mathematics

MDPI

Article

Constraint Unlearnability in Overparameterized Neural Networks

Hongyu Qi and Zhen Tan *

National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, China; qihongyu@nudt.edu.cn
* Correspondence: tanzhen08a@nudt.edu.cn

Abstract

We study whether overparameterized neural networks can learn to satisfy pointwise constraints without inductive bias. One might expect that a sufficiently expressive network would learn to respect such constraints. We prove that this is not the case. For overparameterized feedforward neural networks with real-analytic activation functions, assuming the manifold of global minima is connected, the set of parameters for which the network satisfies a real-analytic pointwise constraint has measure zero on this manifold. We further prove that, under a standard convergence assumption on gradient flow, the attraction basin of these constraint-satisfying parameters has measure zero in the parameter space, so that gradient descent almost surely finds solutions that violate the constraint. The proof applies the identity theorem for real-analytic functions twice: first, to show that a network either satisfies the constraint everywhere or violates it almost everywhere (with respect to Lebesgue measure on the input domain); second, to show that overparameterization excludes the everywhere-satisfied case. For the dynamical part, the stable manifold theorem lifts this zero-measure property from the manifold to the parameter space. We call this phenomenon constraint unlearnability—despite interpolating the training data, the network cannot satisfy the constraint exactly on the entire input domain.

Keywords: neural networks; inductive bias; real-analytic functions; identity theorem; constraint satisfaction; overparameterization

MSC: 68T07; 57R35

1. Introduction

Neural networks have achieved tremendous success across many domains. This success has been accompanied by the widespread use of architectural inductive biases that enforce pointwise constraints: softmax, for instance, ensures that outputs sum to one at every input. These biases are ubiquitous in practice—but given the expressive power of overparameterized networks, are they mathematically necessary, or merely an engineering choice?

Consider probability normalization. A network that outputs a distribution must satisfy $\sum_i f_\theta(x)_i = 1$ for every input x. Softmax guarantees this by construction. Yet universal approximation tells us that a sufficiently wide network can approximate any continuous function, including those that satisfy this constraint. Whether the network could learn to satisfy it on its own, without the bias, is the question we address.

We prove that it cannot. For overparameterized feedforward neural networks with real-analytic activation functions (including GELU and Swish, which dominate modern architectures), in the absence of the corresponding inductive bias (i.e., no architectural guarantee that the constraint holds for all parameters), the set of parameters for which the

Check for updates

Academic Editor: Ioannis Tsoulos

Received: 17 July 2026
Revised: 9 August 2026
Accepted: 13 August 2026
Published: 24 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

network satisfies a real-analytic pointwise constraint has measure zero on the manifold of global minima (assuming this manifold is connected; see Section 3.2). We further prove that the attraction basin of these constraint-satisfying parameters has measure zero in the parameter space, so that gradient descent almost surely finds solutions that violate the constraint. We call this phenomenon *constraint unlearnability*: the network cannot learn to satisfy the constraint exactly on the entire input domain, even though it does so on the training set.

The proof applies the identity theorem for real-analytic functions twice: first, to show that a network either satisfies the constraint everywhere or violates it almost everywhere in Ω; second, to show that the everywhere-satisfied case is non-generic on the manifold of global minima. For the dynamical part, the stable manifold theorem lifts this zero-measure property from the manifold to the parameter space.

Cooper [1] characterized the geometry of the loss landscape for overparameterized networks; we use this geometry but ask a different question—not what the loss landscape looks like, but whether constraint satisfaction is possible within it. Methods that enforce constraints by construction [2] or by penalizing violations [3] implicitly acknowledge that unconstrained training fails; our result provides the mathematical reason.

Our contributions are: (1) a theorem establishing constraint unlearnability for overparameterized networks with real-analytic activations; (2) the implication that architectural inductive biases for pointwise constraints are mathematically necessary for exact constraint satisfaction on the entire input domain, not merely an engineering choice. Section 2 discusses related work. Section 3 introduces the necessary background. Section 4 states and proves the main result. Section 5 discusses implications and limitations.

2. Related Work

Cooper [1] proved that the set of global minima of an overparameterized neural network is a smooth submanifold of parameter space, whose dimension equals the number of excess parameters. This structural result is the geometric foundation of our work: it gives the manifold of global minima on which we perform measure-theoretic analysis. Nguyen & Hein [4] showed that for sufficiently wide networks with analytic activations, every local minimum is global; Simsek et al. [5] further characterized how permutation symmetries shape the loss landscape. These works establish that overparameterization produces a well-behaved manifold of interpolating parameters, but do not address whether these solutions satisfy constraints beyond the training data.

A separate line of work quantifies how many parameters suffice for interpolation. For real-analytic activations—the setting of our main result—Madden & Thrampoulidis [6] proved that two-layer networks can memorize up to $\lfloor md/2 \rfloor$ data points (m hidden neurons, input dimension d), and Madden [7] extended this to deeper architectures. Most relevant to our proof, Constantinescu & Popescu [8] proved that feedforward networks of arbitrary depth with non-affine activations in L^1_{loc} can interpolate arbitrary datasets—a result that subsumes the real-analytic case and is the interpolation theorem we invoke. For threshold and ReLU activations, Vershynin [9] established interpolation with $\Theta(md)$ parameters, and Yun et al. [10] gave tight bounds for shallow ReLU networks; these results show that the interpolation phenomenon is broad, though ReLU networks fall outside the scope of our theorem due to the non-analyticity of the ReLU.

The practical inadequacy of soft constraints is well documented. Beucler et al. [3] systematically compared architectural constraints (ACnet, which enforces conservation laws by construction) with loss-function penalties (LCnet), finding that only the former achieved exact satisfaction. Several methods enforce hard constraints, whether by construction or by optimization: Lu et al. [11] proposed hPINNs, combining penalty and augmented

Lagrangian methods; Zhong et al. [12] developed constrained neural fields for differential constraints of arbitrary order; Balestriero & LeCun [13] provided provably optimal affine constraint enforcement; and Donti et al. [14] introduced a differentiable procedure for constraint satisfaction in optimization. Most recently, Min & Azizan [2] proposed a framework that enforces input-dependent linear equality and inequality constraints while retaining universal approximation guarantees. These methods share a common motivation: unconstrained training fails to satisfy constraints. Our result shows that this failure is not incidental—without architectural inductive bias, the set of parameters satisfying a constraint has measure zero on the manifold of global minima.

The identity theorem for real-analytic functions has appeared in neural network theory in several contexts. Sontag [15] used the fact that the zero set of a non-trivial real-analytic function has measure zero (via analytically thin sets) to prove that, for generic regression data, the set of critical points of the least-squares loss is countable—and finite under stronger assumptions. Madden & Thrampoulidis [6] used the identity theorem in two steps: first, to lift results from polynomial to real-analytic activations (by truncating Taylor expansions and invoking the identity theorem); second, to convert "rank condition function not identically zero" into "generically full rank." Crăciun & Ghoshdastidar [16] proved that the gradient descent map for networks with piecewise analytic activations is non-singular—it preserves sets of measure zero under preimages—establishing this through a layered analysis of almost-everywhere analytic functions, combined with Fubini's theorem and the identity theorem. In contrast, our proof applies the identity theorem twice, in different variables: first to x, showing that constraint violation is all-or-nothing for each parameter, then to θ, showing that the all-satisfied case has measure zero on the manifold of global minima.

A body of work studies what gradient descent implicitly prefers. Jacot et al. [17] showed that in the infinite-width limit, training dynamics are governed by the neural tangent kernel, which characterizes which functions are learned efficiently. Rahaman et al. [18] demonstrated spectral bias: deep networks learn low-frequency components first. Razin & Cohen [19] showed that implicit regularization in deep learning may not be explainable by norm minimization. These works characterize what gradient descent finds. Our result characterizes what it does not: the set of parameters satisfying a given constraint, which has measure zero on the manifold of global minima.

3. Preliminaries

3.1. Real-Analytic Functions and the Identity Theorem

Definition 1. *A function $h : \Omega \to \mathbb{R}$ on an open set $\Omega \subseteq \mathbb{R}^n$ is* real-analytic *if for every $a \in \Omega$, h admits a convergent power series expansion in a neighborhood of a.*

Real-analytic functions are closed under composition and arithmetic operations. Consequently, a feedforward network $f_\theta(x)$ built from real-analytic activations (sigmoid, tanh, GELU, Swish, Mish) and affine maps is real-analytic in x for each fixed θ, and real-analytic in θ for each fixed x.

Theorem 1 (Identity Theorem)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be open and connected, and let $h : \Omega \to \mathbb{R}$ be real-analytic. If h vanishes on a set of positive Lebesgue measure, then $h \equiv 0$ on Ω.*

Proof sketch. For $n = 1$, if $h \not\equiv 0$, then at every zero a there exists a smallest k such that the k-th Taylor coefficient at a is nonzero, so $h(x) = (x - a)^k q(x)$ with $q(a) \neq 0$; by continuity, a is an isolated zero. (If all Taylor coefficients vanished at some point a, the set of such points

would be both open and closed in Ω, forcing $h \equiv 0$ by connectedness.) Isolated zeros form a countable, hence measure-zero, set.

For $n > 1$, work on a local rectangular neighborhood of any point in $h^{-1}(0)$. By Fubini's theorem, the positive measure of $h^{-1}(0)$ implies that for a positive-measure set of x_1-values, the section $x' \mapsto h(x_1, x')$ vanishes on a set of positive measure. By induction, these sections are identically zero. Fixing x', the function $x_1 \mapsto h(x_1, x')$ then vanishes on a positive-measure set; by the one-dimensional case, it is identically zero. Hence, $h \equiv 0$ on this neighborhood, and by analytic continuation and connectedness, $h \equiv 0$ on all of Ω. For a textbook treatment, see Krantz & Parks [20]. □

The contrapositive: A real-analytic function that is not identically zero vanishes only on a set of measure zero. This all-or-nothing property is the engine of our main result.

3.2. The Interpolating Manifold

Consider a feedforward neural network $f_\theta : \Omega \to \mathbb{R}$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$, with real-analytic activation functions. Given training data $\{(x_i, y_i)\}_{i=1}^m \subset \Omega \times \mathbb{R}$, define the evaluation map

$$F : \Theta \to \mathbb{R}^m, \quad F(\theta) = (f_\theta(x_1), \ldots, f_\theta(x_m)), \tag{1}$$

which stacks the m network outputs into a single vector, and is real-analytic in θ. The set of interpolating parameters is the fiber

$$M = F^{-1}(y) = \{\theta \in \Theta : f_\theta(x_i) = y_i,\ i = 1, \ldots, m\}, \tag{2}$$

where $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. We consider the mean squared error loss

$$L(\theta) = \frac{1}{2} \sum_{i=1}^{m} (f_\theta(x_i) - y_i)^2 = \frac{1}{2} \|F(\theta) - y\|^2. \tag{3}$$

Since $L(\theta) \geq 0$, any $\theta \in M$ attains $L(\theta) = 0$, the global minimum. If $M \neq \varnothing$, then the set of global minima of L is exactly M. The nonemptiness of M is not automatic; it is guaranteed by the interpolation capacity assumption (Assumption 1 below).

Theorem 2 (Cooper, 2021 [1])**.** *If $p > m$, then generically (i.e., possibly after an arbitrarily small perturbation of the training label y), y is a regular value of F. Consequently, $M = F^{-1}(y)$ is, if nonempty, a smooth submanifold of $\mathbb{R}^p$ of dimension $p - m$ (possibly disconnected).*

The regular value condition means that the Jacobian $DF(\theta) \in \mathbb{R}^{m \times p}$ has full row rank m at every $\theta \in M$: the m interpolation conditions are independent. Intuitively, overparameterization ($p > m$) provides enough degrees of freedom that these constraints do not degenerate.

By Sard's theorem, for any fixed set of distinct inputs $\{x_i\}_{i=1}^m$, the set of target vectors $y \in \mathbb{R}^m$ for which y fails to be a regular value of F has Lebesgue measure zero. Thus, if the training labels are drawn from a distribution that is absolutely continuous with respect to Lebesgue measure (conditional on the inputs), y is almost surely a regular value.

For our purposes, a stronger property of M is needed. Since F is real-analytic and y is a regular value, the real-analytic implicit function theorem applies: locally, M can be parameterized by real-analytic functions. This makes M a *real-analytic submanifold* of $\mathbb{R}^p$, not merely a smooth one. The consequence is that the restriction of any real-analytic function on $\mathbb{R}^p$ to M remains real-analytic, so the identity theorem applies to functions on M.

Throughout the paper, "measure zero in M" refers to the $(p-m)$-dimensional Hausdorff measure on M, which coincides with the Riemannian volume induced by the Euclidean metric of the ambient space $\mathbb{R}^p$. In any local real-analytic coordinate chart, this measure is equivalent to the Lebesgue measure on $\mathbb{R}^{p-m}$ via a real-analytic density, so the identity theorem for real-analytic functions (which is typically stated for Lebesgue measure on Euclidean space) applies without modification.

Since $p > m$, the manifold M has positive dimension $p - m$ when nonempty. We will additionally assume M is connected (Assumption 2 below), a non-trivial condition discussed in Section 4.

Assumption 1 (Interpolation Capacity)**.** *For any $m+1$ distinct inputs $\{x_1, \ldots, x_m, x_0\} \subset \Omega$ and any desired outputs $y_1, \ldots, y_m, y_0 \in \mathbb{R}$, there exists $\theta \in \Theta$ such that $f_\theta(x_i) = y_i$ for $i = 1, \ldots, m$ and $f_\theta(x_0) = y_0$.*

This assumption is stronger than the overparameterization condition $p > m$ and does not follow from parameter counting alone. It is satisfied by sufficiently wide overparameterized networks: Constantinescu & Popescu [8] proved that feedforward networks of arbitrary depth with non-affine activations in L^1_{loc} (a class that includes all real-analytic activations) can interpolate arbitrary datasets. For vector-valued outputs $f_\theta : \Omega \to \mathbb{R}^k$, this assumption is understood to require interpolation of $m+1$ arbitrary points in $\mathbb{R}^k$; see Section 4.4 for details.

4. Constraint Unlearnability

Let $\Omega \subseteq \mathbb{R}^n$ be open and connected. Let $f_\theta : \Omega \to \mathbb{R}$ and $M \subseteq \mathbb{R}^p$ be the network and interpolating manifold of Section 3. We additionally assume that the network has the interpolation capacity for $m+1$ points (Assumption 1). Let $g : \mathbb{R} \to \mathbb{R}$ be a real-analytic function, defining the pointwise constraint

$$g(f_\theta(x)) = 0, \quad \forall x \in \Omega. \tag{4}$$

We assume throughout that the training labels are consistent with the constraint:

$$g(y_i) = 0, \quad i = 1, \ldots, m. \tag{5}$$

This condition ensures that M_{good} is not trivially empty: if $g(y_i) \neq 0$ for some i, then every $\theta \in M$ (i.e., every exactly interpolating parameter) satisfies $g(f_\theta(x_i)) = g(y_i) \neq 0$, so $M_{\text{good}} = \varnothing$ and the measure-zero conclusion holds vacuously. Under condition (5), exact interpolation ($f_\theta(x_i) = y_i$ for $\theta \in M$) implies $g(f_\theta(x_i)) = 0$ at every training point, so the constraint is satisfied on the training set—yet, as the Main Theorem shows, it is violated almost everywhere else in Ω. We note that the nonemptiness of M itself is guaranteed by Assumption 1, which is independent of the labels' consistency with the constraint.

Assumption 2 (Connectivity of M)**.** *The interpolating manifold $M = F^{-1}(y)$ is connected.*

This assumption is non-trivial. Cooper's theorem gives M as a smooth submanifold, but not its connectivity. Kuditipudi et al. [21] constructed datasets for which the locus of global minima is disconnected, so the assumption does not hold in full generality. However, connectivity has been observed empirically for generic datasets and sufficiently overparameterized networks [22,23], and partial theoretical support is provided by Simsek et al. [5], who proved that the expansion manifold $\Theta_{r^* \to m}(\theta^*)$ is connected for any C^1 activation, and that $M = \Theta_{r^* \to m}(\theta^*)$ under additional technical conditions on the activation. A com-

plete characterization of when M is connected for real-analytic activations remains an open problem.

Definition 2 (Inductive Bias)**.** *The architecture f_θ has an* inductive bias *for the constraint g if $g(f_\theta(x)) = 0$ for all $\theta \in \Theta$ and all $x \in \Omega$—that is, the constraint is satisfied regardless of parameter choice. If there exist $\theta \in \Theta$ and $x \in \Omega$ with $g(f_\theta(x)) \neq 0$, the architecture is* without inductive bias *for g.*

This is a strictly *architectural* notion: the constraint is built into the network structure, not imposed by the learning algorithm. In the broader machine learning literature, "inductive bias" often encompasses algorithmic preferences such as implicit regularization (e.g., spectral bias, the neural tangent kernel), which influence which solutions gradient descent finds among those reachable. Our definition excludes such algorithmic effects; we discuss the relationship between architectural bias and implicit regularization in Section 5.

Define

$$M_{\text{good}} = \{\theta \in M : g(f_\theta(x)) = 0,\ \forall x \in \Omega\}, \tag{6}$$

the subset of M on which the network satisfies the constraint at every input. We prove that without inductive bias, M_{good} has measure zero in M.

4.1. All-or-Nothing for Pointwise Constraints

Proposition 1. *For any fixed $\theta \in \Theta$, define $h_\theta : \Omega \to \mathbb{R}$ by $h_\theta(x) = g(f_\theta(x))$. Then either $h_\theta \equiv 0$ on Ω, or $h_\theta^{-1}(0)$ has Lebesgue measure zero in Ω.*

Proof. The network $f_\theta(x)$ is a composition of affine maps and real-analytic activation functions, and thus is real-analytic in x; g is real-analytic by assumption. Thus, $h_\theta = g \circ f_\theta$ is real-analytic on Ω. Since Ω is open and connected, the identity theorem gives either $h_\theta \equiv 0$ on Ω, or $h_\theta^{-1}(0)$ has measure zero. □

Remark 1. *This proposition holds for every $\theta \in \Theta$, independently of the training data. It rules out partial satisfaction: no parameter can make the network satisfy the constraint on a set of positive measure without satisfying it everywhere.*

4.2. Measure Zero of the Satisfying Parameters

The argument applies the identity theorem a second time, now to a function of θ on the manifold M.

Choose a test point $x_0 \in \Omega \setminus \{x_1, \ldots, x_m\}$. Define the evaluation map

$$\phi : M \to \mathbb{R}, \quad \phi(\theta) = f_\theta(x_0). \tag{7}$$

Lemma 1. *ϕ is real-analytic on M.*

Proof. The network output $f_\theta(x_0)$ is real-analytic in θ on Θ, as a composition of affine maps in θ with real-analytic activation functions. Since M is a real-analytic submanifold of $\mathbb{R}^p$ (Section 3), the restriction $\phi = f_\cdot(x_0)|_M$ preserves real-analyticity. □

Theorem 3 (Main Theorem)**.** *If the training labels satisfy $g(y_i) = 0$ for $i = 1, \ldots, m$ (condition (5)), the architecture is without inductive bias for g (Definition 2), the network has the interpolation capacity for $m + 1$ points (Assumption 1), and M is connected (Assumption 2), then M_{good} has measure zero in M.*

Proof. The strategy is to show that $g \circ \phi$ is real-analytic and not identically zero on M, then apply the identity theorem. By Lemma 1, ϕ is real-analytic on M; since g is real-analytic, the composition $g \circ \phi$ is real-analytic on M.

To see that $g \circ \phi \not\equiv 0$, note that since the architecture is without inductive bias (Definition 2), there exist $\theta' \in \Theta$ and $x_0' \in \Omega$ with $g(f_{\theta'}(x_0')) \neq 0$. Set $y_0 = f_{\theta'}(x_0')$. Consider the augmented dataset $\{(x_i, y_i)\}_{i=1}^m \cup \{(x_0, y_0)\}$ of $m+1$ points, where x_0 is the test point chosen above. By the interpolation capacity assumption (Assumption 1), there exists $\theta^* \in \Theta$ such that

$$f_{\theta^*}(x_i) = y_i \ (i = 1, \ldots, m), \quad f_{\theta^*}(x_0) = y_0. \tag{8}$$

The first m conditions give $\theta^* \in M$; the last gives $\phi(\theta^*) = y_0$. Hence, $g(\phi(\theta^*)) = g(y_0) \neq 0$.

Since M is connected (Assumption 2) and $g \circ \phi$ is real-analytic and not identically zero, the identity theorem gives that $(g \circ \phi)^{-1}(0)$ has measure zero in M. Since $M_{\text{good}} \subseteq (g \circ \phi)^{-1}(0)$—if $\theta \in M_{\text{good}}$, then $g(f_\theta(x)) = 0$ for all $x \in \Omega$, in particular $g(\phi(\theta)) = 0$—we conclude that M_{good} has measure zero in M. □

Remark 2 (Component-wise extension when M is disconnected)**.** *If M is disconnected, the Main Theorem applies on each connected component separately. Let $M = \bigsqcup_\alpha M_\alpha$ be the decomposition into connected components. Since M is a submanifold of $\mathbb{R}^p$, each M_α is open in M, and there are at most countably many components. The proof of the Main Theorem shows that for any component M_α, if there exists $\theta_\alpha \in M_\alpha$ with $g(\phi(\theta_\alpha)) \neq 0$, then $g \circ \phi \not\equiv 0$ on M_α, and by the identity theorem on the connected set M_α, the set $(g \circ \phi)^{-1}(0) \cap M_\alpha$ has measure zero in M_α. Consequently, $M_{\text{good}} \cap M_\alpha$ has measure zero in M_α.*

The no-inductive-bias condition (Definition 2) and the interpolation capacity (Assumption 1) only guarantee the existence of one violating parameter θ^*, which belongs to some component M_{α^*}. These assumptions do not guarantee that every *component contains a violating parameter. A disconnected M could in principle admit a "perfect" component on which $g \circ \phi \equiv 0$, i.e., every interpolating parameter in that component satisfies the constraint everywhere. On such a component, M_{good} would have positive measure (in fact, equal to the entire component). Assumption 2 (connectivity of M) rules out this possibility by ensuring a single component.*

4.3. Attraction Basin and Constraint Unlearnability

We now prove that the set of initial parameters from which gradient flow converges to a constraint-satisfying solution has Lebesgue measure zero. This eliminates the gap between "measure zero on M" and "almost surely unlearnable."

Assumption 3 (Convergence of Gradient Flow)**.** *Consider the gradient flow $\dot{\theta} = -\nabla L(\theta)$ of the loss L defined in Section 3. The set*

$$B(M) := \left\{\theta_0 \in \mathbb{R}^p : \text{the trajectory } \theta(t) \text{ starting from } \theta_0 \text{ is bounded and } \lim_{t\to\infty} \theta(t) \in M\right\} \tag{9}$$

has full Lebesgue measure in $\mathbb{R}^p$.

This assumption formalizes the implicit condition, present in the original corollary, that gradient descent converges to M. It is supported by the large-width and NTK literature [17] and by the saddle-point avoidance framework [24].

Lemma 2 (Morse–Bott Nondegeneracy)**.** *Under the regular value condition of Cooper's theorem, M is a Morse–Bott nondegenerate attracting critical submanifold for the gradient flow: at each $\theta \in M$, the Hessian $\nabla^2 L(\theta)$ is positive-definite on the normal space $N_\theta M$.*

Proof. On M, we have $F(\theta) = y$, so $L|_M = 0$. Differentiating twice yields

$$\nabla^2 L(\theta) = DF(\theta)^\top DF(\theta) + \sum_{i=1}^{m} (f_\theta(x_i) - y_i)\, \nabla^2 f_\theta(x_i). \tag{10}$$

The second term vanishes on M because $f_\theta(x_i) = y_i$ for all i. Since $DF(\theta)$ has full row rank m (the regular value condition), the matrix $DF(\theta)^\top DF(\theta)$ is positive-semidefinite with kernel exactly $T_\theta M = \ker DF(\theta)$ and strictly positive eigenvalues on the normal space $N_\theta M$. □

Theorem 4 (Attraction Basin of Constraint-Satisfying Parameters)**.** *Under condition (5), Assumptions 1–3, and the no-inductive-bias condition (Definition 2), let $B(M_{\text{good}})$ be the set of initial parameters whose ω-limit set under the gradient flow is contained in M_{good}. Then $B(M_{\text{good}})$ has Lebesgue measure zero in $\mathbb{R}^p$.*

Proof. Fix a test point $x_0 \in \Omega \setminus \{x_1, \ldots, x_m\}$ and define $h(\theta) = g(f_\theta(x_0))$, which is real-analytic in θ. □

4.3.1. Local Limit Map

By Lemma 2, M is a normally hyperbolic attracting critical submanifold. For any compact subset $K \subset M$, the stable manifold theorem for real-analytic gradient flows [25] provides an open neighborhood $U_K \supset K$ and a real-analytic submersion $\pi_K : U_K \to M$ such that $\lim_{t\to\infty} \theta(t) = \pi_K(\theta_0)$ for all $\theta_0 \in U_K$.

4.3.2. Global Limit Map

Since M is a normally hyperbolic attracting submanifold (Lemma 2), its stable set $B(M)$ is open in $\mathbb{R}^p$. For any $\theta_0 \in B(M)$, the forward orbit is bounded (Assumption 3), so there exist $R > 0$ and $T \geq 0$ with $\theta(T) \in U_{M_R}$, where $M_R = M \cap \overline{B(0,R)}$. Define $\pi : B(M) \to M$ by $\pi(\theta_0) = \pi_{M_R}(\theta(T))$. Since the gradient flow map ϕ_T is real-analytic and the local projections agree on overlaps, π is a globally defined real-analytic submersion.

4.3.3. Zero Measure of the Basin

Define $H = h \circ \pi$ on $B(M)$; as a composition of real-analytic maps, H is real-analytic. As established in the proof of the Main Theorem, $h|_M \not\equiv 0$: there exists $\theta^* \in M$ with $h(\theta^*) \neq 0$. Since $\theta^* \in M \subset B(M)$ and $\pi(\theta^*) = \theta^*$, we have $H(\theta^*) \neq 0$, so $H \not\equiv 0$ on $B(M)$. Moreover, H is not identically zero on any connected component of $B(M)$: if $H \equiv 0$ on some component C, then $\pi(C)$ is open in M (since π is a submersion, hence an open map), and $h \equiv 0$ on $\pi(C)$; by the connectedness of M (Assumption 2) and the identity theorem, $h \equiv 0$ on M, a contradiction. Since $B(M)$ is open in $\mathbb{R}^p$, it has at most countably many connected components, and the identity theorem applies on each component to give that $H^{-1}(0) \cap C$ has measure zero for each component C. Hence, $H^{-1}(0)$ has Lebesgue measure zero in $\mathbb{R}^p$.

Now observe that $M_{\text{good}} \subseteq \{\theta \in M : h(\theta) = 0\}$, because if $g(f_\theta(x)) = 0$ for all $x \in \Omega$, then in particular $h(\theta) = g(f_\theta(x_0)) = 0$. Consequently,

$$B(M_{\text{good}}) \subseteq \pi^{-1}\big(\{\theta \in M : h(\theta) = 0\}\big) = H^{-1}(0) \cap B(M). \tag{11}$$

Since $H^{-1}(0)$ has measure zero, so does $B(M_{\text{good}})$.

Corollary 1 (Constraint Unlearnability)**.** *Assume the hypotheses of Theorem 4, and let the initial parameters be sampled from a distribution that is absolutely continuous with respect to Lebesgue measure. Then with probability one, the gradient flow converges to a parameter $\theta \in M$ for which the constraint $g(f_\theta(x)) = 0$ is violated for almost every $x \in \Omega$ (with respect to Lebesgue measure on Ω).*

Proof. From Assumption 3, $\Pr(\theta_0 \in B(M)) = 1$. Theorem 4 gives that $B(M_{\text{good}})$ has measure zero, so by absolute continuity of the initialization, $\Pr(\theta_0 \in B(M_{\text{good}})) = 0$. Thus, with probability one, the limit belongs to $M \setminus M_{\text{good}}$. For any $\theta \in M \setminus M_{\text{good}}$, the function $x \mapsto g(f_\theta(x))$ is real-analytic and not identically zero on Ω (otherwise θ would be in M_{good}), so by the identity theorem (Section 3.1), it vanishes on a set of measure zero in Ω. □

Remark 3 (Discrete gradient descent). *For discrete gradient descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$ with a sufficiently small step size $\eta > 0$, the trajectory closely shadows the gradient flow [see, e.g., [26]]. The non-singularity of the gradient descent map [16] ensures that zero-measure sets in the parameter space are preserved under finite-step iteration. Combined with the convergence assumption, the same conclusion follows: the set of initializations leading to M_{good} has measure zero, and constraint violation is almost sure.*

4.4. Extension to Vector-Valued Outputs

We have presented the theory for scalar output $f_\theta : \Omega \to \mathbb{R}$ to keep notation simple. The extension to vector-valued outputs $f_\theta : \Omega \to \mathbb{R}^k$ is straightforward and is outlined below.

4.4.1. Evaluation Map and Global Minimum Manifold

For k-dimensional outputs, the evaluation map stacks all k components at all m training points:

$$F : \Theta \to \mathbb{R}^{mk}, \quad F(\theta) = \left(f_\theta^1(x_1), \ldots, f_\theta^k(x_1), \ldots, f_\theta^1(x_m), \ldots, f_\theta^k(x_m)\right), \tag{12}$$

where f_θ^j denotes the j-th component of the output. The loss is $L(\theta) = \frac{1}{2}\|F(\theta) - y\|^2$, with $y \in \mathbb{R}^{mk}$ stacking the target vectors. The interpolating manifold is $M = F^{-1}(y)$. Cooper's theorem applies with the overparameterization condition $p > mk$, giving $\dim M = p - mk$ when $M \neq \varnothing$. The regular value condition means that $DF(\theta) \in \mathbb{R}^{mk \times p}$ has full row rank mk.

4.4.2. Interpolation Capacity

For vector-valued outputs, Assumption 1 is understood in the natural vector-valued sense: for any $m+1$ distinct inputs $\{x_1, \ldots, x_m, x_0\} \subset \Omega$ and any desired output vectors $y_1, \ldots, y_m, y_0 \in \mathbb{R}^k$, there exists $\theta \in \Theta$ such that $f_\theta(x_i) = y_i$ for $i = 1, \ldots, m$ and $f_\theta(x_0) = y_0$. Adding one test point thus imposes k additional interpolation conditions, requiring the network to accommodate $(m+1)k$ constraints in total. This is stronger than Cooper's dimensional condition $p > mk$, but is still guaranteed for sufficiently wide networks: Constantinescu & Popescu [8] prove interpolation for scalar outputs, and the vector-valued extension follows from the standard multi-output construction in which shared hidden layers are followed by k independent output heads; the hidden representation provides enough capacity to interpolate the augmented dataset, and the k output heads, being linear readouts from a shared feature map, can be chosen independently to match the k target components.

4.4.3. Pointwise Constraint

The constraint is defined by a real-analytic function $g : \mathbb{R}^k \to \mathbb{R}$, and the condition is $g(f_\theta(x)) = 0$ for all $x \in \Omega$. This formulation covers both component-wise and cross-component constraints, such as probability normalization ($g(z) = \sum_{j=1}^k z_j - 1$ for $z = (z_1, \ldots, z_k) \in \mathbb{R}^k$). The composition $h_\theta(x) = g(f_\theta(x))$ remains real-analytic in both x and θ because real-analytic functions are closed under composition.

4.4.4. Proofs Carry over Unchanged

With these adjustments, all proofs in Section 4 apply with m replaced by mk wherever it appears. In particular,

- The identity theorem applies to $h_\theta(x)$ on the connected open set Ω, giving the all-or-nothing dichotomy.
- The test-point construction in the Main Theorem uses the vector-valued interpolation capacity to produce $\theta^* \in M$ with $g(f_{\theta^*}(x_0)) \neq 0$.
- The Morse–Bott nondegeneracy (Lemma 2) holds because $DF(\theta)^\top DF(\theta)$ has kernel of dimension $p - mk$ and is positive-definite on the normal space.
- The stable manifold theorem and the attraction basin argument (Theorem 4) are unchanged, as they depend only on the dimension of M and the real-analyticity of the loss.

Thus, all results extend to the vector-output case under the strengthened dimensional and interpolation assumptions. We restrict the results to scalar output in the main text solely for notational clarity.

5. Discussion

5.1. Inductive Bias in Practice

Definition 2 formalizes a distinction that the practice of deep learning has long relied on: softmax, hard-constraint layers, and architectural enforcement [2,3,11,12] all ensure constraints by construction rather than by training. Our theorem provides the theoretical explanation for the pointwise-equality instances of this practice—without inductive bias, exact constraint satisfaction is a measure-zero event, not merely a training failure. The result clarifies that expressivity (the constraint-satisfying function exists in the hypothesis class) does not imply learnability (gradient descent finds it): the constraint-satisfying parameters occupy measure zero on M, so a convergent trajectory almost surely misses them. Figure 1 illustrates this contrast on a controlled experiment: the same network trained with and without softmax shows constraint violation almost everywhere in Ω versus satisfaction everywhere.

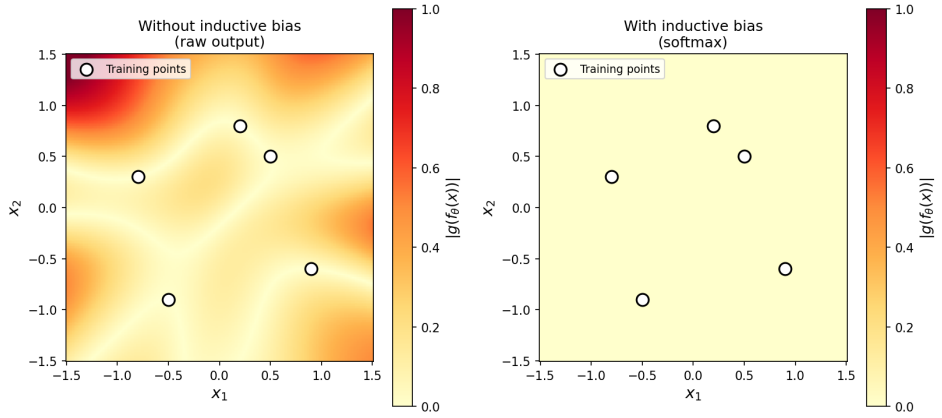


Figure 1. Constraint unlearnability in a controlled experiment. Two networks with identical architecture (two-layer tanh, 64 hidden units, input $\Omega = [-1.5, 1.5]^2 \subset \mathbb{R}^2$) are trained on the same five data points with labels satisfying the probability normalization constraint $p_1 + p_2 = 1$. (**Left**): Without inductive bias, the network outputs raw (p_1, p_2) and is trained with MSE. After convergence (training loss $< 10^{-10}$), the constraint $g(p_1, p_2) = p_1 + p_2 - 1$ is satisfied at the training points (white markers) but violated almost everywhere else on the input domain (brighter colors indicate larger $|g(f_\theta(x))|$). (**Right**): With inductive bias, the network applies softmax to its output. The constraint $p_1 + p_2 = 1$ is satisfied everywhere (to machine precision, $\sim 10^{-17}$), by construction rather than by learning. This contrast illustrates the Main Theorem: without architectural enforcement, exact constraint satisfaction on the entire input domain is a measure-zero event, even though the network interpolates the training data perfectly.

5.2. Implicit Regularization and Constraints

A natural question is whether implicit regularization—the tendency of gradient descent to prefer certain solutions—could substitute for inductive bias. Spectral bias [18] and the neural tangent kernel [17] show that gradient descent preferentially learns low-frequency, simple functions.

Constraint satisfaction, however, is structurally different. The set M_{good} is not a region that gradient descent avoids by preference; it is a measure-zero subset that gradient descent almost surely never reaches. Implicit regularization operates within the positive-measure set $M \setminus M_{\text{good}}$, selecting solutions with certain regularity properties, but it does not redirect the trajectory toward the zero-measure M_{good}.

Whether implicit regularization can achieve *approximate* constraint satisfaction—$\|g \circ f_\theta\|_\infty < \epsilon$ for small ϵ—is a separate question that our results do not address. The ϵ-neighborhood of M_{good} may have positive measure, and characterizing whether gradient descent enters it requires tools beyond the identity theorem.

5.3. Limitations

5.3.1. Activation Functions

The proof relies on the identity theorem for real-analytic functions, which requires real-analytic activations. This class includes sigmoid, tanh, GELU, Swish, and Mish, but excludes ReLU, which is piecewise-linear and not real-analytic. We note that approximate implementations of GELU (e.g., using tanh) are also real-analytic, so the theorem applies to them as well.

For ReLU networks, the identity theorem does not globally apply, but a local analog survives. Fixing θ, the input space is partitioned into finitely many linear regions on each of which $h_\theta(x) = g(f_\theta(x))$ is real-analytic (composition of a real-analytic g with an affine map). The identity theorem holds within each region: either h_θ vanishes identically on that region, or its zero set has measure zero there. The global all-or-nothing dichotomy breaks because h_θ may vanish on some regions and not others. Consequently, a single violating parameter θ^* guarantees non-triviality only on the activation pattern region containing it, not across all regions.

The framework of Crăciun & Ghoshdastidar [16] for piecewise analytic activations establishes non-singularity of the gradient descent map (used in Remark 3 for the dynamical part of our argument), but does not directly address this static question. Whether the measure-zero conclusion for M_{good} can be recovered for ReLU networks—either by a piecewise argument exploiting the local validity of the identity theorem within each activation pattern, or by fundamentally different tools—remains an important open problem.

5.3.2. Pointwise vs. Functional Constraints

Our framework covers pointwise constraints of the form $g(f_\theta(x)) = 0$ for each x independently. Many important constraints are functional: equivariance ($f_\theta(\rho x) = \rho f_\theta(x)$), monotonicity ($x_1 < x_2 \Rightarrow f_\theta(x_1) < f_\theta(x_2)$), and differential constraints ($\nabla \cdot f_\theta = 0$) relate values at different inputs. These require a different mathematical treatment, as they constrain the function as a whole rather than each output independently.

5.3.3. Exact vs. Approximate Satisfaction

The theorem concerns exact constraint satisfaction on the entire input domain. In practice, approximate satisfaction within a tolerance ϵ is often sufficient. Our measure-zero result does not directly imply that approximate satisfaction fails: the ϵ-approximate satisfying set may have positive measure. Extending the analysis to approximate constraints is an open problem.

5.3.4. Interpolation Capacity

The Main Theorem assumes that the network can interpolate $m + 1$ arbitrary points (Assumption 1). This is stronger than the overparameterization condition $p > m$ and does not follow from parameter counting alone. For feedforward networks with non-affine real-analytic activations, this property is guaranteed by Constantinescu & Popescu [8] for sufficiently wide networks. The tight architectural conditions (minimum width or depth) for interpolation capacity remain an active area of study [6,7].

5.3.5. Connectivity of M

The Main Theorem assumes M is connected (Assumption 2); see the discussion following the assumption for its non-triviality and partial theoretical support. When M is disconnected, the measure-zero conclusion applies component-wise (Remark 2): on any connected component containing at least one violating parameter, the set of constraint-satisfying parameters has measure zero. However, a disconnected M could in principle contain a "perfect" component on which the constraint holds identically; such a component would have positive measure. Assumption 2 excludes this scenario. Whether perfect components can actually occur for real-analytic networks without inductive bias, and whether gradient descent can reach them if they do, are open questions.

5.3.6. Convergence of Gradient Flow

Corollary 1 relies on Assumption 3, which requires that gradient flow almost always converges to M. This property is supported by empirical observations and by theoretical results for sufficiently wide networks [17,24], but a complete proof for real-analytic overparameterized networks remains an open problem. The Morse–Bott nondegeneracy of M (Lemma 2) ensures that M is a normally hyperbolic attracting submanifold, so convergence holds locally near M; the global assumption concerns whether trajectories reach this neighborhood.

6. Conclusions

We have shown that overparameterized neural networks with real-analytic activations cannot learn to satisfy pointwise constraints exactly without architectural inductive bias. The proof hinges on a simple geometric fact: the identity theorem for real-analytic functions forces constraint satisfaction to be all-or-nothing, and the no-inductive-bias condition, combined with interpolation capacity, ensures that the "all" case occupies measure zero on the interpolating manifold. By lifting this zero-measure property from M to the parameter space via the Morse–Bott stable manifold theorem, we prove that the attraction basin of constraint-satisfying parameters has measure zero, so gradient descent almost surely converges to parameters that violate the constraint almost everywhere in Ω.

The practical implication is that architectural inductive biases for pointwise equality constraints—such as softmax for probability normalization—are not engineering conveniences but mathematical necessities for exact satisfaction. Their ubiquity in practice reflects this necessity, even when it has not been formally articulated.

The result is exact within its setting: real-analytic activations, pointwise constraints, exact satisfaction, and a connected interpolating manifold (see Section 5.3 for limitations). Extending constraint unlearnability beyond these boundaries defines a program for future work. The most pressing direction is whether an analog holds for ReLU networks, the most widely used in practice; the identity theorem has no direct replacement there, but the framework of Crăciun & Ghoshdastidar [16] for piecewise analytic activations may offer a path. A second direction is approximate satisfaction: whether implicit regularization can

drive the network into an ϵ-neighborhood of M_{good}, even though exact membership is a measure-zero event.

Author Contributions: Conceptualization, H.Q.; methodology, H.Q.; formal analysis, H.Q.; writing—original draft preparation, H.Q.; writing—review and editing, H.Q. and Z.T.; supervision, Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Acknowledgments: The authors thank the National Key Laboratory of Information Systems Engineering for institutional support.

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

References

1. Cooper, Y. Global minima of overparameterized neural networks. *SIAM J. Math. Data Sci.* **2021**, *3*, 676–691. [CrossRef]
2. Min, Y.; Azizan, N. HardNet: Hard-constrained neural networks with universal approximation guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2025.
3. Beucler, T.; Pritchard, M.; Rasp, S.; Ott, J.; Baldi, P.; Gentine, P. Enforcing analytic constraints in neural-networks emulating physical systems. *Phys. Rev. Lett.* **2021**, *126*, 098302. [CrossRef] [PubMed]
4. Nguyen, Q.; Hein, M. The loss surface of deep and wide neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; Volume PMLR 70, pp. 2603–2612.
5. Simsek, B.; Ged, F.; Jacot, A.; Spadaro, F.; Hongler, C.; Gerstner, W.; Brea, J. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume PMLR 139, pp. 9722–9732.
6. Madden, L.; Thrampoulidis, C. Memory capacity of two layer neural networks with smooth activations. *SIAM J. Math. Data Sci.* **2024**, *6*, 679–702. [CrossRef]
7. Madden, L. Interpolation with deep neural networks with non-polynomial activations: Necessary and sufficient numbers of neurons. *arXiv* **2024**, arXiv:2405.13738.
8. Constantinescu, V.-R.; Popescu, I. Approximation and interpolation of deep neural networks. *arXiv* **2024**, arXiv:2304.10552.
9. Vershynin, R. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM J. Math. Data Sci.* **2020**, *2*, 1004–1033. [CrossRef]
10. Yun, C.; Sra, S.; Jadbabaie, A. Small ReLU networks are powerful memorizers: A tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2019.
11. Lu, L.; Pestourie, R.; Yao, W.; Wang, Z.; Verdugo, F.; Johnson, S.G. Physics-informed neural networks with hard constraints for inverse design. *SIAM J. Sci. Comput.* **2021**, *43*, B1105–B1132. [CrossRef]
12. Zhong, F.; Fogarty, K.; Hanji, P.; Wu, T.; Sztrajman, A.; Spielberg, A.; Tagliasacchi, A.; Bosilj, P.; Oztireli, C. Neural fields with hard constraints of arbitrary differential order. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2023.
13. Balestriero, R.; LeCun, Y. POLICE: Provably optimal linear constraint enforcement for deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
14. Donti, P.L.; Rolnick, D.; Kolter, J.Z. DC3: A learning method for optimization with hard constraints. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
15. Sontag, E.D. Critical points for least-squares problems involving certain analytic functions, with applications to sigmoidal nets. *Adv. Comput. Math.* **1996**, *5*, 245–268. [CrossRef]
16. Crăciun, A.; Ghoshdastidar, D. Non-singularity of the gradient descent map for neural networks with piecewise analytic activations. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2025.
17. Jacot, A.; Gabriel, F.; Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2018.
18. Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.A.; Bengio, Y.; Courville, A. On the spectral bias of neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; Volume PMLR 97, pp. 5301–5310.
19. Razin, N.; Cohen, N. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2020.

20. Krantz, S.G.; Parks, H.R. *A Primer of Real Analytic Functions*, 2nd ed.; Birkhäuser: Boston, MA, USA, 2002.
21. Kuditipudi, R.; Wang, X.; Lee, H.; Zhang, Y.; Li, Z.; Hu, W.; Arora, S.; Ge, R. Explaining landscape connectivity of low-cost solutions for multilayer nets. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: San Diego, CA, USA, 2019.
22. Draxler, F.; Veschgini, K.; Salmhofer, M.; Hamprecht, F.A. Essentially no barriers in neural network energy landscape. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume PMLR 80, pp. 1309–1318.
23. Frankle, J.; Dziugaite, G.K.; Roy, D.M.; Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 12–18 July 2020; Volume PMLR 119, pp. 3259–3269.
24. Lee, J.D.; Simchowitz, M.; Jordan, M.I.; Recht, B. Gradient descent only converges to minimizers. In Proceedings of the Conference on Learning Theory (COLT), New York, NY, USA, 23–26 June 2016; Volume PMLR 49, pp. 1246–1257.
25. Hirsch, M.W.; Pugh, C.C.; Shub, M. *Invariant Manifolds*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1977; Volume 583.
26. Hairer, E.; Lubich, C.; Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2006.

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