi) \geq \alpha$ and $\sigma_{\max}(\mathbf{D}_\phi) \leq 1$. For the product $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$:

$$\sigma_{\max}(\mathbf{J}_F) \leq \sigma_{\max}(\mathbf{D}_\phi) \cdot \sigma_{\max}(\mathbf{W}) \leq \sigma_{\max}(\mathbf{W}),$$

$$\sigma_{\min}(\mathbf{J}_F) \geq \sigma_{\min}(\mathbf{D}_\phi) \cdot \sigma_{\min}(\mathbf{W}) \geq \alpha \cdot \sigma_{\min}(\mathbf{W}),$$

where the second inequality uses the general fact that $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\,\sigma_{\min}(B)$ when A is square and nonsingular. Hence $\kappa(\mathbf{J}_F) = \sigma_{\max}(\mathbf{J}_F)/\sigma_{\min}(\mathbf{J}_F) \leq \kappa(\mathbf{W})/\alpha$.

4.4 Sigmoid

Definition 12. *$\sigma(x) = 1/(1+e^{-x})$, so $\sigma'(x) = \sigma(x)(1-\sigma(x))$.*

Proposition 13 (Sigmoid saturation)**.** *The derivative satisfies $\sigma'(x) \in (0, 1/4]$, with:*

- *$\sigma'(x) \to 0$ exponentially as $|x| \to \infty$ (saturation in both tails),*
- *$\sigma'(0) = 1/4$ (maximum sensitivity at the origin).*

For the layer Jacobian $\mathbf{J}_F = \mathbf{D}_\sigma \mathbf{W}$, the maximum singular value is bounded:

$$\sigma_{\max}(\mathbf{J}_F) \leq \frac{1}{4} \cdot \sigma_{\max}(\mathbf{W}). \tag{8}$$

As input magnitudes grow, all diagonal entries of $\mathbf{D}_\sigma$ approach zero and $\mathbf{J}_F \to 0$: the map "freezes" and the representation becomes increasingly ill-conditioned, concentrating near the boundary corners of $[0,1]^m$ determined by the sign pattern of the pre-activations $\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}$. Note that $\sigma'(x) > 0$ everywhere, so exact *rank collapse never occurs—the pathology is one of conditioning, not rank.*

4.5 Tanh

Definition 14. $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$, *so* $\tanh'(x) = 1 - \tanh^2(x)$.

Proposition 15 (Tanh saturation)**.** $\tanh'(x) \in (0, 1]$, *with* $\tanh'(0) = 1$ *(locally identity) and* $\tanh'(x) \to 0$ *as* $|x| \to \infty$. *The same saturation pathology as sigmoid, but centered on the origin: the output is bounded in* $(-1, 1)^m$. *As with sigmoid, exact rank is preserved (*$\tanh'(x) > 0$ *everywhere), but the Jacobian becomes arbitrarily ill-conditioned as inputs grow.*

4.6 Softplus

Definition 16. $\phi(x) = \log(1 + e^x)$, *so* $\phi'(x) = \sigma(x) = 1/(1 + e^{-x})$.

Proposition 17 (Softplus: smooth ReLU)**.** *Softplus is strictly monotone with* $\phi'(x) = \sigma(x) \in (0, 1)$. *The derivative is strictly positive everywhere, so:*

$$\operatorname{rank}(\mathbf{J}_F) = \operatorname{rank}(\mathbf{W}) \quad \textit{at every point.}$$

Moreover, $\phi'(x) \to 1$ *for* $x \to +\infty$ *(linear regime) and* $\phi'(x) \to 0^+$ *for* $x \to -\infty$ *(compression but not collapse).*

4.7 GELU

Definition 18. *The Gaussian error linear unit [3] is* $\mathrm{GELU}(x) = x \cdot \Phi(x)$, *where* Φ *is the CDF of* $\mathcal{N}(0, 1)$. *The derivative is* $\mathrm{GELU}'(x) = \Phi(x) + x \cdot \varphi(x)$, *where* φ *is the PDF of* $\mathcal{N}(0, 1)$.

Proposition 19 (GELU properties)**.** *GELU is smooth, and:*

- $\mathrm{GELU}'(0) = 0.5$ *(half-gate at origin),*
- $\mathrm{GELU}'(x) \to 1$ *as* $x \to +\infty$ *(approaches identity),*
- $\mathrm{GELU}'(x) \to 0$ *as* $x \to -\infty$ *(soft suppression),*
- $\mathrm{GELU}'(x) > 0$ *for all* x *above the unique minimum of* GELU'.

GELU is not globally monotone: GELU' *becomes slightly negative on a bounded interval of negative inputs, with a shallow minimum around* $x \approx -1.4$ *where* $\mathrm{GELU}' \approx -0.13$. *Empirically, trained networks tend not to have large concentrations of pre-activations in this narrow non-monotone region, though we are not aware of a formal result establishing this.*

4.8 Comparison

5 Topological Consequences

5.1 When Is the Map a Homeomorphism?

Theorem 20 (Sufficient condition for topological preservation)**.** *Let* $F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$ *with* $\mathbf{W} \in \mathbb{R}^{m \times n}$ *and* $\phi : \mathbb{R} \to \mathbb{R}$ *continuous and applied coordinatewise, and let* $M \subset \mathbb{R}^n$ *be a compact data manifold. If:*

(a) $\mathbf{W}$ *has full column rank (*$\operatorname{rank}(\mathbf{W}) = n$*, requiring* $m \geq n$*), and*

(b) ϕ *is* strictly monotone *(i.e.,* $\phi(a) < \phi(\mathbf{b})$ *whenever* $a < \mathbf{b}$*),*

then $F|_M : M \to F(M)$ *is a homeomorphism. Smoothness of* ϕ *is* not *required—continuity and strict monotonicity suffice.*

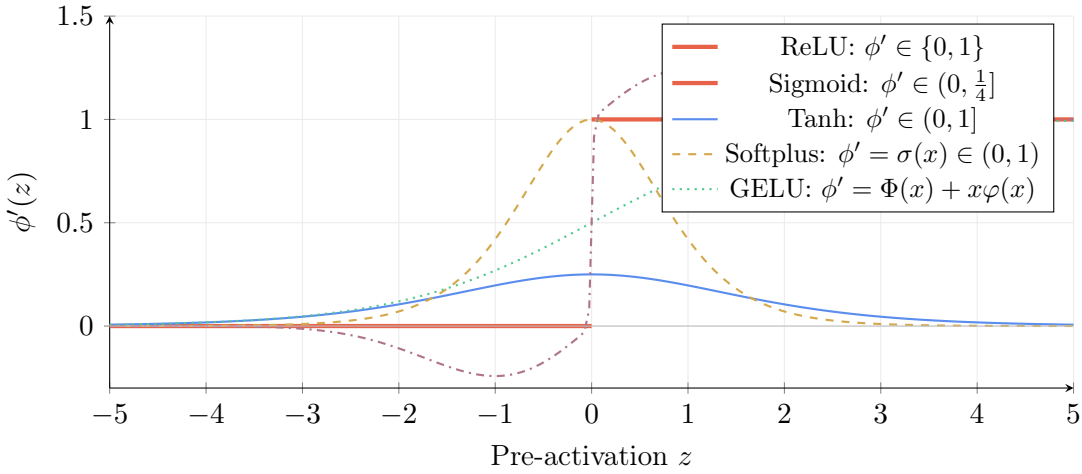

Figure 2: Derivatives $\phi'(z)$ of standard activations. Where $\phi' = 0$, the Jacobian $\mathbf{J}_F = \mathbf{D}_\phi \mathbf{W}$ loses exact rank; where $\phi' \approx 0$ but nonzero, the Jacobian becomes ill-conditioned (effective rank shrinks without exact rank loss). ReLU has exact zeros; sigmoid and tanh saturate smoothly toward zero; softplus remains strictly positive everywhere; GELU has a small negative dip near $z \approx -1.4$.

Proof. If $\mathbf{W}$ has full column rank, then $\mathbf{x} \mapsto \mathbf{W}\mathbf{x} + \mathbf{b}$ is injective. If ϕ is strictly monotone, then ϕ^{-1} exists and $\mathbf{z} \mapsto \phi(\mathbf{z})$ is injective coordinatewise. The composition of injectives is injective. Since ϕ is continuous and strictly monotone, F is a continuous injection. A continuous injection from a compact space into a Hausdorff space is a homeomorphism onto its image (see, e.g., Munkres [12], Theorem 26.6), so $F|_M$ is a homeomorphism onto $F(M)$. □

Remark 2 (Necessity). The converse—that conditions (a) and (b) are *necessary*—holds for the ambient map $F : \mathbb{R}^n \to \mathbb{R}^m$ but not in general for restrictions to a particular compact M. If $\mathbf{W}$ is rank-deficient, then $\ker(\mathbf{W}) \neq \{0\}$, and $F(\mathbf{x}) = F(\mathbf{x} + \mathbf{v})$ for $\mathbf{v} \in \ker(\mathbf{W})$; but $\mathbf{x}$ and $\mathbf{x} + \mathbf{v}$ may not both lie on a given M. Similarly, non-injectivity of ϕ creates collisions only at pairs of points where the pre-activations reach the colliding values, which need not occur on every M. Thus, strictly monotone ϕ plus full-rank $\mathbf{W}$ is sufficient for any compact M, and necessary for F to be globally injective on $\mathbb{R}^n$, but for a specific M the conditions could be relaxed.

Remark 3 (Smoothness and differential geometry). Theorem 20 guarantees topological preservation (no tearing, no folding) but says nothing about the quality of the pullback metric. For the differential-geometric analysis of Sections 3–6 to apply—Jacobians, singular values, pullback metrics—we additionally need ϕ to be differentiable (C^1 or smoother). Leaky ReLU, for example, is strictly monotone and continuous, so it preserves topology, but it is only piecewise smooth (C^0 at the origin), which means the Jacobian is undefined at the kink. Within the framework of this paper: *strict monotonicity governs topology; smoothness governs geometry.* (Other analytical frameworks—Lipschitz analysis, nonsmooth geometry—can handle some of these cases differently, but the slogan captures the dichotomy for the smooth differential-geometric setting we work in.)

Remark. Among the standard activations:

- **Topologically preserving:** Linear, Leaky ReLU, Sigmoid, Tanh, Softplus (all strictly monotone).
- **Topologically destroying:** ReLU (the scalar activation is not injective: all negative inputs map to 0), GELU (not globally monotone; the non-monotone region is narrow, but no formal guarantee prevents pre-activations from reaching it).

Activation	Range of ϕ'	Rank collapse?	ϕ injective?	Smooth?
Linear	$\{1\}$	No	Yes	C^∞
ReLU	$\{0, 1\}$	Yes (hard)	No	pw. C^∞, not C^1
Leaky ReLU	$\{\alpha, 1\}$	No	Yes	pw. C^∞, not C^1
Sigmoid	$(0, \frac{1}{4}]$	Asymptotic	Yes	C^∞
Tanh	$(0, 1]$	Asymptotic	Yes	C^∞
Softplus	$(0, 1)$	No	Yes	C^∞
GELU	$\approx [-0.13, 1.13]$	Near-zero	No (non-monotone)	C^∞

Table 1: Geometric properties of standard activations. "Rank collapse" distinguishes between *exact* collapse ($\phi'(z) = 0$ for some z, as in ReLU) and *asymptotic* ill-conditioning ($\phi'(z) \to 0$ but never reaches it, as in sigmoid/tanh). "ϕ injective?" refers to whether the scalar activation is globally injective (a necessary condition for the layer to be injective).

However, topological preservation is necessary but not sufficient for geometric preservation. Sigmoid and tanh preserve topology but can severely degrade metric structure through saturation.

5.2 The Pullback Metric Under Activation Warping

Even when ϕ is invertible, the pullback metric can be severely distorted.

Theorem 21 (Metric warping bound)**.** *Let $F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$ with $g_N = \mathbf{I}$ (Euclidean metric on N). In ambient coordinates, the pullback Gram matrix at $\mathbf{x}$ satisfies:*

$$g_M(\mathbf{x}) = \mathbf{W}^\top \mathbf{D}_\phi(\mathbf{z})^2 \mathbf{W}, \qquad \mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}. \tag{9}$$

The eigenvalues of g_M are bounded by:

$$\min_i |\phi'(z_i)|^2 \cdot \sigma_{\min}^2(\mathbf{W}) \le \lambda_k(g_M) \le \max_i |\phi'(z_i)|^2 \cdot \sigma_{\max}^2(\mathbf{W}). \tag{10}$$

Comparing the largest eigenvalue of $g_M(\mathbf{x})$ with the smallest eigenvalue of $g_M(\mathbf{x}')$ gives a worst-case bound on relative local stretching between the two points:

$$\frac{\lambda_{\max}(g_M(\mathbf{x}))}{\lambda_{\min}(g_M(\mathbf{x}'))} \le \frac{\max_i |\phi'(z_i)|^2}{\min_j |\phi'(\mathbf{z}'_j)|^2} \cdot \kappa^2(\mathbf{W}), \tag{11}$$

which can be arbitrarily large when ϕ saturates at one point but not the other. This is not a canonical global distortion invariant, but it bounds how differently the map stretches infinitesimal neighborhoods at $\mathbf{x}$ versus $\mathbf{x}'$.

Proof. $g_M = \mathbf{J}_F^\top \mathbf{J}_F = (\mathbf{D}_\phi \mathbf{W})^\top (\mathbf{D}_\phi \mathbf{W}) = \mathbf{W}^\top \mathbf{D}_\phi^2 \mathbf{W}$. Let $d_{\min}^2 = \min_i \phi'(z_i)^2$ and $d_{\max}^2 = \max_i \phi'(z_i)^2$. Since $\mathbf{D}_\phi^2$ is positive semidefinite diagonal, we have the matrix inequality $d_{\min}^2 \mathbf{I} \preceq \mathbf{D}_\phi^2 \preceq d_{\max}^2 \mathbf{I}$, which gives:

$$d_{\min}^2 \, \mathbf{W}^\top \mathbf{W} \preceq \mathbf{W}^\top \mathbf{D}_\phi^2 \mathbf{W} \preceq d_{\max}^2 \, \mathbf{W}^\top \mathbf{W}.$$

By the Courant–Fischer minimax characterization, $\lambda_k(g_M) = \min_{\dim(S)=k} \max_{\mathbf{v} \in S, \|\mathbf{v}\|=1} \mathbf{v}^\top g_M \mathbf{v}$. Applying this to both sides of the matrix inequality yields $d_{\min}^2 \, \lambda_k(\mathbf{W}^\top \mathbf{W}) \le \lambda_k(g_M) \le d_{\max}^2 \, \lambda_k(\mathbf{W}^\top \mathbf{W})$. Since $\lambda_k(\mathbf{W}^\top \mathbf{W}) = \sigma_k^2(\mathbf{W})$, the bounds follow. The distortion ratio bound is obtained by comparing the largest eigenvalue at $\mathbf{x}$ with the smallest at $\mathbf{x}'$. □

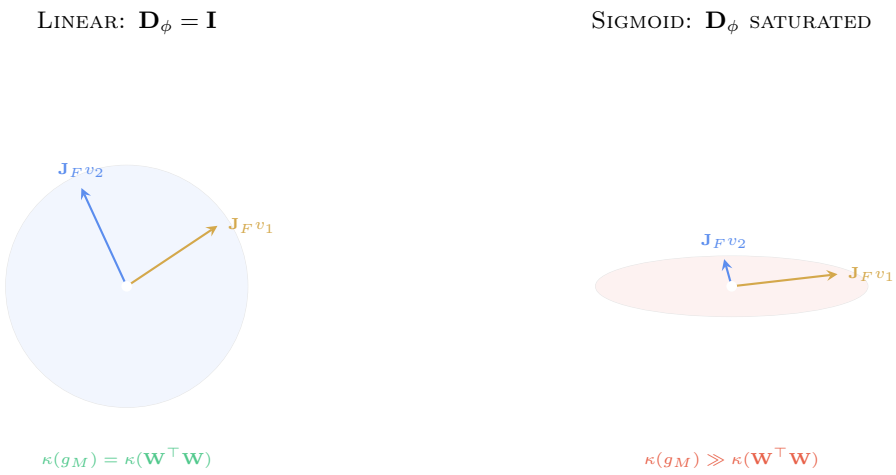

Figure 3: The pullback metric as an ellipsoid of tangent vector lengths. **Left:** Linear activation preserves the conditioning of $\mathbf{W}$. **Right:** Sigmoid saturation compresses one axis dramatically, worsening the condition number and distorting distances.

6 The Expressivity–Geometry Tradeoff

The analysis above lets us articulate a central design tension. What follows is not a theorem but a conjectural design principle—one that the formal results to this point make plausible, but that admits sharper formulation when "expressivity" is given a precise definition.

6.1 Why Activations Exist: Simulated Locality

Without activations, a stack of L layers produces a single linear map $\mathbf{W}_L \cdots \mathbf{W}_2 \mathbf{W}_1$—no increased expressivity, no ability to separate nonlinearly entangled classes. Activations provide *selectivity*: each neuron's pre-activation $z_i = \mathbf{w}_i^\top \mathbf{x} + b_i$ measures similarity to a learned prototype $\mathbf{w}_i$, and the activation ϕ decides which similarities to pass through.

Definition 22 (Prototype selectivity)**.** *The* selectivity *of neuron i at input $\mathbf{x}$ is $\phi'(z_i)$. When $\phi'(z_i) \approx 0$, the neuron is "blind" to its prototype—it contributes no information to the downstream representation. When $\phi'(z_i) \approx 1$, the neuron faithfully transmits its projection.*

This selectivity is the source of expressivity: it allows the network to implement nonlinear decision boundaries by activating different subsets of neurons in different input regions. But it comes at a cost.

6.2 The Tradeoff

Remark. **The expressivity–geometry tradeoff** (conjectural). For a layer $F(\mathbf{x}) = \phi(\mathbf{W}\mathbf{x} + \mathbf{b})$:

(i) **More selectivity** $\Rightarrow$ entries of $\mathbf{D}_\phi$ closer to $\{0, 1\}$ $\Rightarrow$ sharper nonlinearity $\Rightarrow$ better separation of classes $\Rightarrow$ worse geometry (rank collapse or ill-conditioning, non-invertibility).

(ii) **More geometry preservation** $\Rightarrow \mathbf{D}_\phi \approx \mathbf{I} \Rightarrow$ near-linear behavior $\Rightarrow$ less expressive $\Rightarrow$ better conditioning and invertibility.

We conjecture that the optimal activation navigates this tradeoff: near-linear for inputs close to the decision boundary (where geometry matters), selective elsewhere (where expressivity matters). This remains a design intuition rather than a formal result; making it precise would require a rigorous definition of "expressivity" that accounts for depth and width.

From the perspective developed here, GELU and softplus occupy an appealing middle ground: both are smooth, both avoid exact zeros in ϕ', and both transition gradually from suppression to transmission. ReLU is a corner solution that maximizes selectivity at the cost of geometry.

6.3 The Philosophical Point

The tradeoff reveals something deeper: *representation learning is a geometric problem.* The goal is not merely to find features that separate classes—it is to find a map $F|_M : M \to \mathbb{R}^m$ that preserves sufficient geometric structure for downstream tasks to work.

Cosine similarity, Euclidean k-NN, clustering—all of these downstream operations assume that the metric structure of N reflects meaningful relationships. If the map F has destroyed that structure through rank collapse or metric warping, no amount of downstream sophistication can recover it.

Remark. **Task-relevant geometry must be preserved end-to-end.** Real models often benefit from collapsing nuisance directions; the goal is not to preserve all geometry, but to preserve the structure that downstream tasks rely on.

(a) Choose activations that minimize Jacobian rank deficiency along task-relevant directions.

(b) Control input magnitudes (via normalization) to avoid saturation.

(c) Regularize the Jacobian's condition number when geometric fidelity matters (e.g., normalizing flows, metric learning).

(d) Match the evaluation metric to the geometry preserved by the network.

7 High-Dimensional Scaling

The geometric pathologies described above become dramatically worse in high dimensions.

Proposition 23 (ReLU collapse fraction—idealized)**.** *Let $\mathbf{z} \in \mathbb{R}^n$ with each z_i drawn independently from a symmetric distribution (mean zero). Then the probability that at least one coordinate is zeroed by ReLU is:*

$$P\big(\exists\, i : z_i \le 0\big) = 1 - 2^{-n}. \tag{12}$$

For $n = 768$ (a typical transformer hidden dimension), this is $1 - 2^{-768} \approx 1$.

Proof. Each coordinate is positive with probability $1/2$ by the symmetry assumption. The events are independent, so $P(\text{all positive}) = (1/2)^n = 2^{-n}$. □

Remark 4. The independence and symmetry assumptions are idealized: in practice, the pre-activations $z_i = \mathbf{w}_i^\top \mathbf{x} + b_i$ are correlated through the shared input $\mathbf{x}$, and biases shift the distribution away from zero. Nevertheless, the qualitative conclusion is robust: under symmetric pre-activation models the expected active fraction is $1/2$, and empirically, ReLU networks often exhibit substantial instantaneous sparsity [7], though the exact fraction depends on architecture, bias terms, normalization, and training. The idealized argument captures the qualitative scaling pressure toward sparsity.

Remark 5. This does not mean ReLU is useless in high dimensions—the network learns to route information through the active coordinates. But it does mean that at every point, a substantial fraction of output coordinates are zeroed (*sparsification*). Whether this sparsification causes actual rank collapse depends on whether the surviving row submatrix $\mathbf{W}_{S(\mathbf{x}),:}$ retains full column rank (Theorem 9). When $m \gg n$, many rows can be zeroed while $\mathbf{W}_{S(\mathbf{x}),:}$ still spans $\mathbb{R}^n$; when $m \approx n$, sparsification and rank collapse are closely linked.

Proposition 24 (Expected active fraction)**.** *For the setup of Proposition 23, the expected fraction of active coordinates is exactly* $1/2$*, and the standard deviation of the active fraction is* $O(n^{-1/2})$*. Thus for large* n*, almost exactly half the coordinates are zeroed at every point.*

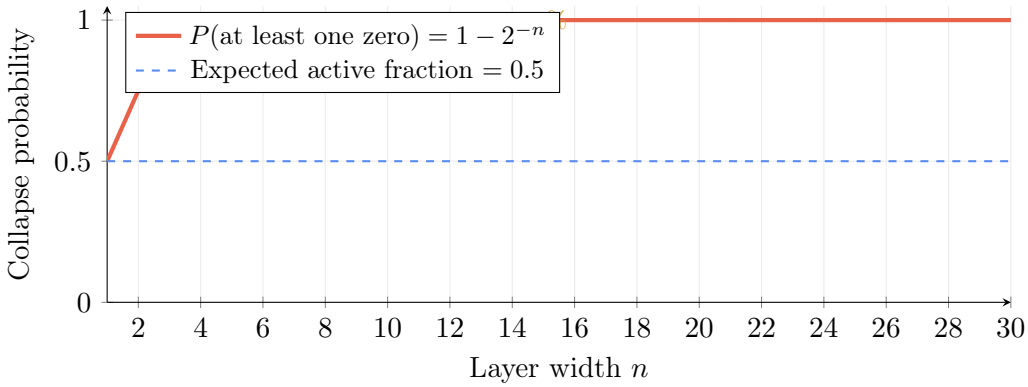

Figure 4: In n dimensions (under the idealized iid symmetric model), the probability that ReLU zeros at least one coordinate approaches 1 exponentially fast. By $n = 10$, it is already 99.9%. The expected active fraction is always 50%. Under these assumptions, approximately half the rows of $\mathbf{J}_F$ are zeroed at any point; whether this translates to rank loss depends on the structure of the surviving submatrix $\mathbf{W}_{S(\mathbf{x}),:}$.

7.1 Multi-Layer Composition

The single-layer analysis extends naturally to deep networks. For an L-layer network $F = F_L \circ \cdots \circ F_1$ with $F_\ell(\mathbf{x}) = \phi(\mathbf{W}_\ell \mathbf{x} + \mathbf{b}_\ell)$, the chain rule gives:

$$\mathbf{J}_F(\mathbf{x}) = \prod_{\ell=L}^{1} \mathbf{D}_\phi(\mathbf{z}^{(\ell)})\, \mathbf{W}_\ell, \tag{13}$$

where $\mathbf{z}^{(\ell)}$ is the pre-activation at layer ℓ. Rank defects *compound*: if $\mathbf{D}_\phi(\mathbf{z}^{(\ell)})$ drops k_ℓ dimensions at layer ℓ, the end-to-end rank satisfies $\operatorname{rank}(\mathbf{J}_F) \le \min_\ell \operatorname{rank}(\mathbf{D}_\phi(\mathbf{z}^{(\ell)})\mathbf{W}_\ell)$. For ReLU networks, this means the effective dimension of the representation can only decrease or stay the same through depth—geometry lost at any layer is lost permanently. This compounding effect is a key motivation for residual connections [4], which bias the Jacobian toward $\mathbf{I}$ + (perturbation) and can mitigate—though not guarantee prevention of—cumulative degeneracy in very deep networks. The complementary architectural strategy of *enforcing* invertibility, through unitary parameterizations [5] or invertible residual networks [6], removes the rank-loss failure mode at the cost of additional constraints on $\mathbf{W}$. A related strand of analysis studies the spectral and expressive geometry of deep networks under standard activations [8, 9, 10, 11].

8 Discussion: Activations as Geometry

8.1 The Map Is the Representation

A representation is not a set of features—it is a *map* $F|_M : M \to \mathbb{R}^m$. The quality of the representation is the quality of this map: does it preserve the geometric structure needed for downstream tasks?

Remark 6 (Standard techniques as Jacobian regularization). This perspective reframes several standard deep learning practices as forms of Jacobian regularization: **(i)** Residual connections

add $\mathbf{x}$ to $F(\mathbf{x})$, keeping the Jacobian near $\mathbf{I} + \mathbf{J}_{\text{branch}}$ and helping mitigate degeneracy. **(ii)** Batch normalization rescales activations to prevent saturation, keeping $\mathbf{D}_\phi$ away from zero. **(iii)** Layer normalization constrains representations to near a sphere, connecting to geometry-objective alignment. **(iv)** Skip and dense connections increase the effective rank of the end-to-end Jacobian. A full treatment of these connections is beyond the scope of this paper, but the pattern is consistent: techniques that stabilize training also stabilize the geometry of the learned map.

8.2 The Riemannian View

If we equip the representation space N with the standard Euclidean metric, the pullback induced by the restricted differential is a Riemannian metric on M wherever $d(F|_M)_{\mathbf{x}}$ is injective. This metric encodes the network's "belief" about which data points are close and which are far.

Remark. The optimal learned metric g_M^* should satisfy:

- Points from the same class should have small geodesic distance under g_M^*.
- Points from different classes should have large geodesic distance under g_M^*.
- The metric should be smooth and well-conditioned (no singularities or degeneracies).

Activations that cause rank collapse create degeneracies in the pullback metric, where the Riemannian description breaks down and distances along some directions collapse.

8.3 Implications for Evaluation Metrics

This geometric perspective connects directly to our earlier work on cosine similarity [2]. If the network's Jacobian is well-conditioned, then cosine similarity in representation space reflects genuine angular relationships on the data manifold. If the Jacobian has collapsed, cosine similarity in representation space is comparing artifacts, not features.

Remark. **The full pipeline.** For cosine similarity to be meaningful:

(i) Train with normalization (so the gauge freedom doesn't arise).

(ii) Use activations that preserve Jacobian rank (so the map doesn't destroy geometry).

(iii) Control input magnitudes (so activations don't saturate).

Condition (i) was the subject of our companion paper. Conditions (ii) and (iii) are the subject of this one.

Summary. The mathematical content of the paper is compact enough to fit in a single table.

Remark. Activations are not nonlinearities added for expressivity and then forgotten. They are **geometric operators** that control the Jacobian, the pullback metric, and the topology of the learned representation. When they are many-to-one and induce zero derivatives (ReLU), the Jacobian may lose exact rank and information along the killed directions is destroyed. When they saturate (sigmoid, tanh), exact rank is preserved but the Jacobian becomes ill-conditioned: effective dimension shrinks and metric structure is distorted. When they are non-smooth (C^0 only), the Jacobian is undefined at the non-differentiable points, and the differential-geometric framework simply does not apply there.

The decomposition to keep in mind is

$$\underbrace{F = \phi \circ (\mathbf{W} \cdot + \mathbf{b})}_{\text{layer}}, \qquad \mathbf{J}_F = \underbrace{\mathbf{D}_\phi}_{\text{geometric modulation}} \cdot \underbrace{\mathbf{W}}_{\text{learned geometry}}.$$

Choose the activation with the same care you choose the loss; both decide the geometry of the representation.

Concept	Formal statement
Layer Jacobian	$\mathbf{J}_F = \mathbf{D}_\phi(\mathbf{z}) \cdot \mathbf{W}$: diagonal modulation of the linear map
Exact rank collapse (ReLU)	$\text{rank}(\mathbf{J}_F) = \text{rank}(\mathbf{W}_{S(\mathbf{x}),:})$ where $S(\mathbf{x}) = \{i : z_i > 0\}$
Ill-conditioning (sigmoid, tanh)	$\phi'(z_i) \to 0$ degrades $\kappa(\mathbf{J}_F)$ without exact rank loss
Volume collapse	$\text{vol}_n(\mathbf{J}_F) = 0$ when $\text{rank}(\mathbf{J}_F) < n$ (exact, e.g. ReLU); $\text{vol}_n(\mathbf{J}_F) \to 0$ under saturation
Pullback metric	$g_M = \mathbf{W}^\top \mathbf{D}_\phi^2 \mathbf{W}$: activation warps distances
Topological preservation	Sufficient: ϕ strictly monotone + $\mathbf{W}$ full column rank
High-dim collapse (ReLU)	$P(\text{at least 1 dead neuron}) = 1 - 2^{-n}$

Table 2: Summary of the geometric analysis of activation functions.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[2] T. Bouhsine. In defense of cosine similarity: Normalization eliminates the gauge freedom. *arXiv preprint arXiv:2602.19393*, 2026. https://arxiv.org/abs/2602.19393

[3] D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[5] M. Arjovsky, A. Shah, and Y. Bengio. Unitary evolution recurrent neural networks. In *Proc. ICML*, 2016.

[6] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *Proc. ICML*, 2019.

[7] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. ICML*, 2010.

[8] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Proc. NeurIPS*, 2014.

[9] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Proc. NeurIPS*, 2016.

[10] J. Pennington, S. Schoenholz, and S. Ganguli. The emergence of spectral universality in deep networks. In *Proc. AISTATS*, 2018.

[11] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proc. NeurIPS*, 2017.

[12] J. R. Munkres. *Topology*. Prentice Hall, 2nd edition, 2000.