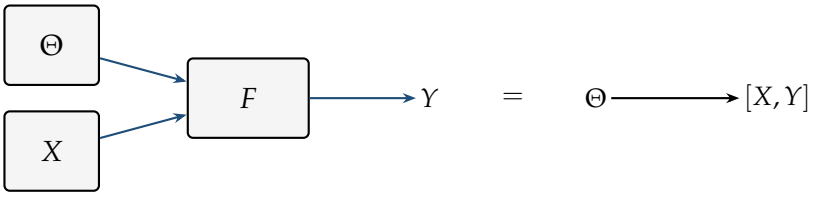
Higher Categorical Semantics

Parametrized Neural Networks and Their Representations in Differential Topology

Families of maps, transversality, surgery along training, and isotopy



Anonymous

2026

About this text. This monograph applies the results of *Differential Topology in Higher Categories* [DTHC], the higher-categorical treatment of M. W. Hirsch's *Differential Topology* [Hir76] in the same series, to smooth families of maps defined by neural network architectures. It is written to be read after that volume and cites it by theorem number, in the form [DTHC, 4.5]. It is an independent work and is not affiliated with or endorsed by the publisher of Hirsch's book.

Status of results. Every numbered statement carries a tag: classical (standard mathematics, attributed), literature (established in the machine learning or mathematical literature, attributed), reformulation (a known result restated or specialized, with the derivation given here), observation (an elementary consequence we have not found stated in this form), or programme (a direction, not a result). The statements tagged reformulation and observation are corollaries of classical differential topology applied to parametrized families; none is claimed as a new theorem of differential topology, and none says anything quantitative about training or generalization.

Authorship and assistance. The text was drafted and typeset with a large language model (Anthropic's Claude) under the direction of the submitting author, who is responsible for it. Generative AI assisted with drafting, literature recall, mathematical consistency checks, computation, and typesetting. The numerical examples and finite instances of the structural claims are recomputed by the accompanying script `verify-parametrized-networks.py` (Appendix B), which uses small networks implemented in pure Python. Proofs and bibliographic details have not been independently refereed; corrections are welcome.

Open access. Rights remain with the author.

Contents

Preface v

Introduction vii

I Families 1

1 Families of Maps 2
- 1.1 Parametrized maps . . . 2
- 1.2 Architectures . . . 3
- 1.3 Data, loss, and training . . . 3
- 1.4 The dictionary in brief . . . 4

2 Derivatives 6
- 2.1 Backpropagation . . . 6
- 2.2 The Jacobian of the evaluation map . . . 7
- 2.3 Rank . . . 7
- 2.4 Second derivatives . . . 8

3 Symmetries and the Moduli of Networks 9
- 3.1 Symmetries . . . 9
- 3.2 One hidden layer . . . 9
- 3.3 Homogeneous activations . . . 11
- 3.4 Invariant functions . . . 11

II Level Sets 13

4 Transversality and Decision Boundaries 14
- 4.1 The boundary is generically a hypersurface . . . 14
- 4.2 Several classes . . . 15
- 4.3 Chambers . . . 15
- 4.4 Every hypersurface is a decision boundary . . . 16
- 4.5 Piecewise-linear networks . . . 17

5 Topology Change along Paths: Surgery and Cobordism **19**
5.1 The trace of a path . . . 19
5.2 Surgery . . . 19
5.3 Genericity of paths . . . 21
5.4 What cobordism does and does not see . . . 22

6 The Loss Landscape **24**
6.1 The interpolation manifold . . . 24
6.2 Sublevel sets and barriers . . . 25
6.3 Why the loss is rarely a Morse function . . . 26
6.4 Critical points at infinity . . . 27

III Maps **28**

7 Embeddings, Isotopy, and Depth **29**
7.1 The first layer generically embeds . . . 29
7.2 What injective networks cannot do . . . 30
7.3 Neural ordinary differential equations . . . 30
7.4 Augmentation . . . 31
7.5 The space of representations . . . 32

8 Degree and Invertibility **34**
8.1 Global invertibility . . . 34
8.2 Topological mismatch . . . 35
8.3 Equilibrium layers . . . 35

IV Assessment **38**

9 What Was Used, and What Is Missing **39**
9.1 The inputs . . . 39
9.2 What is and is not new . . . 39
9.3 Limits . . . 40
9.4 Directions . . . 41

A Dictionary **42**

B Computations **44**

C Solutions to the Exercises **46**

References **52**

Preface

A neural network architecture is a smooth map $F\colon \Theta \times X \to Y$: a family of maps from inputs to outputs, parametrized by weights. Differential topology, in the form Hirsch gave it, is the study of what is true of a generic member of a family of maps and of what happens along a generic path in one. The fit is close enough that most of the qualitative questions asked about networks have answers that are corollaries of theorems proved before 1976, once the questions are phrased as questions about families.

Is the decision boundary of a classifier a manifold? For almost every value of the parameters, yes, by the parametric transversality theorem, provided the last layer has a bias. How does the boundary change during training? By ambient isotopy, except at isolated times, at each of which it undergoes one surgery whose type is the index of a critical point; this is Morse theory on the trace of the boundary. What is the set of parameters that fit the data exactly? For almost every labelling, a submanifold of dimension equal to the number of parameters minus the number of constraints, by the regular value theorem. Can a neural ordinary differential equation separate two linked circles? Not in dimension three, because its time-one map is ambient isotopic to the identity; in dimension four always, because there any two embeddings of a one-manifold are isotopic and isotopies extend. Each of these is a paragraph of argument on top of the companion volume, and this book supplies the paragraphs.

The higher-categorical layer of the companion volume enters in three places. A parametrized network is a plot of the mapping object $[X, Y]$ in smooth sets, and composition of layers is composition in a category of parametrized maps; backpropagation is the functoriality of the reverse derivative. The symmetries of an architecture make the parameter space an atlas for a stack whose points are networks up to relabelling, and the loss is a function on the stack. And the level sets that fail to be manifolds, at the exceptional parameters, are derived manifolds whose cobordism classes are still defined; the surgery description is what one sees of them from nearby.

The results are qualitative and the last chapter is frank about what that leaves out. Nothing here bounds a rate of convergence or a generalization error. Activation functions are assumed smooth except where stated, and the piecewise-linear case, which needs a different transversality theory, is treated only by reference. What the book offers is a dictionary, a set of theorems with

honest hypotheses, and a script that checks the examples.

Introduction

What is assumed

The companion volume [DTHC], at least Chapters 1, 3, 4, 8 and 10, or the corresponding chapters of Hirsch's book. From machine learning, only the definition of a feedforward network and of gradient descent.

How the book is organized

Part I sets up families, their derivatives and their symmetries. Part II studies level sets: of the network in input space (decision boundaries), of the network along a path of parameters (surgery), and of the loss in parameter space (the interpolation manifold). Part III studies the maps themselves: when they are embeddings, when they are isotopic, when they are invertible. Part IV is an assessment. Appendix A gives the dictionary between the two subjects in one table.

Conventions

Manifolds and maps are smooth. A *family* is a smooth map $F\colon \Theta \times X \to Y$; we write $f_\theta = F(\theta, -)$. $P = \dim \Theta$ is the number of parameters, $d = \dim X$ the input dimension, n the number of data points. "Almost every" refers to Lebesgue measure on a Euclidean parameter space, and "generic" means belonging to a residual set whose complement has measure zero. An activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is smooth unless it is explicitly $\mathrm{ReLU}(s) = \max(s, 0)$. Citations of the form [DTHC, 4.5] refer to numbered statements of the companion volume.

Part I

Families

Chapter 1

Families of Maps

1.1 Parametrized maps

Definition 1.1. A ***parametrized map*** from X to Y is a pair (Θ, F) of a manifold Θ and a smooth map $F\colon \Theta \times X \to Y$. Its ***members*** are the maps $f_\theta = F(\theta, -)$, and its ***realization map*** is $\theta \mapsto f_\theta$, from Θ to $C^\infty(X, Y)$.

By [DTHC, Thm. 1.6] a parametrized map is exactly a plot of the mapping object $[X, Y]$ in smooth sets, with domain Θ; the realization map is the function it induces on points. Everything in this book is a statement about plots of mapping objects. The topologies on $C^\infty(X, Y)$ play no direct role, since parameter spaces are finite-dimensional; what is used is the relation between properties of the total map F and properties of its members.

Definition 1.2. Para is the bicategory whose objects are manifolds, whose morphisms $X \to Y$ are parametrized maps, and whose 2-cells $(\Theta', F') \Rightarrow (\Theta, F)$ are smooth maps $r\colon \Theta' \to \Theta$ with $F' = F \circ (r \times \mathrm{id}_X)$. The composite of $(\Theta, F)\colon X \to Y$ and $(\Theta', G)\colon Y \to Z$ is

$$\big(\Theta \times \Theta',\ (\theta, \theta', x) \mapsto G(\theta', F(\theta, x))\big).$$

Composition is associative up to the canonical reassociation of products, and the identity of X is parametrized by a point. The construction is due to Fong, Spivak and Tuyéras [FST19] and is developed in [CGGWZ22]. A 2-cell is a reparametrization: sharing weights between layers is the 2-cell induced by a diagonal map, and freezing a weight is the one induced by the inclusion of a slice.

Proposition 1.3. [reformulation; cf. [DTHC, Thm. 1.6]; [FST19]] *Composition in* Para *is the internal composition $[X, Y] \times [Y, Z] \to [X, Z]$ of smooth sets, evaluated on plots: the composite of two parametrized maps is the plot $\Theta \times \Theta' \to [X, Y] \times [Y, Z] \to [X, Z]$. Hence* Para *is the bicategory of plots of the category of manifolds enriched in smooth sets.*

Proof. The internal composition is the map of sheaves which on plots with domain U sends (F, G) to $(u, x) \mapsto G(u, F(u, x))$. Precomposing with the product plot, whose components depend on separate factors of $U = \Theta \times \Theta'$, gives the formula of Definition 1.2. □

1.2 Architectures

Definition 1.4. A ***feedforward architecture*** with widths $(n_0, \dots, n_L)$ and activation σ is the composite in Para of the layers

$$\ell_i \colon \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}, \qquad \ell_i(W_i, b_i; x) = \sigma(W_i x + b_i) \qquad (i < L),$$
$$\ell_L \colon \mathbb{R}^{n_{L-1}} \to \mathbb{R}^{n_L}, \qquad \ell_L(W_L, b_L; x) = W_L x + b_L,$$

with σ applied to each coordinate. Its parameter space is $\Theta = \prod_i (\mathbb{R}^{n_i \times n_{i-1}} \times \mathbb{R}^{n_i})$, of dimension $P = \sum_i n_i(n_{i-1} + 1)$. The number of hidden units is $H = n_1 + \cdots + n_{L-1}$.

If σ is smooth the family is smooth, and real-analytic if σ is (as tanh, the logistic function and softplus are). For $\sigma = \mathrm{ReLU}$ the family is continuous and piecewise polynomial in (θ, x), and it is not a morphism of Para as defined; statements about it are made separately. Convolutional and attention layers are parametrized maps in the same sense, with 2-cells expressing weight sharing. Nothing below depends on the feedforward form except where an explicit hypothesis on the last or first layer is made.

The image of the realization map, the set $\mathcal{F} = \{f_\theta\}$ of functions the architecture can compute, is a subset of $C^\infty(X, Y)$. With the plots that factor locally through Θ it is a diffeological space, the quotient of Θ by the relation $f_\theta = f_{\theta'}$. It is in general not a manifold; Chapter 3 describes its singularities.

1.3 Data, loss, and training

Fix inputs $x_1, \dots, x_n \in X$ and targets $y_1, \dots, y_n \in Y = \mathbb{R}^m$. The ***evaluation map*** is

$$\mathrm{ev} \colon \Theta \to Y^n = \mathbb{R}^{nm}, \qquad \mathrm{ev}(\theta) = \big(f_\theta(x_1), \dots, f_\theta(x_n)\big),$$

and the squared loss is $\mathcal{L}(\theta) = \frac{1}{2}|\mathrm{ev}(\theta) - y|^2$. More generally a loss is $\mathcal{L} = \lambda \circ \mathrm{ev}$ for a smooth $\lambda \colon \mathbb{R}^{nm} \to \mathbb{R}$. A ***training path*** is a smooth curve $\gamma \colon [0, T] \to \Theta$; gradient flow is the curve with $\dot\gamma = -\nabla\mathcal{L}(\gamma)$, for the Euclidean metric on Θ. A path in Θ is a plot $[0, T] \to [X, Y]$: a one-parameter family of maps. When every member is an embedding of a submanifold $K \subset X$ it is an isotopy of K in the sense of [DTHC, Def. 10.1].

Three kinds of question organize the rest of the book.

(Q1) What does a generic member f_θ look like? (Chapters 4, 7 and 8)

(Q2) What happens to the members along a generic path? (Chapter 5)

(Q3) What does the set of parameters satisfying a condition, such as $\mathcal{L}(\theta) = 0$, look like? (Chapters 3 and 6)

In each case the answer comes from the same mechanism. A condition on a member, or on a parameter, is the condition that a point lies in the preimage of a submanifold under a map built from F; the preimage is a manifold when the map is transverse to the submanifold; and transversality of the map built from F can be checked on the architecture, usually by exhibiting one direction in parameter space, a bias, in which the map moves freely.

Principle 1.5 (Richness). [reformulation; cf. [DTHC, Thm. 4.5]] *Let $F\colon \Theta \times X \to Y$ be a family and $A \subset Y$ a submanifold. If F is transverse to A, in particular if F is a submersion, then f_θ is transverse to A for almost every θ. A feedforward architecture is a submersion onto $Y = \mathbb{R}^{n_L}$ because of the last-layer bias: $\partial F/\partial b_L = \mathrm{id}$.*

This is [DTHC, Thm. 4.5] with its hypothesis verified. It is worth stating separately because the hypothesis is about the family and not about function space. A statement that a property is generic among all smooth maps (for instance, that Morse functions are dense) says nothing about a given architecture, whose realization is a thin subset of all maps. Genericity transfers to an architecture only through a transversality property of F itself, and each use of Principle 1.5 below names the parameters that provide it.

1.4 The dictionary in brief

network notion	object of differential topology	Ch.
architecture	plot of a mapping object; morphism of Para	1
backpropagation	functoriality of the reverse derivative	2
weight-space symmetry	group action with invariant realization	3
decision boundary	preimage of a regular value	4
change of topology in training	handle attachment on the trace	5
interpolating parameters	fibre of the evaluation map	6
barrier between minima	death level in the sublevel barcode	6
hidden representation of data	immersion or embedding of a manifold	7
neural ODE, residual network	diffeotopy; isotopy of embeddings	7
invertible network	diffeomorphism of degree one	8
equilibrium of an implicit layer	transverse fixed point	8

Notes

The categorical treatment of learners as parametrized maps with update and request functions begins with [FST19]; the bicategory Para and its combination with lenses and reverse derivatives is in [CGGWZ22]. The point of view that a network is a family in the sense of singularity theory, to which transversality applies, is implicit in much work on loss landscapes, and explicit for piecewise-linear networks in [GL22].

Exercises

Exercise 1.1. Compute P and H for the architectures with widths $(2, 3, 1)$, $(2, 4, 4, 1)$ and $(d, k, 1)$.

Exercise 1.2. Verify that composition in Para is associative up to the reassociation $(\Theta \times \Theta') \times \Theta'' \cong \Theta \times (\Theta' \times \Theta'')$, and that this isomorphism is a 2-cell in both directions.

Exercise 1.3. Let $(\Theta, F) \colon X \to X$ be a layer. Express the weight-tied network $f_\theta \circ f_\theta$ as the source of a 2-cell into the composite $(\Theta \times \Theta, F \circ F)$, and show that its derivative with respect to θ is the sum of the two partial derivatives of the untied network on the diagonal.

Exercise 1.4. Give an example of a family $F \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ no member of which is a Morse function, although Morse functions are dense in $C^\infty(\mathbb{R}, \mathbb{R})$. Which hypothesis of Principle 1.5 fails for the relevant jet condition?

Chapter 2

Derivatives

Two derivatives of a family matter: with respect to the input, which governs the geometry of each member, and with respect to the parameters, which governs training. Both are instances of the tangent functor of [DTHC, Ch. 2], and the algorithm that computes the second efficiently is the functoriality of its transpose.

2.1 Backpropagation

Recall from [DTHC, Def. 2.10] the category Lens and from [DTHC, Prop. 2.11] the functor $f \mapsto (f, R[f])$, $R[f](x, w) = (Df_x)^{\mathsf{T}} w$, from Euclidean spaces to lenses. For a parametrized map $F\colon \Theta \times X \to Y$ the reverse derivative has two components,

$$R[F](\theta, x, w) = \big(R_\Theta[F](\theta, x, w),\ R_X[F](\theta, x, w)\big) = \big((D_\theta F)^{\mathsf{T}} w,\ (D_x F)^{\mathsf{T}} w\big),$$

the first being the update direction for the parameters and the second the covector passed back to the previous layer.

Theorem 2.1 (Backpropagation is a functor)**.** [literature; Fong–Spivak–Tuyéras [FST19]; Cruttwell et al. [CGGWZ22]] *The assignment* $(\Theta, F) \mapsto (\Theta, (F, R[F]))$ *is a functor of bicategories from* Para *over Euclidean spaces to parametrized lenses. In particular, for a composite of layers* $F = \ell_L \circ \cdots \circ \ell_1$ *the component* $R_{\Theta_i}[F]$ *is obtained by composing the lenses of the layers: a forward pass computing the activations* $x_i = \ell_i(\theta_i, x_{i-1})$*, and a backward pass computing* $w_{i-1} = R_X[\ell_i](\theta_i, x_{i-1}, w_i)$ *and emitting* $R_\Theta[\ell_i](\theta_i, x_{i-1}, w_i)$*.*

Proof. Composition in Para is composition of maps after taking products with identities, and R is a product-preserving functor to lenses by [DTHC, Prop. 2.11]. □

The gradient of a loss $\mathcal{L} = \lambda \circ \mathrm{ev}$ is $R[\mathrm{ev}](\theta, \nabla\lambda)$. The theorem says the cost of the gradient is that of one forward and one backward pass, whatever the number of parameters, which is the practical content of reverse-mode differentiation. The script implements both modes for a small network and checks them against each other and against finite differences.

2.2 The Jacobian of the evaluation map

Definition 2.2. For data $x_1, \dots, x_n$ the ***parameter Jacobian*** is $J(\theta) = D\mathrm{ev}_\theta$, an $nm \times P$ matrix. The ***tangent kernel*** is the $nm \times nm$ matrix $K(\theta) = J(\theta)J(\theta)^{\mathsf{T}}$.

Proposition 2.3. [classical; Jacot–Gabriel–Hongler [JGH18]] *$K(\theta)$ is symmetric positive semidefinite of rank* $\operatorname{rk} J(\theta)$. *Along gradient flow of $\mathcal{L} = \lambda \circ \mathrm{ev}$ the outputs $u(t) = \mathrm{ev}(\gamma(t))$ satisfy $\dot{u} = -K(\gamma)\nabla\lambda(u)$.*

Proof. $\dot{u} = J\dot{\gamma} = -J\nabla\mathcal{L} = -JJ^{\mathsf{T}}\nabla\lambda$. □

So the tangent kernel is the Riemannian metric that the Euclidean metric of Θ induces, by pushing forward along ev, on the cotangent directions of output space; where J is onto, K is invertible and the outputs follow a gradient flow for the metric K^{-1}.

2.3 Rank

The rank of J is the first invariant of a parameter. It is lower semicontinuous, and for analytic architectures its generic value is attained almost everywhere.

Proposition 2.4. [classical] *Let F be real-analytic in θ on a connected Θ, and let r be the maximum of* $\operatorname{rk} J(\theta)$. *Then $\{\theta : \operatorname{rk} J(\theta) = r\}$ is open and dense and its complement has measure zero. On it the fibres of* ev *are submanifolds of dimension $P - r$, and* ev *is locally a submersion onto an r-dimensional submanifold of $\mathbb{R}^{nm}$.*

Proof. The complement is the common zero set of the $r \times r$ minors of J, which are analytic and not all identically zero. The zero set of a non-zero analytic function on a connected domain is closed with empty interior and measure zero. The rest is the constant rank theorem. □

Definition 2.5. The ***functional dimension*** of the architecture at θ is the supremum over all finite sets of inputs of $\operatorname{rk} J(\theta)$: the rank of the differential of the realization map [GLMW22].

The functional dimension is at most P, and Chapter 3 shows when it is smaller. The strata $\{\operatorname{rk} J = s\}$ for $s < r$ are where the fibres of ev jump in dimension. They contain the parameters at which hidden units are redundant, and they are responsible for most of what is singular in the loss landscape (Chapter 6).

2.4 Second derivatives

By the composition law for 2-jets [DTHC, Prop. 2.9], the Hessian of a composite $\lambda \circ \mathrm{ev}$ is

$$\nabla^2 \mathcal{L}(\theta) = J^{\mathsf{T}} \, \nabla^2 \lambda \, J + \sum_k \frac{\partial \lambda}{\partial u_k} \, \nabla^2 \mathrm{ev}_k(\theta),$$

a Gauss–Newton term and a term weighted by the gradient of λ in output space.

Proposition 2.6. [classical] *At a parameter where $\nabla \lambda(\mathrm{ev}(\theta)) = 0$, in particular at a zero of the squared loss, $\nabla^2 \mathcal{L} = J^{\mathsf{T}} \nabla^2 \lambda \, J$. For the squared loss this is $J^{\mathsf{T}} J$, positive semidefinite of rank* $\operatorname{rk} J$*, with kernel* $\ker J$.

The non-zero eigenvalues of $J^{\mathsf{T}} J$ and of the tangent kernel $J J^{\mathsf{T}}$ coincide. So at an interpolating parameter the curvature of the loss in the directions normal to the solution set is the spectrum of the tangent kernel.

Notes

Reverse-mode differentiation has been rediscovered several times; Griewank's history is the standard reference. The compositional formulation is from [FST19] and [CGGWZ22], and the axiomatics of reverse derivatives from [CCG20]. The tangent kernel was introduced in [JGH18]. Functional dimension for ReLU networks is studied in [GLMW22].

Exercises

Exercise 2.1. Write out the lens of a layer $\ell(W, b; x) = \sigma(Wx + b)$, and compose two of them to obtain the gradient of $\frac{1}{2}(a^{\mathsf{T}} \sigma(Wx + b) + c - y)^2$ with respect to (W, b, a, c).

Exercise 2.2. Show that $\operatorname{rk} K(\theta) = \operatorname{rk} J(\theta)$ and that, for n data points and scalar output, $K(\theta)$ is invertible if and only if the functions $\partial F / \partial \theta_k(\theta, \cdot)$, $k = 1, \dots, P$, restricted to the data, span $\mathbb{R}^n$.

Exercise 2.3. Derive the formula for $\nabla^2(\lambda \circ \mathrm{ev})$ from the composition of 2-jets.

Exercise 2.4. For the linear family $F(\theta, x) = \langle \theta, \phi(x) \rangle$ with a fixed feature map $\phi \colon X \to \mathbb{R}^P$, compute J, K, the fibres of ev, and the functional dimension.

Chapter 3

Symmetries and the Moduli of Networks

The realization map $\theta \mapsto f_\theta$ is never injective. Its fibres are the first thing to understand about an architecture, since every function on parameter space that is defined in terms of the network's input–output behaviour, such as the loss, is constant on them. This chapter describes the fibres for networks with one hidden layer, where the theory is complete, and draws the consequences that hold in general.

3.1 Symmetries

Definition 3.1. A ***symmetry*** of a family $F\colon \Theta \times X \to Y$ is a diffeomorphism g of Θ with $F(g\theta, x) = F(\theta, x)$ for all θ, x: an invertible 2-cell of (Θ, F) in Para.

For a hidden unit j in layer i of a feedforward architecture, call its ***incoming*** parameters the j-th row of W_i and the j-th entry of b_i, and its ***outgoing*** parameters the j-th column of W_{i+1}.

Proposition 3.2. [classical] *Every feedforward architecture has the following symmetries.*

(i) Permutations of the hidden units of a layer, acting on incoming and outgoing parameters together.

(ii) If σ is odd: for each hidden unit, the sign change of its incoming and outgoing parameters.

(iii) If σ is positively homogeneous of degree one, as ReLU *is: for each hidden unit and $\lambda > 0$, multiplication of the incoming parameters by λ and of the outgoing ones by λ^{-1}.*

For $\sigma = \tanh$ these generate a finite group G of order $\prod_{i=1}^{L-1} 2^{n_i} n_i!$.

3.2 One hidden layer

Consider $F(\theta, x) = \sum_{j=1}^{H} a_j \sigma(\langle w_j, x\rangle + b_j) + c$ on $X = \mathbb{R}^d$, with $\theta = (w_j, b_j, a_j)_{j \le H}$ and c, so $P = H(d+2)+1$, and $\sigma = \tanh$.

Definition 3.3. The parameter θ is ***irreducible*** if every $a_j \neq 0$, every $w_j \neq 0$, and $(w_j, b_j) \neq \pm(w_k, b_k)$ for $j \neq k$. The set of irreducible parameters is Θ^{irr}.

Theorem 3.4 (Identifiability). [literature; Sussmann [Sus92]; Fukumizu [Fuk96]] *Let θ be irreducible. Then $f_{\theta'} = f_\theta$ for a parameter θ' with the same number of hidden units if and only if $\theta' = g\theta$ for some $g \in G$. Moreover θ is irreducible if and only if the functions $\partial F/\partial\theta_k(\theta, \cdot)$, $k = 1, \dots, P$, are linearly independent; so the functional dimension is P exactly on Θ^{irr}.*

For deep networks with tanh the corresponding statement holds for generic parameters, by a theorem of Fefferman [Fef94]; for ReLU networks see [PL20]. The complement of Θ^{irr} is a finite union of proper linear subspaces, so Θ^{irr} is open and dense, and G acts freely on it.

Proposition 3.5 (The moduli of irreducible networks). [observation] *The quotient Θ^{irr}/G is a manifold of dimension P on which the realization map is injective. Let $Z = \{(w, b, a) : w \neq 0,\ a > 0\}$ and $Z' = \{(w, b) : w \neq 0\}/\pm$. Then*

$$\Theta^{\mathrm{irr}}/G \;\cong\; \big\{\text{sets of } H \text{ points of } Z \text{ with pairwise distinct images in } Z'\big\} \times \mathbb{R}.$$

If $d \geq 2$ it is connected and its fundamental group maps onto the symmetric group S_H.

Proof. A free action of a finite group has a manifold quotient. Using the sign change to make each a_j positive identifies $\Theta^{\mathrm{irr}}/(\mathbb{Z}/2)^H$ with the ordered H-tuples in Z having distinct images in Z', times the line of the output bias; dividing by S_H forgets the order. For $d \geq 2$, Z is connected, and the tuples with a coincidence in Z' form a closed subset of codimension $d + 1 \geq 2$ in Z^H, so the space of ordered tuples is connected. A connected covering with group S_H gives a surjection from the fundamental group of the base. □

So there are loops of networks, along which the function computed changes continuously and returns to itself, that cannot be lifted to loops of parameters: at the end of the loop two hidden units have exchanged roles. An explicit one moves (w_1, b_1, a_1) to (w_2, b_2, a_2) and back along two different arcs. The space of functions an architecture computes has topology that its parameter space, a Euclidean space, does not.

Proposition 3.6 (Fibres at reducible parameters). [classical; cf. [Sus92], [Fuk96], [SGJ21]] *At a reducible parameter the fibre of the realization map has positive dimension:*

(i) if $a_j = 0$, the fibre contains the affine space of all values of (w_j, b_j), of dimension $d + 1$;

(ii) *if* $(w_j, b_j) = (w_k, b_k)$, *it contains the line* $a_j + a_k = \text{const}$;

(iii) *if* $w_j = 0$, *it contains the surface* $a_j \sigma(b_j) + c = \text{const}$ *in the coordinates* (a_j, b_j, c).

Each case is a way for a network with H hidden units to compute a function that needs fewer. The set of parameters in Θ_H computing a given function of an irreducible network with $H' < H$ units is a union of such families, permuted by G; it is the subject of a detailed combinatorial study in [SGJ21]. At these parameters the rank of J drops, by Theorem 3.4. The image $\mathcal{F}$ of the realization map is therefore a P-dimensional manifold Θ^{irr}/G together with a singular set, of lower dimension, onto which the reducible strata of Θ are collapsed.

3.3 Homogeneous activations

Proposition 3.7. [literature; cf. [GLMW22]] *Let* σ *be positively homogeneous of degree one. For each hidden unit* j *let* ξ_j *be the vector field on* Θ *equal to the incoming parameters of* j *in the incoming coordinates, minus the outgoing parameters in the outgoing coordinates, and zero elsewhere. Then* $D_\theta F(\theta, x)\, \xi_j(\theta) = 0$ *wherever* F *is differentiable. Consequently the functional dimension of the architecture is at most* $P - H$ *at every parameter.*

Proof. ξ_j is the generator of the one-parameter group of symmetries of Proposition 3.2(iii); differentiate $F(g_\lambda \theta, x) = F(\theta, x)$ at $\lambda = 1$. The ξ_j for different units have disjoint supports in the incoming coordinates, so they are independent wherever each unit has a non-zero incoming parameter, and the bound holds on the closure of that set by semicontinuity of rank. □

The script confirms both statements on small networks: for tanh the Jacobian has full rank P at a random parameter and loses rank at reducible ones; for ReLU its rank never exceeds $P - H$ and the fields ξ_j lie in its kernel.

For ReLU the symmetry group is not compact, and its quotient is not a manifold or even a Hausdorff space: the orbit of a unit with zero outgoing weight is not closed, since $(\lambda w, \lambda b, 0) \to (0, 0, 0)$ as $\lambda \to 0$. The correct object is the quotient stack, in which such points are remembered with their stabilizers. This is the same reason that vector bundles in [DTHC, Ch. 5] formed a stack and not a space, and the loss is a function on it.

3.4 Invariant functions

Proposition 3.8. [classical] *Let* $\mathcal{L} \colon \Theta \to \mathbb{R}$ *be invariant under a group* G *of symmetries. Then the critical set of* $\mathcal{L}$ *is a union of* G*-orbits. If* G *contains a one-parameter*

group with generator ξ, then $d\mathcal{L}(\xi) = 0$ identically and, at every critical point, ξ lies in the kernel of the Hessian. If moreover $\nabla\lambda$ vanishes at $\mathrm{ev}(\theta)$ for a loss $\mathcal{L} = \lambda \circ \mathrm{ev}$, then the whole fibre of ev *through θ is critical.*

Proof. Differentiating $\mathcal{L}(g_s\theta) = \mathcal{L}(\theta)$ in s gives $d\mathcal{L}_\theta(\xi) = 0$ for all θ; differentiating that identity in a direction v at a critical point gives $\nabla^2\mathcal{L}(\xi, v) + d\mathcal{L}(D_v\xi) = \nabla^2\mathcal{L}(\xi, v) = 0$. The last statement is $\nabla\mathcal{L} = J^{\mathsf{T}}\nabla\lambda$. □

Corollary 3.9. [observation] *For an architecture with a positively homogeneous activation and at least one hidden unit, no loss of the form $\lambda \circ \mathrm{ev}$ is a Morse function on any open set containing a critical point at which some hidden unit has a non-zero incoming parameter: the Hessian at every such critical point has nullity at least the number of such units.*

So Morse theory in the form of [DTHC, Ch. 8] never applies directly to the loss of a ReLU network. It can apply to the function induced on a slice transverse to the orbits, or in Morse–Bott form, with critical manifolds in place of critical points. For tanh networks there is no continuous symmetry and no such obstruction on Θ^{irr}; the degeneracies there come from the reducible strata and from overparametrization (Chapter 6).

Notes

Identifiability of networks up to the obvious symmetries was established for one hidden layer by Sussmann and by Albertini and Sontag, and for deep tanh networks by Fefferman. The relation between reducibility and singularity of the Fisher information matrix is due to Fukumizu; it is the starting point of Watanabe's singular learning theory [Wato9]. The use of permutation symmetry to compare trained networks is discussed in Chapter 6.

Exercises

Exercise 3.1. Verify the three symmetries of Proposition 3.2 for a network with one hidden layer.

Exercise 3.2. For the one-hidden-layer tanh network, show that if $a_j = 0$ then $\operatorname{rk} J(\theta) \le P - (d+1)$ for every set of inputs, and if $(w_j, b_j) = (w_k, b_k)$ then $\operatorname{rk} J(\theta) \le P - 1$.

Exercise 3.3. For $F(\theta, x) = a\ \mathrm{ReLU}(wx + b) + c$ with $d = 1$, write down ξ and verify $D_\theta F \cdot \xi = 0$ by hand. What is the functional dimension?

Exercise 3.4. Compute the order of G for the tanh architectures with widths $(2, 3, 1)$ and $(2, 4, 4, 1)$.

Part II

Level Sets

Chapter 4

Transversality and Decision Boundaries

A binary classifier is a function $f_\theta \colon X \to \mathbb{R}$ whose sign is the predicted class, and its ***decision boundary*** is $B_\theta = f_\theta^{-1}(0)$. This chapter answers question (Q1) of Chapter 1 for decision boundaries: for almost every parameter the boundary is a smooth hypersurface; its isotopy type is constant on the components of the complement of a closed null set of parameters; and every isotopy type of compact hypersurface occurs for wide enough networks.

4.1 The boundary is generically a hypersurface

Theorem 4.1 (Generic decision boundaries). [reformulation; cf. [DTHC, Thm. 4.1, Thm. 1.12]] *Let $F \colon \Theta \times X \to \mathbb{R}$ be a family of the form $F((\theta', c), x) = G(\theta', x) + c$.*

(a) For every θ', the set of c for which 0 fails to be a regular value of $f_{(\theta', c)}$ has measure zero.

(b) For almost every θ, the decision boundary B_θ is a smooth hypersurface in X, closed as a subset and possibly empty, with trivial normal bundle oriented by ∇f_θ; and $X \setminus B_\theta$ is the disjoint union of the open sets $\{f_\theta > 0\}$ and $\{f_\theta < 0\}$.

(c) If X is a compact manifold with boundary, then for almost every θ the boundary B_θ is moreover a neat submanifold: it meets ∂X transversally, in ∂B_θ.

Proof. (a) 0 is a regular value of $G(\theta', \cdot) + c$ exactly when $-c$ is a regular value of $G(\theta', \cdot)$, and the critical values of a smooth function have measure zero by [DTHC, Thm. 4.1]. (b) follows from (a) by Fubini's theorem and [DTHC, Thm. 1.12]. (c) Apply (a) also to the restriction to ∂X and use [DTHC, Thm. 1.17]. □

Every feedforward architecture with a bias in its last layer has this form, whatever its depth and whatever smooth activation it uses. Part (a) is stronger than the conclusion of Principle 1.5, since it holds for each value of the other parameters separately; it says that a non-smooth decision boundary can always be repaired by an arbitrarily small change of the output bias alone. The level sets $f_\theta^{-1}(s)$ for other thresholds s are covered by the same statement.

Example 4.2. Let $f(x) = x_1^2 - x_2^2 + c$. For $c \neq 0$ the boundary is a hyperbola, a 1-manifold with two components. For $c = 0$ it is a pair of crossing lines,

not a manifold. In the language of [DTHC, Ch. 4], the boundary at $c = 0$ is a derived fibre product that is not a manifold; its cobordism class is that of either neighbour.

4.2 Several classes

A classifier with k classes is a map $f_\theta \colon X \to \mathbb{R}^k$, and the predicted class is the index of the largest coordinate. For a non-empty $I \subset \{1, \dots, k\}$ let $A_I \subset \mathbb{R}^k$ be the set of points whose largest coordinate is attained exactly on I. Then A_I is an open subset of the linear subspace $\{y_i = y_j : i, j \in I\}$, a submanifold of codimension $|I| - 1$, and the A_I partition $\mathbb{R}^k$.

Proposition 4.3. [reformulation; cf. [DTHC, Thm. 4.5]] *Let $F \colon \Theta \times X \to \mathbb{R}^k$ be a family whose last layer is affine with a bias. For almost every θ, f_θ is transverse to every A_I. Then the decision region of class i is the open set $f_\theta^{-1}(A_{\{i\}})$; the set where exactly the classes in I tie is a submanifold of codimension $|I| - 1$; and no more than $d + 1$ classes tie anywhere.*

Proof. F is a submersion, so [DTHC, Thm. 4.5] applies to each of the finitely many A_I, and a finite intersection of sets of full measure has full measure. A submanifold of codimension greater than d is empty. □

In the plane, then, a generic classifier has decision regions bounded by smooth arcs meeting three at a time at isolated points, like the cells of a foam, and four regions never meet at a point.

4.3 Chambers

Let X be compact, and for simplicity without boundary. Define

$$\Sigma = \{(\theta, x) : F(\theta, x) = 0,\ d_x F(\theta, x) = 0\} \subset \Theta \times X, \qquad \Delta = \pi(\Sigma) \subset \Theta,$$

the set of parameters for which 0 is not a regular value: the ***discriminant*** of the family.

Theorem 4.4 (Chambers). [reformulation; cf. [DTHC, Lem. 4.4, Thm. 8.5, Thm. 10.3]] *Let F be as in Theorem 4.1 with X compact. Then Δ is closed and of measure zero. If γ is a path in $\Theta \setminus \Delta$, there is a diffeotopy φ_t of X with $\varphi_t(B_{\gamma(0)}) = B_{\gamma(t)}$. In particular the diffeomorphism type of the pair (X, B_θ) is constant on each connected component of $\Theta \setminus \Delta$.*

Proof. Σ is closed and $\pi|_\Sigma$ is proper because X is compact, so Δ is closed; its complement has full measure by Theorem 4.1. Over $\Theta \setminus \Delta$ the set $W = F^{-1}(0)$

is a submanifold of $\Theta \times X$, since $d_x F \neq 0$ there, and by [DTHC, Lem. 4.4] the projection $W \to \Theta \setminus \Delta$ is a submersion; it is proper. By Ehresmann's theorem [DTHC, Thm. 8.5] it is a fibre bundle. Pulled back along γ it is trivial, which gives an isotopy of embeddings $B_{\gamma(0)} \to X$ with image $B_{\gamma(t)}$ at time t, and the isotopy extends to a diffeotopy of X by [DTHC, Thm. 10.3]. □

The components of $\Theta \setminus \Delta$ are the ***chambers*** of the architecture. The theorem says that a classifier has a well-defined topology (the number of components of its decision boundary and of its decision regions, their genera, how they are nested) which does not change under small changes of the parameters, and can change only when the parameter crosses Δ. What happens at a crossing is the subject of the next chapter. The symmetry group of Chapter 3 permutes the chambers.

For manifolds with boundary the same holds with Δ enlarged by the parameters for which B_θ is tangent to ∂X.

Proposition 4.5 (Walls). [observation] *Consider the one-hidden-layer architecture $F(\theta, x) = \sum_{j \leq H} a_j \sigma(\langle w_j, x \rangle + b_j) + c$ on $\mathbb{R}^d$ with $\sigma' > 0$. At every point where the matrix W with rows w_j has rank d, the map $\Phi(\theta, x) = (F, d_x F) \in \mathbb{R} \times \mathbb{R}^d$ is a submersion. Hence, over the open set of such parameters, Σ is a submanifold of dimension $P - 1$, and Δ is the image of a manifold of dimension $P - 1$.*

Proof. $d_x F = W^\mathsf{T}(a \odot \sigma'(z))$ with $z = Wx + b$. A variation δa changes $d_x F$ by $W^\mathsf{T}(\delta a \odot \sigma'(z))$, and $\delta a \mapsto \delta a \odot \sigma'(z)$ is a linear isomorphism since $\sigma' > 0$; so if W^T is onto, every variation of $d_x F$ is produced. The resulting change in F is compensated by δc. □

So the discriminant is no larger than a hypersurface, and the picture of parameter space is that of open chambers separated by walls of codimension one. The proposition is stated for the simplest architecture; the argument needs only that the parameters can move the value and the input-gradient of the network independently at each point, which holds for deeper networks at parameters where the layer Jacobians have full rank.

4.4 Every hypersurface is a decision boundary

Theorem 4.6 (Realization). [reformulation; cf. [DTHC, Thm. 6.2]; Hornik [Hor91]] *Let $K \subset \mathbb{R}^d$ be a compact domain with smooth boundary and $B \subset \operatorname{int} K$ a closed hypersurface. Let σ be smooth, bounded and non-constant. Then for all sufficiently large H there is a non-empty open set of parameters θ of the one-hidden-layer architecture with H units for which $B_\theta \cap K$ is carried to B by a diffeomorphism of K isotopic to the identity.*

Proof. A closed hypersurface in $\mathbb{R}^d$ separates and so has a trivial normal bundle; using a tubular neighbourhood [DTHC, Thm. 6.2] one constructs a smooth $g\colon K \to \mathbb{R}$ with $g^{-1}(0) = B$, 0 a regular value, and $g \neq 0$ on ∂K. Since B is compact, there is $\varepsilon > 0$ such that every h with $\|h - g\|_{C^1(K)} < \varepsilon$ has 0 as a regular value and does not vanish on ∂K. By Hornik's theorem, networks with one hidden layer are dense in $C^1(K)$, so some f_{θ_0} is within ε of g, and so is f_θ for θ near θ_0, the realization map being continuous into $C^1(K)$. The family $(1-s)g + sf_\theta$, $s \in [0,1]$, stays within ε of g, and the argument of Theorem 4.4 applied to it gives the diffeotopy. □

So there is no topological restriction on decision boundaries beyond those satisfied by every compact hypersurface in $\mathbb{R}^d$: it is orientable, it separates, and it bounds (Proposition 5.6). The theorem is silent on how large H must be. Upper bounds for the Betti numbers of decision regions in terms of the architecture, which give lower bounds for H, are due to Bianchini and Scarselli [BS14].

4.5 Piecewise-linear networks

For $\sigma = \mathrm{ReLU}$ a network is piecewise linear, its decision boundary is a polyhedral complex, and the smooth theory does not apply as it stands. The counterpart of Theorem 4.1 is due to Grigsby and Lindsey [GL22]. They define what it means for a ReLU network to be generic and transversal, in terms of the bent hyperplane arrangement formed by the zero sets of its units, and prove that almost all ReLU networks have both properties. As an application they show that for the architecture with one hidden layer of $d+1$ units, a decision region has at most one bounded connected component: a restriction on topology coming from the architecture, of the kind Theorem 4.6 rules out for unbounded width.

Notes

That decision boundaries are generically manifolds is folklore, and is usually justified by Sard's theorem in the form (a). The statement about chambers is the standard structure theory of discriminants in singularity theory, applied to this family. For topological complexity of decision regions see [BS14], and for an empirical study of how networks change the topology of data, [NZL20].

Exercises

Exercise 4.1. Let $d = 1$ and $X = [-R, R]$. Show that for almost every θ the decision boundary is a finite set at which f_θ changes sign, and that the parity of its cardinality is determined by the signs of $f_\theta(\pm R)$.

Exercise 4.2. Let X be a closed disc in the plane. Describe the generic decision boundary up to diffeomorphism.

Exercise 4.3. For the family $F(c, x) = x_1^2 - x_2^2 + c$ on the plane, determine Σ and Δ, and describe the chambers.

Exercise 4.4. For a three-class classifier in the plane satisfying the conclusion of Proposition 4.3, show that near a point where all three classes tie the decision regions look like three sectors meeting at a point.

Chapter 5

Topology Change along Paths: Surgery and Cobordism

By Theorem 4.4 the topology of a decision boundary can change during training only when the parameter crosses the discriminant. This chapter answers question (Q2): for a generic path the crossings are isolated, and at each of them the boundary changes by one surgery, determined by the index of a critical point. The union of the boundaries along the path is a cobordism, and the surgeries are the critical points of its height function. This is the Morse theory of [DTHC, Ch. 8] applied to a manifold that the path itself creates.

5.1 The trace of a path

Let $F\colon \Theta \times X \to \mathbb{R}$ be a family as in Theorem 4.1, with X compact of dimension d, and let $\gamma\colon [0,1] \to \Theta$ be a path. Write $f_t = f_{\gamma(t)}$ and $B_t = f_t^{-1}(0)$, and define

$$\Phi_\gamma\colon [0,1] \times X \to \mathbb{R}, \quad \Phi_\gamma(t,x) = F(\gamma(t),x); \qquad W_\gamma = \Phi_\gamma^{-1}(0) \subset [0,1] \times X.$$

Theorem 5.1 (The trace is a cobordism). [reformulation; cf. [DTHC, Thm. 1.17, Thm. 4.5]] *If 0 is a regular value of Φ_γ and of f_0 and f_1, then W_γ is a compact neat hypersurface in $[0,1] \times X$ with $\partial W_\gamma = B_0 \sqcup B_1$: a cobordism from B_0 to B_1 embedded in $[0,1] \times X$. For a path with endpoints outside Δ, this holds after an arbitrarily small perturbation of the bias component of γ fixed at the endpoints.*

Proof. The first statement is [DTHC, Thm. 1.17]. For the second, replace the bias component $c(t)$ by $c(t) + s\,\beta(t)$ with $\beta > 0$ on $(0,1)$ and vanishing with all derivatives at the ends. The map $(s,t,x) \mapsto \Phi_\gamma(t,x) + s\beta(t)$ is a submersion on $\{0 < t < 1\}$, so by [DTHC, Thm. 4.5] zero is a regular value for almost every s; near $t = 0, 1$ it is one already because f_0, f_1 have 0 as a regular value. □

The times at which γ meets Δ are, by [DTHC, Lem. 4.4], exactly the critical values of the height function $\tau\colon W_\gamma \to [0,1]$, $\tau(t,x) = t$. At such a point (t_0, x_0) we have $d_x f_{t_0}(x_0) = 0$: x_0 is a critical point of f_{t_0} with critical value 0.

5.2 Surgery

Theorem 5.2 (One surgery at a time). [reformulation; cf. [DTHC, Thm. 8.6, Cor. 8.7]] *Let γ be a path for which W_γ is a manifold and τ is a Morse function with distinct critical values $t_1 < \cdots < t_r$ in $(0,1)$. Then:*

(a) On each interval of $[0,1] \setminus \{t_1, \dots, t_r\}$ the boundary B_t moves by an ambient isotopy of X.

(b) At each t_i there is exactly one point $x_i \in B_{t_i}$ at which B_{t_i} fails to be a manifold. It is a non-degenerate critical point of f_{t_i} with $\partial_t \Phi_\gamma(t_i, x_i) \neq 0$, and near (t_i, x_i) there are coordinates on X in which

$$\Phi_\gamma(t, x) = \pm(t - t_i) - x_1^2 - \cdots - x_k^2 + x_{k+1}^2 + \cdots + x_d^2,$$

with k the index of x_i as a critical point of f_{t_i}.

(c) $B_{t_i+\varepsilon}$ is obtained from $B_{t_i-\varepsilon}$ by a surgery: a copy of $S^{\lambda-1} \times D^{d-\lambda}$ is removed and $D^\lambda \times S^{d-\lambda-1}$ is glued in, where λ is the index of τ at (t_i, x_i); $\lambda = k$ or $d - k$ according to the sign in (b). Correspondingly the region $\{f_t \leq 0\}$ gains or loses a handle of index k.

Proof. (a) is Theorem 4.4. For (b): at a critical point of τ on W_γ the tangent space of W_γ is $\{dt = 0\}$, so $d_x \Phi_\gamma = 0$ and, W_γ being a manifold, $\partial_t \Phi_\gamma \neq 0$. Solving $\Phi_\gamma = 0$ for t as a function of x near the point exhibits W_γ as a graph over X on which τ is the function $t(x)$, whose Hessian at x_i is $-\nabla^2 f_{t_i}(x_i)/\partial_t \Phi_\gamma$. So τ is non-degenerate there exactly when x_i is a non-degenerate critical point of f_{t_i}, and the normal form follows from the Morse lemma with parameters. (c) is [DTHC, Thm. 8.6] for τ on W_γ: the level sets of τ are the B_t, and passing a critical point of index λ changes the level set by the stated surgery. □

In words: the topology of a decision boundary changes exactly when a critical point of the network function crosses the threshold, and the change is read off from the index of the critical point and the direction of crossing. Table 5.1 lists the cases for inputs of dimension two and three.

Proposition 5.3. [observation] *In the situation of Theorem 5.2(c), $\chi(B_{t_i+\varepsilon}) - \chi(B_{t_i-\varepsilon})$ equals $2(-1)^\lambda$ if d is odd and 0 if d is even, and the Euler characteristic of the region $\{f_t \leq 0\}$ changes by $\pm(-1)^k$.*

Proof. The change is $\chi(D^\lambda \times S^{d-\lambda-1}) - \chi(S^{\lambda-1} \times D^{d-\lambda}) = \chi(S^{d-\lambda-1}) - \chi(S^{\lambda-1}) = (-1)^{d-\lambda-1} - (-1)^{\lambda-1}$. The second statement is additivity of χ under attaching a k-cell. □

Example 5.4. Let $F(t, x) = (x_1^2 - 1)^2 + x_2^2 - t$ on the plane. The function $(x_1^2 - 1)^2 + x_2^2$ has minima at $(\pm 1, 0)$ with value 0 and a saddle at the origin with value 1. For $t < 0$ the region $\{F \leq 0\}$ is empty. At $t = 0$ two discs are

d	index k of x_i	effect on the region $\{f_t \le 0\}$ when the critical value becomes negative
2	0 (minimum)	a new disc appears; the boundary gains a circle
2	1 (saddle)	two parts of the region join, or one part meets itself; two boundary arcs reconnect, so the number of boundary circles changes by ± 1
2	2 (maximum)	a hole in the region is filled; a boundary circle disappears
3	0	a new ball appears; the boundary gains a sphere
3	1	a tube joins two parts of the region, or adds a handle to one
3	2	a tunnel through the region is closed by a disc, or a cavity is split in two
3	3	a cavity is filled; a boundary sphere disappears

Table 5.1: Elementary changes of topology of a decision region, by the index of the critical point crossing zero from above. Crossing in the other direction reverses each.

born (two critical points cross at the same time, which is not generic; a small asymmetry separates them). At $t = 1$ the saddle crosses and the two discs merge. The numbers of components are 0, 2, 1, and the Euler characteristics of the region 0, 2, 1. The script counts components on a grid.

5.3 Genericity of paths

Are the hypotheses of Theorem 5.2 generic? For paths in the space of all functions, yes: this is the content of [DTHC, Thm. 8.12]. For paths in a given architecture one needs enough parameters to move critical points and critical values independently.

Proposition 5.5. [reformulation; cf. [DTHC, Thm. 4.6, Thm. 8.12]] *For the architecture of Proposition 4.5, with $H \ge d$, let γ be a path with endpoints in $\Theta \setminus \Delta$ along which W has rank d. Then arbitrarily close to γ, with the same endpoints, there is a path satisfying the hypotheses of Theorem 5.2.*

Sketch. By Proposition 4.5 the map $(\theta, x) \mapsto (F, d_x F)$ is a submersion along the path, so the family of perturbations of γ by compactly supported variations of (a, c) induces a family of maps $(t, x) \mapsto (\Phi, d_x \Phi)$ transverse to the zero section; by [DTHC, Thm. 4.5] a generic member has Σ pulled back to a finite set of points at which the crossing is transverse. Non-degeneracy of the Hessian at those points, and distinctness of the times, are further conditions of positive

codimension on the 2-jet, which a variation of W and b moves freely; [DTHC, Thm. 4.6] applies. □

A training path produced by gradient descent is one particular path, and nothing guarantees that it is generic in this sense. The proposition says that if it is not, an arbitrarily small perturbation is, and that the perturbed path sees a finite sequence of elementary surgeries. Paths with symmetry are the typical exception (Example 5.4): a symmetric network crossing a wall has several critical points crossing together.

5.4 What cobordism does and does not see

Proposition 5.6. [classical] *Let $d \geq 2$ and let $B = f^{-1}(0) \subset \mathbb{R}^d$ be compact with 0 a regular value. Then B is the boundary of a compact region of $\mathbb{R}^d$; it is orientable, its normal bundle is trivial, and $TB \oplus \mathbb{R}$ is trivial. In particular B is null-cobordant, all its Stiefel–Whitney numbers vanish, and no closed non-orientable surface is a decision boundary in $\mathbb{R}^3$.*

Proof. Choose a closed ball D containing B in its interior. Its boundary sphere is connected and f does not vanish on it, so f has a constant sign there, say positive. Then $\{f \leq 0\} \cap D$ is a compact manifold with boundary B. The normal bundle is trivialized by ∇f, and $TB \oplus \nu = T\mathbb{R}^d|_B$ is trivial. □

So the cobordism class of a decision boundary carries no information: it is always zero, as [DTHC, Thm. 4.16] predicts for a fibre of a map to a contractible space. What carries information is finer: the diffeomorphism type of the pair (X, B_t) in each chamber, and the sequence of surgeries along a path.

That sequence is not an invariant of the endpoints, but its ambiguity is controlled. Parameter space is contractible, so any two paths with the same endpoints are homotopic, and a generic homotopy between generic paths meets a short list of codimension-two events.

Proposition 5.7 (Moves). [reformulation; cf. Cerf; [DTHC, Thm. 8.12]] *Let γ_0, γ_1 be generic paths with the same endpoints in an architecture rich enough for jet transversality in two parameters. Their surgery sequences differ by a finite sequence of moves of three kinds: exchanging the order of two consecutive surgeries at different critical points; inserting or deleting a pair of consecutive surgeries of indices $\lambda, \lambda + 1$ that cancel (a birth–death point of f_t on the zero level); and inserting or deleting a surgery followed by its inverse (a critical value that touches zero and returns).*

The proposition is the two-parameter analogue of Theorem 5.2, and is the form Cerf's theorem takes for level sets. It says that the surgery sequence of a training run is well defined modulo these moves, and that anything invariant

under them, for instance the parity of the number of surgeries of each index modulo the cancellation of adjacent pairs, depends only on the endpoints.

Notes

The description of how level sets change in one-parameter families is classical Morse and Cerf theory; see Milnor's lectures on the h-cobordism theorem. Its use for decision boundaries during training appears in empirical form in studies that track Betti numbers of decision regions and of data representations, such as [NZL20]. Persistent homology of the family $\{f_t \leq s\}$ in the two parameters (t, s) is the natural computational tool, and the events of Proposition 5.7 are the generic singularities of its bifiltration.

Exercises

Exercise 5.1. Verify Example 5.4: find the critical points and indices of $(x_1^2 - 1)^2 + x_2^2$ and the topology of the sublevel sets.

Exercise 5.2. For $d = 1$, show that along a generic path the points of the decision boundary are born and die in pairs, and that the parity of their number is constant if X is a circle.

Exercise 5.3. In $d = 3$, a decision region evolves from empty to a solid torus. What is the smallest number of surgeries, and what are their indices?

Exercise 5.4. Prove the formula for the Hessian of τ used in the proof of Theorem 5.2(b).

Chapter 6

The Loss Landscape

Question (Q3) asks for the shape of a set of parameters. The most important such set is the zero set of the loss, the parameters that fit the data exactly, and the regular value theorem describes it. The rest of the landscape is organized by the sublevel sets of the loss, to which the Morse theory of [DTHC, Ch. 8] applies with two caveats established in Chapter 3 and below: the loss is rarely a Morse function, and parameter space is not compact.

6.1 The interpolation manifold

Let $\mathrm{ev}\colon \Theta \to \mathbb{R}^N$ be the evaluation map on n data points, $N = nm$, and $\mathcal{L} = \frac{1}{2}|\mathrm{ev} - y|^2$. Write $M_y = \mathrm{ev}^{-1}(y) = \mathcal{L}^{-1}(0)$.

Theorem 6.1 (Interpolating parameters). [reformulation; cf. [DTHC, Thm. 1.12, Thm. 4.1]; Cooper [Coo21]]

(a) For almost every $y \in \mathbb{R}^N$, M_y is a submanifold of Θ of dimension $P - N$, possibly empty. If $P < N$ it is empty for almost every y.

(b) If y is a regular value and $\theta \in M_y$, then $T_\theta M_y = \ker J(\theta)$ and $\nabla^2\mathcal{L}(\theta) = J^{\mathsf{T}}J$ has rank N: it vanishes on $T_\theta M_y$ and is positive definite on its orthogonal complement.

(c) M_y is non-empty for y in the image of ev*, and that image has non-empty interior if and only if J has rank N somewhere.*

Proof. (a) By [DTHC, Thm. 4.1] almost every y is a regular value of ev, and then [DTHC, Thm. 1.12] applies; if $P < N$ every point of the image is a critical value. (b) $T_\theta M_y = \ker D\mathrm{ev}_\theta$ by [DTHC, Thm. 1.12], and the Hessian is given by Proposition 2.6. (c) If J has rank N at θ then ev is open near θ; conversely if the rank is everywhere less than N the image has measure zero. □

Cooper proved the statement in this form for feedforward networks with smooth activations, and showed in addition that for sufficiently wide networks M_y is non-empty for generic data. In the overparametrized regime $P > N$, then, the global minima of the loss are not isolated: they form a manifold whose dimension is the excess of parameters over constraints, and the loss is non-degenerate in the directions normal to it. In the terminology of Morse

theory the loss is a Morse–Bott function near M_y. By Proposition 2.6 its normal curvature is the spectrum of the tangent kernel.

Proposition 6.2 (Convergence near the interpolation manifold). [literature; Liu–Zhu–Belkin [LZB22]] *Let $U \subset \Theta$ be an open set on which the smallest eigenvalue of the tangent kernel $K = JJ^{\mathsf{T}}$ is at least $\mu > 0$. Then $|\nabla \mathcal{L}|^2 \geq 2\mu\, \mathcal{L}$ on U. A gradient flow line starting at θ_0 satisfies $\mathcal{L}(\gamma(t)) \leq \mathcal{L}(\theta_0) e^{-2\mu t}$ and has length at most $\sqrt{2\mathcal{L}(\theta_0)/\mu}$ for as long as it stays in U; if the ball of that radius about θ_0 lies in U, the flow line converges to a point of M_y.*

Proof. $|\nabla \mathcal{L}|^2 = (\mathrm{ev} - y)^{\mathsf{T}} JJ^{\mathsf{T}} (\mathrm{ev} - y) \geq \mu |\mathrm{ev} - y|^2 = 2\mu\mathcal{L}$. Then $\frac{d}{dt}\mathcal{L} = -|\nabla\mathcal{L}|^2 \leq -2\mu\mathcal{L}$, and $\frac{d}{dt}\sqrt{\mathcal{L}} = -|\nabla\mathcal{L}|^2/(2\sqrt{\mathcal{L}}) \leq -\sqrt{\mu/2}\,|\nabla\mathcal{L}|$, which integrates to the bound on the length $\int |\dot\gamma| = \int |\nabla\mathcal{L}|$. □

The inequality is the quantitative form of transversality: $\mu > 0$ says that ev is a submersion on U with a uniform bound. It is the only quantitative statement in this book, and it is quoted to show where the qualitative theory hands over to analysis.

6.2 Sublevel sets and barriers

For $c \in \mathbb{R}$ let $\Theta^c = \{\mathcal{L} \leq c\}$. Two parameters are ***connected at level*** c if they lie in the same component of Θ^c, and the ***barrier*** between them is the infimum of such c. This is the notion studied empirically under the name of mode connectivity [GIP18, DVSH18].

Proposition 6.3 (Barriers). [reformulation; cf. [DTHC, Cor. 8.7, Prop. 8.10]] *Let $\mathcal{L}\colon \Theta \to \mathbb{R}$ be a Morse function with compact sublevel sets. Then the components of Θ^c change only at critical values: a component is born at each local minimum, and two components merge at a critical point of index 1 whose two descending directions lead to different components. The barrier between two local minima is the level at which their components merge; it is attained at a critical point of index 1, and equals the minimum over paths joining them of the maximum of $\mathcal{L}$ along the path.*

So barriers are the death times in the degree-zero barcode of the sublevel functor [DTHC, Prop. 8.10], and the barcode is the complete record of connectivity at all levels. The hypotheses hold, for instance, for a generic analytic loss with weight decay, $\mathcal{L} + \frac{\varepsilon}{2}|\theta|^2$, on an architecture without continuous symmetries. Without weight decay the sublevel sets are not compact and the statement can fail (Proposition 6.5).

Example 6.4 (Two parameters). Let $F(a, b; x) = abx$ with one data point $(x, y) = (1, 1)$, so $\mathcal{L} = \frac{1}{2}(ab - 1)^2$. The interpolation manifold is the hyperbola

$ab = 1$, of dimension $P - N = 1$, with two components exchanged by the symmetry $(a, b) \mapsto (-a, -b)$. The only other critical point is the origin, a saddle with $\mathcal{L} = \frac{1}{2}$ and Hessian eigenvalues ± 1. For $c < \frac{1}{2}$ the sublevel set has two components; at $c = \frac{1}{2}$ they merge. So the barrier between $(1, 1)$ and $(-1, -1)$ is $\frac{1}{2}$, attained on the straight segment between them, and no path does better. After applying the symmetry to one of the two points they coincide. The script checks the component counts on a grid.

The example is the smallest instance of a general phenomenon. By Proposition 3.8 minima come in orbits of the symmetry group, and two minima in the same orbit compute the same function but may be separated by a barrier. The barrier between two trained networks should therefore be measured after choosing representatives of their orbits, which is the content of the conjecture of Entezari et al. that barriers largely disappear once permutations are accounted for [EKSN22]. In the language of Chapter 3, the relevant sublevel sets are those of the function induced on the quotient Θ/G.

Two results from the literature describe the overparametrized case. For a class of deep networks with piecewise-linear activations in which one hidden layer is wider than the number of training points, Nguyen proved that every sublevel set of the loss is connected [Ngu19]. And Şimşek et al. describe the critical manifolds that the reducible strata of Proposition 3.6 contribute, and show how adding hidden units connects the $|G|$ copies of a minimum [SGJ21].

6.3 Why the loss is rarely a Morse function

Three mechanisms make the loss degenerate, and they are independent.

(i) *Overparametrization.* If $P > N$ the zero-loss minima form the manifold M_y of positive dimension (Theorem 6.1).

(ii) *Continuous symmetry.* For homogeneous activations every critical point has a Hessian of nullity at least H (Corollary 3.9).

(iii) *Reducible strata.* A critical point of a smaller network gives a positive-dimensional family of critical points of a larger one, along the fibres of Proposition 3.6; these are typically saddles [FA00].

In each case the degeneracy is of Morse–Bott type on an open dense part of the critical set, and the handle theorem [DTHC, Thm. 8.6] has a Morse–Bott version in which a critical manifold contributes a bundle of cells over itself. But the density of Morse functions [DTHC, Cor. 4.8] is of no direct use: a generic perturbation of the loss as a function on Θ is not the loss of any data set, and perturbing the data preserves all three mechanisms. This is the point

of Principle 1.5 again. Genericity must be established within the family one actually has.

6.4 Critical points at infinity

Proposition 6.5. [observation] *For architectures with a bounded activation the loss is not proper, and its infimum need not be attained. For $F(w;x)=\tanh(wx)$ with the single data point $(1,1)$, $\mathcal{L}(w)=\frac{1}{2}(\tanh w-1)^2$ has no critical point, its infimum 0 is not attained, and every sublevel set Θ^c with $0<c<2$ is a half-line.*

Proof. $\mathcal{L}'(w)=(\tanh w-1)\operatorname{sech}^2 w<0$ for all w. □

So the topology of sublevel sets can change with no critical point to account for it, the change being carried off to infinity, and the regular interval theorem [DTHC, Thm. 8.4], which needs compactness, does not apply. Any use of Morse theory for losses must either restore compactness (weight decay, a bounded parameter domain) or account for these ends explicitly, as one does for the Palais–Smale condition in infinite dimensions.

Notes

The manifold of global minima was described by Cooper [Coo21]; the Polyak–Łojasiewicz condition in this setting is due to Liu, Zhu and Belkin. Mode connectivity was observed by Garipov et al. and Draxler et al., and analysed for wide networks by Freeman and Bruna and by Nguyen; permutation alignment is studied in [EKSN22]. For the singular structure of the landscape at reducible parameters see [FA00, SGJ21] and Watanabe's theory [Wat09].

Exercises

Exercise 6.1. For the linear family of Exercise 2.4 with $P>n$ and features in general position, describe M_y and the Hessian of the loss.

Exercise 6.2. Verify Example 6.4: find all critical points of $\frac{1}{2}(ab-1)^2$, their Hessians, and the components of the sublevel sets.

Exercise 6.3. In the setting of Proposition 6.2, show that the limit point of the flow line lies within $\sqrt{2\mathcal{L}(\theta_0)/\mu}$ of θ_0.

Exercise 6.4. For $F(a_1,\dots,a_k;x)=a_1\cdots a_k x$ with data $(1,1)$, show that M_y has 2^{k-1} components and dimension $k-1$, and that the origin is a degenerate critical point for $k\geq 3$.

Part III

Maps

Chapter 7

Embeddings, Isotopy, and Depth

The previous three chapters studied level sets. This one studies the maps themselves, and in particular what a network does to a data set regarded as a subset of input space. The standing assumption, usually called the manifold hypothesis [FMN16], is that the data lie on or near a compact submanifold $K \subset \mathbb{R}^D$ of dimension m much smaller than D, with the classes occupying disjoint closed subsets K_+, K_-. A hidden layer of width k is a map $h_\theta \colon \mathbb{R}^D \to \mathbb{R}^k$, and the question is what $h_\theta(K)$ can be.

7.1 The first layer generically embeds

Proposition 7.1. [reformulation; cf. Whitney; [DTHC, Thm. 4.5, Cor. 4.8, Prop. 3.12]] *Let $K \subset \mathbb{R}^D$ be a compact submanifold of dimension m, let $k \geq 2m+1$, and let σ be a diffeomorphism of $\mathbb{R}$ onto an open interval. Then for almost every $W \in \mathbb{R}^{k \times D}$ and every b, the map $x \mapsto \sigma(Wx + b)$ restricted to K is an embedding. The set of such (W, b) is open.*

Proof. Applying σ coordinatewise after a translation is a diffeomorphism onto its image, so the question is whether $W|_K$ is an embedding. The map $(W, (x, y)) \mapsto W(x - y)$ on $\mathbb{R}^{k \times D} \times (K \times K \setminus \Delta)$ is a submersion to $\mathbb{R}^k$, so by [DTHC, Thm. 4.5] for almost every W its restriction to $K \times K \setminus \Delta$, a manifold of dimension $2m < k$, is transverse to 0 and hence misses it: $W|_K$ is injective. Likewise $(W, (x, v)) \mapsto Wv$ on the unit tangent bundle of K, of dimension $2m - 1$, is a submersion, so for almost every W no unit tangent vector is killed: $W|_K$ is an immersion. An injective immersion of a compact manifold is an embedding. Openness is [DTHC, Prop. 3.12]. □

So a hidden layer wider than twice the intrinsic dimension of the data loses no topological information for almost every choice of weights, however small it is compared with the ambient dimension D. This is Whitney's embedding theorem read as a statement about random projections. Narrower layers may fail to be injective on K, and that is where the topology of a representation can change.

7.2 What injective networks cannot do

A classifier ends with a linear map, so the last hidden representation must make the classes linearly separable. If the network up to that point is injective, its restriction to K is a homeomorphism onto the image, and separability becomes a constraint on how K_+ and K_- sit in the ambient space.

Theorem 7.2 (Nested classes). [classical; invariance of domain; cf. [Joh19]] *Let $h\colon \mathbb{R}^d \to \mathbb{R}^d$ be continuous and injective, let $\beta \subset \mathbb{R}^d$ be a closed topological ball with boundary sphere S, and let p be an interior point of β. Then $h(p)$ lies in the convex hull of $h(S)$. In particular no hyperplane separates $h(p)$ from $h(S)$.*

Proof. Suppose a linear functional λ satisfies $\lambda > 0$ on $h(S)$ and $\lambda(h(p)) \leq 0$. Let q be a point of the compact set $h(\beta)$ at which λ is smallest. Then $\lambda(q) \leq 0$, so $q \notin h(S)$ and $q \in h(\operatorname{int}\beta)$, which is open in $\mathbb{R}^d$ by invariance of domain. But a linear functional has no local minimum on an open set. □

Corollary 7.3. [classical; [Joh19, HS17]] *A feedforward network on $\mathbb{R}^d$ in which every hidden layer has width d, every weight matrix is invertible, and the activation is injective, cannot separate a point from a sphere surrounding it. To do so a network must either have a layer of width at least $d+1$ or use a layer that is not injective.*

Both escapes are used in practice. A wider layer provides the extra dimension into which the inner class can be lifted. A ReLU layer is not injective: it folds space along hyperplanes, and a network of ReLU layers can map a nested data set to a separable one by collapsing. Naitzat, Zhitnikov and Lim [NZL20] measured the Betti numbers of the two classes layer by layer and found that well-trained networks reduce them steadily, faster with ReLU than with smooth invertible activations. In the terms of this chapter: smooth invertible layers of constant width act by diffeomorphisms and cannot change topology, and the network must then rely on the few layers where injectivity fails.

7.3 Neural ordinary differential equations

A ***neural ODE*** [CRBD18] is the time-T map φ_θ of the equation $\dot{x} = v_\theta(x,t)$, where v_θ is a network with input (x,t) and output in $\mathbb{R}^d$.

Proposition 7.4. [reformulation; cf. [DTHC, Thm. 10.2]] *If v_θ is smooth with bounded velocity, as it is when its last hidden layer has a bounded activation, then $t \mapsto \varphi_\theta^t$ is a diffeotopy of $\mathbb{R}^d$. Hence φ_θ is a diffeomorphism isotopic to the identity, it preserves orientation, and for every compact $K \subset \mathbb{R}^d$ the image $\varphi_\theta(K)$ is ambient isotopic to K.*

Corollary 7.5 (Obstructions). [classical; cf. Dupont–Doucet–Teh [DDT19]] *A neural ODE on $\mathbb{R}^d$ cannot:*

(a) for $d = 1$, reverse the order of two points;

(b) map a sphere and a point inside it to linearly separable sets (Theorem 7.2);

(c) for $d = 3$, map two linked circles to unlinked ones, or a knotted circle to a round one.

Proof. (a) A diffeomorphism of $\mathbb{R}$ isotopic to the identity is increasing. (c) The linking number is the degree of the Gauss map [DTHC, Ex. 7.12] and is invariant under isotopy by [DTHC, Thm. 7.2]; the knot type is an isotopy invariant by definition. □

These are statements about exact representation. A neural ODE can still approximate a map that violates them, at the cost of a vector field with large derivatives, and [DDT19] observed that training then becomes slow and unstable. Their remedy was to append extra coordinates to the state.

7.4 Augmentation

Theorem 7.6 (Flows in the stable range). [reformulation; cf. [DTHC, Thm. 10.9, Thm. 10.3, Thm. 10.2]] *Let K be a compact manifold of dimension m and let $e_0, e_1 \colon K \to \mathbb{R}^N$ be embeddings, with $N \geq 2m + 2$. Then there is a smooth, compactly supported, time-dependent vector field on $\mathbb{R}^N$ whose time-one map φ satisfies $\varphi \circ e_0 = e_1$.*

Proof. By [DTHC, Thm. 10.9] the two embeddings are isotopic. By the isotopy extension theorem [DTHC, Thm. 10.3] the isotopy is induced by a diffeotopy of $\mathbb{R}^N$ with compact support, and by [DTHC, Thm. 10.2] the diffeotopy is the flow of its velocity field. □

Corollary 7.7 (Separation after augmentation). [observation] *Let $K = K_+ \sqcup K_- \subset \mathbb{R}^d$ be a compact submanifold of dimension m, the union of two disjoint compact submanifolds, and let $N \geq \max(d + 1, 2m + 2)$. Then there is a compactly supported flow on $\mathbb{R}^N$ whose time-one map carries $K_+ \times \{0\}$ and $K_- \times \{0\}$ into the open half-spaces $\{x_N > 0\}$ and $\{x_N < 0\}$. The vector field may be taken to be a neural network with one hidden layer and input (x, t).*

Proof. Take e_0 the inclusion of K in $\mathbb{R}^d \times \{0\} \subset \mathbb{R}^N$ and $e_1 = e_0 \pm (0, \dots, 0, 1)$ on $K_\pm$, and apply Theorem 7.6 to get a field v with flow φ. For the last statement, approximate v in C^1 on a compact set containing the support, by Hornik's theorem [Hor91]; by Grönwall's inequality the time-one maps are then close in C^1 on compact sets, and the conclusion, being an open condition on the images of the compact sets $K_\pm$, persists. □

So the obstructions of Corollary 7.5 disappear once the state space has dimension at least $2m + 2$, where m is the dimension of the data and not of the ambient space. For finite data sets $m = 0$ and two dimensions suffice, in agreement with (a): points on a line cannot be reordered by a flow, points in the plane can. For curves the bound is 4, and it is sharp, since circles link in $\mathbb{R}^3$. For connected data of dimension $m \geq 2$ the bound can be lowered to $2m + 1$ by the second part of [DTHC, Thm. 10.9]. Zhang et al. [ZGUA20] prove a related statement about maps: every homeomorphism of $\mathbb{R}^d$ can be embedded in a flow on $\mathbb{R}^{2d}$, and hence approximated by a neural ODE or an invertible residual network acting there. The dimension count of Theorem 7.6 depends on the data alone.

Residual networks fit the same description. A block $x \mapsto x + g(x)$ with g of Lipschitz constant less than one is a diffeomorphism isotopic to the identity (Proposition 8.2), and a composite of such blocks is a discretized isotopy. Depth, for such architectures, is time.

7.5 The space of representations

Open Problem 7.8. [programme] For a compact data manifold K of dimension m and width $k \geq m + 3$, the space $\mathrm{Emb}(K, \mathbb{R}^k)$ of representations that preserve the topology of the data is described by manifold calculus [DTHC, Thm. 10.10]: it is the limit of a tower whose first stage is the space of immersions and whose higher stages are built from configuration spaces of K. Determine which components and which homotopy classes of this space are reached by the realization map of a given architecture, as a function of width and depth.

For $k \geq 2m + 2$ the space is connected and the question is about higher homotopy; in the range $m + 3 \leq k \leq 2m + 1$ already the set of components is interesting, and Corollary 7.5(c) is the first example outside that range.

Notes

The picture of a network as progressively untangling data manifolds goes back at least to expository writing of Olah in 2014, where Theorem 7.2 for $d = 2$ and the role of an extra dimension are described. Minimal width for universal approximation is treated in [HS17, Joh19]. Neural ODEs are from [CRBD18], their limitations and augmentation from [DDT19], and approximation properties from [ZGUA20].

Exercises

Exercise 7.1. Show that the time-one map of a bounded smooth vector field on $\mathbb{R}$ is increasing.

Exercise 7.2. Write down a compactly supported vector field on $\mathbb{R}^2$ whose time-one map exchanges the points $(\pm 1, 0)$.

Exercise 7.3. Show that the linking number of two disjoint closed curves in $\mathbb{R}^3$ is unchanged when both are moved by a diffeomorphism isotopic to the identity.

Exercise 7.4. Let K be the curve $t \mapsto (\cos t, \sin t, \cos 2t, \sin 2t, \cos 3t)$ in $\mathbb{R}^5$. Show that the projection to the first two coordinates is an embedding of K, and that the projection to the third and fourth is not.

Chapter 8

Degree and Invertibility

Some architectures are designed so that every member is invertible: normalizing flows, invertible residual networks, the neural ODEs of the last chapter. Others are designed around a fixed-point equation. In both cases the relevant invariants are the degree of a map and the signed count of solutions of an equation, from [DTHC, Ch. 7].

8.1 Global invertibility

Theorem 8.1 (Hadamard). [classical] *A local diffeomorphism $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism if and only if it is proper.*

Proof. A proper local diffeomorphism is a covering map, and $\mathbb{R}^d$ is simply connected. In terms of degree: every point is a regular value, the sign of $\det Df$ is constant, and so the number of preimages of a point is the absolute value of the degree, which is constant; it is 1 because a covering of a simply connected space is trivial. □

A network all of whose layer Jacobians are invertible is a local diffeomorphism. The theorem says that the only way it can fail to be globally invertible is by sending a sequence tending to infinity to a bounded one.

Proposition 8.2 (Invertible residual blocks). [classical; Banach; Behrmann et al. [BGC19]] *Let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ have Lipschitz constant $L < 1$ and let $f = \mathrm{id} + g$.*

(a) f is a bijection, its inverse is Lipschitz with constant $(1-L)^{-1}$, and $x_{j+1} = y - g(x_j)$ converges to $f^{-1}(y)$ with $|x_j - f^{-1}(y)| \le L^j |x_0 - f^{-1}(y)|$.

(b) If g is smooth then f is a diffeomorphism with $\det Df > 0$.

(c) $f_s = \mathrm{id} + sg$, $0 \le s \le 1$, is a diffeotopy from the identity to f. In particular f is proper of degree 1.

Proof. (a) $x \mapsto y - g(x)$ is a contraction; and $|f(x) - f(x')| \ge (1-L)|x - x'|$. (b), (c): the eigenvalues of Dg have modulus at most L, so $1 + sDg$ is invertible for $s \in [0,1]$, and its determinant is continuous, non-zero, and equal to 1 at $s = 0$. □

Coupling layers and autoregressive layers are also isotopic to the identity, by scaling their conditioning networks to zero. So, apart from fixed permutations and reflections, the invertible architectures in common use all realize diffeomorphisms in the identity component of $\mathrm{Diff}(\mathbb{R}^d)$, and by [DTHC, Thm. 10.2] every such diffeomorphism with compact support is conversely the time-one map of a flow.

8.2 Topological mismatch

A normalizing flow models a probability measure ν on $\mathbb{R}^d$ as the image $h_*\mu$ of a simple measure μ, usually a Gaussian, under a diffeomorphism h.

Proposition 8.3. [classical; cf. Cornish et al. [CCDD20]] *For a homeomorphism h of $\mathbb{R}^d$ and a Borel probability measure μ, $\operatorname{supp}(h_*\mu) = h(\operatorname{supp}\mu)$. Hence if the supports of μ and ν are not homeomorphic, no homeomorphism satisfies $h_*\mu = \nu$.*

A Gaussian has support $\mathbb{R}^d$. A target supported on two disjoint balls, on a sphere, or on any proper closed subset cannot be represented exactly, and Cornish et al. show that a sequence of flows approximating such a target must have bi-Lipschitz constants tending to infinity. The remedies proposed in the literature are the ones differential topology suggests: a target supported on a submanifold $M^m \subset \mathbb{R}^d$ is modelled by a flow on $\mathbb{R}^m$ followed by an embedding $\mathbb{R}^m \to \mathbb{R}^d$, and when M is not diffeomorphic to an open subset of $\mathbb{R}^m$, by several such charts with a discrete index, which is an atlas in the sense of [DTHC, Ch. 1]. The minimal number of charts is an invariant of M; for a sphere it is two.

8.3 Equilibrium layers

A ***deep equilibrium layer*** outputs a solution z of $z = g_\theta(z, x)$ for a network g_θ. The existence and number of solutions is a fixed-point problem, and [DTHC, Prop. 7.11] describes its three aspects: a generic signed count, a derived fibre product, and a trace.

Theorem 8.4 (Equilibria of a bounded layer)**.** [reformulation; cf. [DTHC, Thm. 4.5, Thm. 7.2, Thm. 7.9]] *Let $g(z) = \tanh(Az + u)$ on $\mathbb{R}^k$, with $A \in \mathbb{R}^{k\times k}$ and $u \in \mathbb{R}^k$.*

(a) For every A and almost every u, every fixed point of g is non-degenerate: 1 is not an eigenvalue of Dg there. The fixed points are then finite in number.

(b) In that case $\sum_{g(z)=z} \operatorname{sign}\det(1 - Dg(z)) = 1$. In particular the number of fixed points is odd, and at least one exists.

(c) *Each non-degenerate fixed point is locally a smooth function of* (A, u)*, with* $dz = (1 - D_z g)^{-1}\, d_{(A,u)} g$.

Proof. (a) The map $\Psi(u, z) = z - \tanh(Az + u)$ has $D_u \Psi = -\operatorname{diag}(\operatorname{sech}^2(Az + u))$, which is invertible; so Ψ is a submersion and by [DTHC, Thm. 4.5] zero is a regular value of $\Psi(u, \cdot)$ for almost every u. Regularity at a zero z means that $1 - Dg(z)$ is invertible. All fixed points lie in the open cube $(-1, 1)^k$; being isolated in a compact set they are finite. (b) For $0 \le s \le 1$ the map $g_s(z) = \tanh(s(Az + u))$ sends the closed cube $Q = [-1, 1]^k$ into its interior, so $\mathrm{id} - g_s$ has no zero on ∂Q, and the degree at 0 of $\mathrm{id} - g_s \colon (Q, \partial Q) \to (\mathbb{R}^k, \mathbb{R}^k \setminus 0)$ is independent of s by [DTHC, Thm. 7.2]. At $s = 0$ it is the degree of the identity. At $s = 1$, zero being a regular value, it is the stated sum. (c) is the implicit function theorem. □

In one dimension, $z = \tanh(az + u)$ has one solution for $a \le 1$, and for $a > 1$ and small $|u|$ three, with signs $+, -, +$: two stable equilibria and an unstable one between them. The script finds the fixed points for several (a, u) and for a two-dimensional example and checks that the signed count is 1.

Part (c) is the formula used to differentiate through an equilibrium layer. It is valid exactly at the transverse fixed points, and (a) says that for almost every bias all fixed points are transverse. The exceptional parameters form the discriminant of the fixed-point problem, the analogue of the set Δ of Chapter 4; crossing it creates or destroys a pair of equilibria of opposite sign, as in the birth–death family of [DTHC, Thm. 8.12], and it is there that the output of an equilibrium layer can jump.

Notes

The global inverse function theorem is due to Hadamard and has many refinements (Caccioppoli, Banach and Mazur, Palais). Invertible residual networks are from [BGC19]. The observation that the topology of the support obstructs normalizing flows, with a quantitative form, is in [CCDD20]. Deep equilibrium models were introduced by Bai, Kolter and Koltun; conditions for uniqueness of the equilibrium (monotone operator parametrizations) are a separate line of work, and Theorem 8.4 describes what happens without them.

Exercises

Exercise 8.1. Prove Proposition 8.2(a).

Exercise 8.2. Let g be smooth with $\|Dg\| \le L < 1$ everywhere. Show $\det(1 + Dg) \ge (1 - L)^d$.

Exercise 8.3. For $k = 1$ and $u = 0$, find the fixed points of $z = \tanh(az)$ as a function of a and verify Theorem 8.4(b). What happens at $a = 1$?

Exercise 8.4. Prove Proposition 8.3.

Part IV

Assessment

Chapter 9

What Was Used, and What Is Missing

9.1 The inputs

Table 9.1 lists the main statements of this book with the theorems of the companion volume from which they follow. Every entry in the right-hand column was known by the mid-1970s.

statement here	from [DTHC]
backpropagation composes (Theorem 2.1)	reverse chain rule, Prop. 2.11
generic decision boundaries are hypersurfaces (Theorem 4.1)	Sard, Thm. 4.1; regular values, Thm. 1.12
chambers and walls (Theorem 4.4 and Proposition 4.5)	fibrewise lemma 4.4; Ehresmann, Thm. 8.5; isotopy extension, Thm. 10.3
every hypersurface is realized (Theorem 4.6)	tubular neighbourhoods, Thm. 6.2; openness, Thm. 3.2
surgery along paths (Theorems 5.1 and 5.2)	neat preimages, Thm. 1.17; handles, Thm. 8.6
moves between paths (Proposition 5.7)	Cerf, Thm. 8.12
interpolation manifold (Theorem 6.1)	regular values, Thm. 1.12; Sard, Thm. 4.1
barriers (Proposition 6.3)	sublevel functor, Cor. 8.7, Prop. 8.10
first layer embeds (Proposition 7.1)	parametric transversality, Thm. 4.5
neural ODEs are diffeotopies (Proposition 7.4)	flows, Thm. 10.2
augmentation (Theorem 7.6)	isotopy in the stable range, Thm. 10.9; Thm. 10.3
odd number of equilibria (Theorem 8.4)	degree, Thm. 7.2; Thm. 4.5

Table 9.1: What is used.

9.2 What is and is not new

Nothing in this book is a new theorem of differential topology, and most of its statements about networks have appeared in the machine learning literature in

some special case, with the references given in the text: generic smoothness of decision boundaries, the manifold of global minima, the obstruction to neural ODEs and its removal by augmentation, the mismatch of supports for flows. What the book adds is the uniform derivation, with the hypothesis that makes each statement true stated as a transversality property of the architecture (Principle 1.5), and a small number of statements tagged as observations that we have not found in this form: the description of the moduli of irreducible one-hidden-layer networks as a configuration space, with its monodromy (Proposition 3.5); the wall structure of the discriminant (Proposition 4.5); the surgery description of training with its Euler characteristic count (Theorem 5.2 and Proposition 5.3) and its moves (Proposition 5.7); the dimension count $2m + 2$ for augmentation in terms of the data (Theorem 7.6 and Corollary 7.7); and the parity of the number of equilibria (Theorem 8.4). Each is a short corollary and is offered as such.

The higher-categorical layer was used as language more than as a tool. The bicategory Para and the lens functor organize Chapters 1 and 2; the quotient by symmetries is a stack and not a space for homogeneous activations (Chapter 3); and a degenerate level set is a derived manifold with a cobordism class (Example 4.2). The theorems that do the work are those of Table 9.1.

9.3 Limits

Smoothness. The standing assumption excludes ReLU. For piecewise-linear networks the analogue of Theorem 4.1 is the theorem of Grigsby and Lindsey [GL22]; analogues of Theorems 4.4 and 5.2 would require stratified Morse theory, and are not proved here.

Genericity within the family. "Almost every parameter" is with respect to Lebesgue measure on Θ, so the statements apply to parameters drawn from a distribution with a density, as at a random initialization. They do not apply to distinguished parameters. In particular, if g is an orthogonal symmetry of the architecture, the gradient of an invariant loss is equivariant, and so the fixed-point set of g is invariant under gradient flow: a network initialized with two identical hidden units keeps them identical for ever, and stays in a reducible stratum of measure zero where none of the generic statements need hold.

Paths. Theorem 5.2 describes generic paths. A training trajectory is a particular path, the solution of a differential equation, and we have proved nothing about whether it is generic. The honest statement is Proposition 5.5: arbitrarily close to it there is a path with the stated behaviour.

Compactness. Input domains are assumed compact for the isotopy statements, and parameter space is never compact, which restricts the use of Morse

theory for losses (Proposition 6.5).

No rates. With the single exception of Proposition 6.2, the statements are qualitative. They say that a set has measure zero and not how small its neighbourhoods are; that an isotopy exists and not how large its velocity must be; that a boundary of a given topology is realizable and not at what width. Each of these has a quantitative counterpart in the literature that the present methods do not reach.

No learning theory. Nothing here concerns generalization.

9.4 Directions

Open Problem 9.1 (Piecewise-linear version)**.** [programme] Prove the analogues of Theorems 4.4 and 5.2 for ReLU networks that are generic and transversal in the sense of [GL22], with handle attachments replaced by the elementary moves of stratified Morse theory on the canonical polyhedral complex.

Open Problem 9.2 (Quantitative discriminant)**.** [programme] For a given architecture and initialization distribution, estimate the measure of the ε-neighbourhood of the discriminant Δ of Chapter 4, and hence the expected number of surgeries along a path of given length. Proposition 4.5 suggests that Δ is a hypersurface and that the answer is governed by its volume.

Open Problem 9.3 (The stack of networks)**.** [programme] For homogeneous activations, describe the quotient stack $[\Theta/G]$ of Chapter 3 and the critical locus of a loss on it as a derived object, so that Theorem 6.1 and the degeneracies listed in Chapter 6 become statements about a single space with a virtual dimension.

Together with Open Problem 7.8, these are the places where the methods of the companion volume, and in particular its higher-categorical parts (derived manifolds, stacks, the calculus of embeddings), would be doing more than providing a vocabulary.

Appendix A

Dictionary

Network notion	Differential topology	Where
architecture $F\colon \Theta \times X \to Y$	plot of the mapping object $[X, Y]$; morphism of Para	Definition 1.1 and Proposition 1.3
composition of layers	composition of plots	Definition 1.2
weight sharing, freezing	2-cells of Para	Exercise 1.3
forward-mode derivative	tangent functor	[DTHC, Prop. 2.1]
backpropagation	reverse derivative as a functor to lenses	Theorem 2.1
tangent kernel	Gram matrix JJ^{T} of the evaluation map	Proposition 2.3
functional dimension	rank of the realization map	Definition 2.5 and Proposition 2.4
permutation, sign, scaling symmetry	group of invertible 2-cells	Proposition 3.2
redundant hidden unit	reducible parameter; positive-dimensional fibre	Proposition 3.6
space of networks	quotient manifold, or stack	Proposition 3.5
decision boundary	preimage of a regular value	Theorem 4.1
multi-class boundaries	preimage of a stratification	Proposition 4.3
robust topology of a classifier	chamber of the discriminant complement	Theorem 4.4
change of boundary topology	surgery; critical point of the trace	Theorem 5.2
training run, seen in input space	embedded cobordism with a Morse function	Theorem 5.1
two runs with the same endpoints	Cerf moves	Proposition 5.7
interpolating parameters	fibre of the evaluation map over a regular value	Theorem 6.1
sharpness at a minimum	normal Hessian $J^{\mathsf{T}}J$	Proposition 2.6
loss barrier, mode connectivity	death level in the degree-zero barcode	Proposition 6.3

Network notion	Differential topology	Where
saturation, diverging weights	failure of properness	Proposition 6.5
hidden representation	immersion or embedding of the data manifold	Proposition 7.1
minimal width	invariance of domain	Theorem 7.2
neural ODE	diffeotopy	Proposition 7.4
augmented neural ODE	isotopy of embeddings in the stable range	Theorem 7.6
residual block with small branch	diffeomorphism isotopic to the identity	Proposition 8.2
normalizing flow	diffeomorphism; pushforward of measures	Proposition 8.3
equilibrium layer	transverse fixed point; Lefschetz count	Theorem 8.4

Appendix B

Computations

The script `verify-parametrized-networks.py` uses only the Python standard library. It implements small feedforward networks (widths $(2,3,1)$, $(2,4,4,1)$ and a few others) with forward evaluation, forward-mode differentiation by dual numbers and reverse-mode differentiation by backpropagation, a Runge–Kutta integrator, a Jacobi eigenvalue routine and a flood-fill component counter. It performs 49 checks in about three seconds, each printing `PASS` or `FAIL`, and exits with status zero if and only if all pass. The checks recompute the examples in the text and test numerical instances of the structural statements on random parameters with a fixed seed. They are not proofs, and a statement about almost every parameter is tested only at the parameters sampled.

Chapter	What is checked
1	parameter counts (Exercise 1.1); associativity of composition in Para; the derivative of a weight-tied network (Exercise 1.3); the family $\theta + x^3$ (Exercise 1.4)
2	backpropagation against forward mode and finite differences for two architectures, and for the input covector (Theorem 2.1); the tangent kernel is positive semidefinite of rank rk J; the output dynamics $\dot{u} = -K(u - y)$ (Proposition 2.3); the loss gradient; the Hessian as Gauss–Newton plus curvature term (Exercise 2.3)
3	all 48 symmetries of the tanh network $(2,3,1)$ and the size of an orbit (Proposition 3.2); the scaling symmetry for ReLU; rank $P = 13$ at a random parameter and ranks 10 and 9 at reducible ones (Theorem 3.4, Exercise 3.2); for ReLU, the generators ξ_j in $\ker J$ and $\operatorname{rk} J \leq P - H$ (Proposition 3.7); a path of irreducible networks exchanging two hidden units (Proposition 3.5)
4	$\partial F/\partial c = 1$; degenerate zeros of a one-dimensional network occur only at the critical values (Theorem 4.1(a)); the two chambers of $x_1^2 - x_2^2 + c$ (Exercise 4.3); three sectors at a triple point (Exercise 4.4); the submersion of Proposition 4.5
5	component counts, critical points and indices in Example 5.4; the Euler characteristic formula of Proposition 5.3 for $d \leq 6$; the Hessian of the height function on the trace (Exercise 5.4); births and deaths of boundary points on a circle (Exercise 5.2)

Chapter	What is checked
6	rk $J = N$ and the dimension $P - N$ of the solution set; the Hessian $J^{\mathsf{T}}J$ with 4 positive and 9 zero eigenvalues (Theorem 6.1); the inequality of Proposition 6.2; sublevel components, the barrier and the Hessians of Example 6.4; Proposition 6.5; Exercise 6.4; invariance of the loss under the symmetries
7	random first layers of width 3 embed a circle in $\mathbb{R}^5$ (Proposition 7.1), and a coordinate projection that does not (Exercise 7.4); the convex hull property of Theorem 7.2 for random injective networks; order preservation on the line; positivity of the Jacobian determinant and Liouville's formula for a planar neural ODE (Proposition 7.4); the swap of Exercise 7.2; invariance of the linking number under a neural ODE in $\mathbb{R}^3$ (Corollary 7.5); separation of nested circles after augmentation (Corollary 7.7)
8	convergence of the fixed-point inversion and the determinant bound for a residual block (Proposition 8.2, Exercise 8.2); a non-proper local diffeomorphism (Theorem 8.1); odd numbers of equilibria with signed count 1 in one and two dimensions (Theorem 8.4)

Appendix C

Solutions to the Exercises

Solutions to all exercises follow, by chapter. Where a solution relies on a computation, the computation is also carried out by `verify-parametrized-networks.py`.

Chapter 1

1.1. $(2,3,1)$: $P = 3\cdot 3 + 1\cdot 4 = 13$, $H = 3$. $(2,4,4,1)$: $P = 4\cdot 3 + 4\cdot 5 + 1\cdot 5 = 37$, $H = 8$. $(d,k,1)$: $P = k(d+1) + (k+1) = k(d+2) + 1$, $H = k$.

1.2. Let $(\Theta, F)\colon X \to Y$, $(\Theta', G)\colon Y \to Z$, $(\Theta'', K)\colon Z \to V$. Both bracketings have underlying function $(\theta, \theta', \theta'', x) \mapsto K(\theta'', G(\theta', F(\theta, x)))$, on the parameter spaces $(\Theta \times \Theta') \times \Theta''$ and $\Theta \times (\Theta' \times \Theta'')$ respectively. The reassociation diffeomorphism r satisfies $F_{\text{left}} = F_{\text{right}} \circ (r \times \mathrm{id}_X)$, so r and r^{-1} are 2-cells, mutually inverse.

1.3. The diagonal $\delta\colon \Theta \to \Theta \times \Theta$ gives a 2-cell $(\Theta, (F \circ F) \circ (\delta \times \mathrm{id})) \Rightarrow (\Theta \times \Theta, F \circ F)$, and its source is $(\theta, x) \mapsto F(\theta, F(\theta, x))$, the weight-tied network. By the chain rule, the derivative of $\theta \mapsto G(\delta(\theta))$ is $\partial_1 G + \partial_2 G$ evaluated on the diagonal, for $G(\theta_1, \theta_2) = F(\theta_2, F(\theta_1, x))$.

1.4. $F(\theta, x) = \theta + x^3$. Every member has the degenerate critical point $x = 0$. A function is Morse when its 1-jet prolongation is transverse to the set of jets with vanishing derivative; the family of prolongations $(\theta, x) \mapsto (x, \theta + x^3, 3x^2)$ is not transverse to $\{\text{third coordinate} = 0\}$, since the third coordinate has vanishing derivative in both θ and x at $x = 0$. The parameter moves the value of the function but not its derivative, so the family is rich for conditions on values (Principle 1.5) and not for conditions on first derivatives. Adding a parameter $\theta_1 x$ repairs this: $\theta_0 + \theta_1 x + x^3$ is Morse for $\theta_1 \neq 0$.

Chapter 2

2.1. Let $z = Wx + b$ and $h = \sigma(z)$. The lens of the layer has forward part $(W, b, x) \mapsto h$ and backward part $(W, b, x, w) \mapsto \left(\delta x^{\mathsf{T}},\ \delta,\ W^{\mathsf{T}}\delta\right)$ with $\delta = \sigma'(z) \odot w$: the three components are the covectors for W, b and x. The output layer $(a, c; h) \mapsto a^{\mathsf{T}}h + c$ has backward part $(a, c, h, w) \mapsto (wh,\ w,\ wa)$. The

loss $\frac{1}{2}(u-y)^2$ has gradient $e = u - y$ in u. Composing: the covector entering the output layer is e, so $\nabla_a = e\,h$, $\nabla_c = e$, and the covector passed back is $e\,a$; then $\delta = \sigma'(z) \odot (e\,a)$, $\nabla_W = \delta x^\mathsf{T}$, $\nabla_b = \delta$.

2.2. $\ker JJ^\mathsf{T} = \ker J^\mathsf{T}$ since $\langle v, JJ^\mathsf{T}v\rangle = |J^\mathsf{T}v|^2$; so $\operatorname{rk} K = \operatorname{rk} J^\mathsf{T} = \operatorname{rk} J$. For scalar output K is $n \times n$, invertible if and only if $\operatorname{rk} J = n$, that is, if and only if the columns of J span $\mathbb{R}^n$; the k-th column is the vector $\big(\partial F/\partial\theta_k(\theta, x_i)\big)_i$.

2.3. By [DTHC, Prop. 2.9] the 2-jet of $\lambda \circ \mathrm{ev}$ at θ has linear part CA and quadratic part $C \circ B + E \circ (A \times A)$, where (A, B) is the 2-jet of ev (so $A = J$ and $B = (\nabla^2 \mathrm{ev}_k)_k$) and (C, E) that of λ (so $C = \nabla\lambda^\mathsf{T}$, $E = \nabla^2\lambda$). The quadratic part is $\sum_k \partial_k\lambda\, \nabla^2\mathrm{ev}_k + J^\mathsf{T}\nabla^2\lambda\, J$.

2.4. $\mathrm{ev}(\theta) = \Phi\theta$ with Φ the $n \times P$ matrix of rows $\phi(x_i)^\mathsf{T}$; so $J = \Phi$ for all θ and $K = \Phi\Phi^\mathsf{T}$, the Gram matrix of the features. The fibres of ev are the affine subspaces $\theta_0 + \ker\Phi$, of dimension $P - \operatorname{rk}\Phi$. The realization map is linear with kernel $\{\theta : \langle\theta, \phi(x)\rangle = 0\ \forall x\}$, so the functional dimension is the dimension of the span of $\phi(X)$.

Chapter 3

3.1. $F(\theta, x) = \sum_j a_j\sigma(\langle w_j, x\rangle + b_j) + c$. (i) Permuting the index j reorders the sum. (ii) If σ is odd, $(-a_j)\sigma(-\langle w_j, x\rangle - b_j) = a_j\sigma(\langle w_j, x\rangle + b_j)$. (iii) If $\sigma(\lambda s) = \lambda\sigma(s)$ for $\lambda > 0$, then $(a_j/\lambda)\sigma(\lambda\langle w_j, x\rangle + \lambda b_j) = a_j\sigma(\langle w_j, x\rangle + b_j)$.

3.2. The columns of J belonging to (w_j, b_j) are $\big(a_j\sigma'(\langle w_j, x_i\rangle + b_j)\, x_i\big)_i$ and $\big(a_j\sigma'(\langle w_j, x_i\rangle + b_j)\big)_i$; if $a_j = 0$ these $d+1$ columns vanish, so $\operatorname{rk} J \le P - (d+1)$. If $(w_j, b_j) = (w_k, b_k)$, the columns for a_j and a_k are both $\big(\sigma(\langle w_j, x_i\rangle + b_j)\big)_i$, so two columns coincide and $\operatorname{rk} J \le P - 1$. (In the second case the columns for (w_j, b_j) and (w_k, b_k) are also proportional, with ratio a_j/a_k, so the rank drops by $d+2$ in all.)

3.3. $\theta = (w, b, a, c)$ and $\xi = (w, b, -a, 0)$. Where $wx + b > 0$, $F = a(wx+b) + c$ and $D_\theta F = (ax,\ a,\ wx + b,\ 1)$, so $D_\theta F \cdot \xi = awx + ab - a(wx+b) = 0$. Where $wx + b < 0$, $F = c$ and $D_\theta F = (0, 0, 0, 1)$, again orthogonal to ξ. Here $P = 4$, $H = 1$, so the functional dimension is at most 3; it equals 3 when $a \neq 0$ and $w \neq 0$, since the function $x \mapsto a\ \mathrm{ReLU}(wx + b) + c$ is determined by its breakpoint $-b/w$, the slope aw on the active side, and c (and the side, which is discrete).

3.4. $(2, 3, 1)$: one hidden layer of 3 units, $|G| = 2^3 \cdot 3! = 48$. $(2, 4, 4, 1)$: two hidden layers of 4 units, $|G| = (2^4 \cdot 4!)^2 = 384^2 = 147456$.

Chapter 4

4.1. By Theorem 4.1(c), for almost every θ the set B_θ is a compact neat 0-dimensional submanifold of $[-R,R]$: a finite set of interior points at which $f'_\theta \neq 0$, with $f_\theta(\pm R) \neq 0$. At each such point f_θ changes sign. So the number of zeros is even if $f_\theta(-R)$ and $f_\theta(R)$ have the same sign and odd otherwise.

4.2. A compact neat 1-dimensional submanifold of the disc: a finite disjoint union of circles in the interior and arcs with both endpoints on the boundary circle, meeting it transversally. Any such configuration, up to isotopy, is determined by the nesting of the circles and the pairing of the endpoints by non-crossing arcs.

4.3. $d_x F = (2x_1, -2x_2)$ vanishes only at $x = 0$, where $F = c$. So $\Sigma = \{(0,0)\}$ and $\Delta = \{0\}$. There are two chambers, $c > 0$ and $c < 0$. In each the boundary is a hyperbola; for $c < 0$ its branches open along the x_1-axis ($x_1^2 = x_2^2 - c \geq -c$) and for $c > 0$ along the x_2-axis. The region $\{f < 0\}$ is connected for $c < 0$ (it contains the origin) and has two components for $c > 0$. (The plane is not compact; the statement about chambers holds here because the family is a product near infinity up to scaling. Restricted to a disc the same description holds.)

4.4. At a triple point p, f_θ is transverse to $A_{\{1,2,3\}}$, the line $y_1 = y_2 = y_3$ in $\mathbb{R}^3$, of codimension 2. So the map $x \mapsto (f_1 - f_2, f_2 - f_3)$ from the plane to $\mathbb{R}^2$ is a local diffeomorphism at p, and in the coordinates $(u,v) = (f_1 - f_2, f_2 - f_3)$ the decision regions are $\{u > 0,\ u + v > 0\}$, $\{u < 0,\ v > 0\}$ and $\{v < 0,\ u + v < 0\}$: three sectors bounded by the rays $\{u = 0, v > 0\}$, $\{v = 0, u > 0\}$ and $\{u + v = 0, u < 0\}$.

Chapter 5

5.1. $g = (x_1^2 - 1)^2 + x_2^2$ has $\nabla g = (4x_1(x_1^2 - 1), 2x_2)$, vanishing at $(0,0)$ and $(\pm 1, 0)$. The Hessian is $\operatorname{diag}(12x_1^2 - 4,\ 2)$: at $(\pm 1, 0)$ it is $\operatorname{diag}(8, 2)$, index 0, value 0; at the origin $\operatorname{diag}(-4, 2)$, index 1, value 1. The sublevel set $\{g \leq t\}$ is empty for $t < 0$, two discs around $(\pm 1, 0)$ for $0 < t < 1$ (on the line $x_2 = 0$ it is $\{|x_1^2 - 1| \leq \sqrt{t}\}$, two intervals when $t < 1$), and one disc for $t > 1$. Its Euler characteristic is $0, 2, 1$, consistent with attaching two 0-cells and then one 1-cell.

5.2. Here B_t is a finite set and W_γ is a compact 1-manifold with boundary $B_0 \sqcup B_1$. A critical point of τ on a 1-manifold is a local minimum or maximum of t along a curve: two points of B_t appear together or disappear together (surgery with $\lambda = 0$ or 1 on a 0-manifold: add or remove an S^0). Between

critical times the points move continuously. If X is a circle, f_t changes sign at each point of B_t and returns to its original sign after one circuit, so $\#B_t$ is even for every regular t.

5.3. Two. By Proposition 5.3 with $d = 3$, the Euler characteristic of the region changes by $\pm(-1)^k$ at each surgery; the solid torus has $\chi = 0$ and needs at least one 0-handle to be non-empty, so at least one more handle of odd index is needed. A 0-handle (a ball appears, $k = 0$) followed by a 1-handle attached to it (the ball becomes a solid torus, $k = 1$) suffices. The boundary goes from empty to a sphere to a torus.

5.4. Write $\Phi = \Phi_\gamma$ and let $t(x)$ be defined by $\Phi(t(x), x) = 0$ near x_i, with $t(x_i) = t_i$. Differentiating once, $\Phi_t\,\partial_j t + \partial_j \Phi = 0$, so $\partial_j t(x_i) = 0$ because $d_x\Phi = 0$ there. Differentiating again and evaluating at x_i, where the terms containing first derivatives of t vanish, gives $\Phi_t\,\partial_j\partial_k t + \partial_j\partial_k\Phi = 0$. Hence $\nabla^2 t(x_i) = -\nabla_x^2\Phi(t_i, x_i)/\Phi_t(t_i, x_i)$.

Chapter 6

6.1. $\mathrm{ev}(\theta) = \Phi\theta$ with Φ of rank n. Every y is a regular value; $M_y = \theta^* + \ker\Phi$ is an affine subspace of dimension $P - n$, where θ^* is any solution, for instance the minimum-norm one $\Phi^{\mathsf{T}}(\Phi\Phi^{\mathsf{T}})^{-1}y$. The Hessian is the constant matrix $\Phi^{\mathsf{T}}\Phi$, of rank n, with kernel $\ker\Phi$.

6.2. $\nabla\mathcal{L} = (ab-1)(b, a)$, which vanishes when $ab = 1$ or $(a, b) = 0$. The Hessian is $\left(\begin{smallmatrix} b^2 & 2ab-1 \\ 2ab-1 & a^2 \end{smallmatrix}\right)$. On $ab = 1$ it is $\left(\begin{smallmatrix} b^2 & 1 \\ 1 & a^2 \end{smallmatrix}\right)$, of determinant $a^2b^2 - 1 = 0$ and trace $a^2 + b^2 > 0$: rank 1, with kernel spanned by $(a, -b)$, the tangent to the hyperbola. At the origin it is $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$, with eigenvalues ± 1: a saddle of index 1 with value $\frac{1}{2}$. The sublevel set is $\{1 - \sqrt{2c} \le ab \le 1 + \sqrt{2c}\}$. For $c < \frac{1}{2}$ the lower bound is positive, and the set lies in the two quadrants $ab > 0$, one component in each. For $c \ge \frac{1}{2}$ it contains the axes and is connected. Any path from $(1, 1)$ to $(-1, -1)$ must cross the set $ab = 0$, where $\mathcal{L} = \frac{1}{2}$.

6.3. The flow line γ has length at most $\ell = \sqrt{2\mathcal{L}(\theta_0)/\mu}$ while it stays in U. If the closed ball of radius ℓ about θ_0 lies in U, the flow line cannot leave U: at the first exit time it would be at distance at most its length, which is at most ℓ, from θ_0, hence still inside. So the flow line is defined for all time, has finite length, and therefore converges; the limit θ_∞ satisfies $|\theta_\infty - \theta_0| \le \ell$ and $\mathcal{L}(\theta_\infty) = \lim \mathcal{L}(\gamma(t)) = 0$.

6.4. $\mathcal{L} = \frac{1}{2}(\prod a_i - 1)^2$ and $M_y = \{\prod a_i = 1\}$, the preimage of the regular value 1 of the product map (its gradient at a point with all $a_i \neq 0$ is non-zero); so $\dim M_y = k - 1$. On M_y every $a_i \neq 0$ and the number of negative a_i is even;

each admissible sign pattern gives one component (a graph over the positive orthant in the first $k-1$ coordinates), and there are 2^{k-1} sign patterns with an even number of minus signs. The gradient is $(\prod a_i - 1)\big(\prod_{j\neq i} a_j\big)_i$, which vanishes at the origin for $k \geq 2$; the Hessian at the origin has entries $-\partial_i \partial_j \prod a_l$, which are zero for $k \geq 3$. So the origin is a critical point with zero Hessian.

Chapter 7

7.1. Let φ^t be the flow. If $x < y$ then $\varphi^t(x) \neq \varphi^t(y)$ for all t by uniqueness of solutions, so $\varphi^t(y) - \varphi^t(x)$ is continuous in t, never zero, and positive at $t = 0$; hence positive at $t = 1$.

7.2. Let $\rho\colon [0,\infty) \to [0,1]$ be smooth with $\rho = 1$ on $[0,2]$ and $\rho = 0$ on $[3,\infty)$, and let $v(x) = \pi\rho(|x|)\,(-x_2, x_1)$. The flow preserves $|x|$, since v is tangent to circles, and on $\{|x| \leq 2\}$ it is rotation with angular speed π. At time one the points $(\pm 1, 0)$ have been rotated by the angle π and are exchanged.

7.3. Let φ_t be the diffeotopy and $K_t = \varphi_t(K)$, $L_t = \varphi_t(L)$; they are disjoint for every t because φ_t is injective. The Gauss maps $G_t\colon K \times L \to S^2$, $G_t(x,y) = (\varphi_t(y) - \varphi_t(x))/|\varphi_t(y) - \varphi_t(x)|$, form a homotopy, and the degree is a homotopy invariant by [DTHC, Thm. 7.2].

7.4. The first projection is $t \mapsto (\cos t, \sin t)$, the standard embedding of the circle. The second is $t \mapsto (\cos 2t, \sin 2t)$, which takes the same value at t and $t + \pi$; it is an immersion but not injective, a double cover of the circle.

Chapter 8

8.1. For fixed y the map $T(x) = y - g(x)$ satisfies $|T(x) - T(x')| \leq L|x - x'|$, so it has a unique fixed point x^*, which is the unique solution of $x + g(x) = y$, and $|x_j - x^*| = |T^j(x_0) - T^j(x^*)| \leq L^j|x_0 - x^*|$. So f is a bijection. And $|f(x) - f(x')| \geq |x - x'| - |g(x) - g(x')| \geq (1 - L)|x - x'|$, which says that f^{-1} is Lipschitz with constant $(1 - L)^{-1}$.

8.2. The singular values of $1 + Dg$ are at least $1 - L$, since $|(1 + Dg)v| \geq |v| - |Dg\,v| \geq (1 - L)|v|$. The absolute value of the determinant is the product of the singular values, so it is at least $(1 - L)^d$; and the determinant is positive by Proposition 8.2(b).

8.3. $z = 0$ is always a fixed point, with $1 - g'(0) = 1 - a$. For $a \leq 1$ it is the only one: $\tanh(az)$ is strictly concave for $z > 0$ with slope $a \leq 1$ at 0, so $\tanh(az) < z$ for $z > 0$, and symmetrically for $z < 0$. For $a > 1$, $\tanh(az) > z$ for small $z > 0$ and $\tanh(az) < 1 < z$ for $z > 1$, so there is a fixed point

$z^* \in (0,1)$, unique by concavity, and $-z^*$ by symmetry. At z^* the graph of $\tanh(az)$ crosses the diagonal from above, so $g'(z^*) < 1$ and the sign is $+$; at 0 the sign of $1 - a$ is $-$. The signed count is $1 - 1 + 1 = 1$. For $a < 1$ it is the single $+1$. At $a = 1$ the fixed point 0 is degenerate ($1 - g'(0) = 0$): this is the exceptional parameter at which the pair $\pm z^*$ is born, a pitchfork, which is not a generic birth–death because of the symmetry $z \mapsto -z$ present when $u = 0$.

8.4. A point y lies in $\operatorname{supp}(h_*\mu)$ if and only if every open neighbourhood V of y has $\mu(h^{-1}(V)) > 0$. As h is a homeomorphism, the sets $h^{-1}(V)$ are exactly the open neighbourhoods of $h^{-1}(y)$; so the condition is $h^{-1}(y) \in \operatorname{supp}\mu$. If $h_*\mu = \nu$ then h restricts to a homeomorphism from $\operatorname{supp}\mu$ to $\operatorname{supp}\nu$.

References

[BGC19] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, J.-H. Jacobsen, Invertible residual networks, in *ICML 2019*, PMLR 97, 573–582.

[BS14] M. Bianchini, F. Scarselli, On the complexity of neural network classifiers: a comparison between shallow and deep architectures, *IEEE Trans. Neural Netw. Learn. Syst.* 25 (2014), 1553–1565.

[CCDD20] R. Cornish, A. Caterini, G. Deligiannidis, A. Doucet, Relaxing bijectivity constraints with continuously indexed normalising flows, in *ICML 2020*, PMLR 119.

[CCG20] R. Cockett, G. Cruttwell, J. Gallagher, J.-S. P. Lemay, B. MacAdam, G. Plotkin, D. Pronk, Reverse derivative categories, in *Computer Science Logic 2020*, LIPIcs 152, 18:1–18:16.

[CGGWZ22] G. S. H. Cruttwell, B. Gavranović, N. Ghani, P. Wilson, F. Zanasi, Categorical foundations of gradient-based learning, in *ESOP 2022*, Lecture Notes in Computer Science 13240, Springer, 1–28.

[Coo21] Y. Cooper, Global minima of overparameterized neural networks, *SIAM J. Math. Data Sci.* 3 (2021), 676–691.

[CRBD18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, Neural ordinary differential equations, in *NeurIPS 2018*.

[DDT19] E. Dupont, A. Doucet, Y. W. Teh, Augmented neural ODEs, in *NeurIPS 2019*.

[DTHC] Anonymous, *Differential Topology in Higher Categories*, Higher Categorical Semantics, 2026. Companion volume.

[DVSH18] F. Draxler, K. Veschgini, M. Salmhofer, F. Hamprecht, Essentially no barriers in neural network energy landscape, in *ICML 2018*, PMLR 80.

[EKSN22] R. Entezari, H. Sedghi, O. Saukh, B. Neyshabur, The role of permutation invariance in linear mode connectivity of neural networks, in *ICLR 2022*.

[FA00] K. Fukumizu, S. Amari, Local minima and plateaus in hierarchical structures of multilayer perceptrons, *Neural Networks* 13 (2000), 317–327.

[Fef94] C. Fefferman, Reconstructing a neural net from its output, *Rev. Mat. Iberoamericana* 10 (1994), 507–555.

[FMN16] C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesis, *J. Amer. Math. Soc.* 29 (2016), 983–1049.

[FST19] B. Fong, D. I. Spivak, R. Tuyéras, Backprop as functor: a compositional perspective on supervised learning, in *LICS 2019*, IEEE, 2019.

[Fuk96] K. Fukumizu, A regularity condition of the information matrix of a multilayer perceptron network, *Neural Networks* 9 (1996), 871–879.

[GIP18] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, A. G. Wilson, Loss surfaces, mode connectivity, and fast ensembling of DNNs, in *NeurIPS 2018*.

[GL22] J. E. Grigsby, K. Lindsey, On transversality of bent hyperplane arrangements and the topological expressiveness of ReLU neural networks, *SIAM J. Appl. Algebra Geom.* 6 (2022), 216–242.

[GLMW22] J. E. Grigsby, K. Lindsey, R. Meyerhoff, C. Wu, Functional dimension of feedforward ReLU neural networks, preprint, arXiv:2209.04036.

[Hir76] M. W. Hirsch, *Differential Topology*, Graduate Texts in Mathematics 33, Springer, 1976.

[Hor91] K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural Networks* 4 (1991), 251–257.

[HS17] B. Hanin, M. Sellke, Approximating continuous functions by ReLU nets of minimal width, preprint, arXiv:1710.11278.

[JGH18] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, in *NeurIPS 2018*.

[Joh19] J. Johnson, Deep, skinny neural networks are not universal approximators, in *ICLR 2019*.

[LZB22] C. Liu, L. Zhu, M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks, *Appl. Comput. Harmon. Anal.* 59 (2022), 85–116.

[Ngu19] Q. Nguyen, On connected sublevel sets in deep learning, in *ICML 2019*, PMLR 97.

[NZL20] G. Naitzat, A. Zhitnikov, L.-H. Lim, Topology of deep neural networks, *J. Mach. Learn. Res.* 21 (2020), no. 184, 1–40.

[PL20] M. Phuong, C. H. Lampert, Functional vs. parametric equivalence of ReLU networks, in *ICLR 2020*.

[SGJ21] B. Şimşek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, J. Brea, Geometry of the loss landscape in overparameterized neural networks: symmetries and invariances, in *ICML 2021*, PMLR 139.

[Sus92] H. J. Sussmann, Uniqueness of the weights for minimal feedforward nets with a given input-output map, *Neural Networks* 5 (1992), 589–593.

[Wat09] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

[ZGUA20] H. Zhang, X. Gao, J. Unterman, T. Arodz, Approximation capabilities of neural ODEs and invertible residual networks, in *ICML 2020*, PMLR 119.