Abide by the Law and Follow the Flow:
Conservation Laws for Gradient Flows

Sibylle Marcotte
ENS - PSL Univ.
sibylle.marcotte@ens.fr

Rémi Gribonval
Univ Lyon, EnsL, UCBL,
CNRS, Inria, LIP,
remi.gribonval@inria.fr

Gabriel Peyré
CNRS, ENS - PSL Univ.
gabriel.peyre@ens.fr

Abstract

Understanding the geometric properties of gradient descent dynamics is a key ingredient in deciphering the recent success of very large machine learning models. A striking observation is that trained over-parameterized models retain some properties of the optimization initialization. This "implicit bias" is believed to be responsible for some favorable properties of the trained models and could explain their good generalization properties. The purpose of this article is threefold. First, we rigorously expose the definition and basic properties of "conservation laws", which are maximal sets of independent quantities conserved during gradient flows of a given model (e.g. of a ReLU network with a given architecture) with any training data and any loss. Then we explain how to find the exact number of these quantities by performing finite-dimensional algebraic manipulations on the Lie algebra generated by the Jacobian of the model. Finally, we provide algorithms (implemented in SageMath) to: a) compute a family of polynomial laws; b) compute the number of (not necessarily polynomial) conservation laws. We provide showcase examples that we fully work out theoretically. Besides, applying the two algorithms confirms for a number of ReLU network architectures that all known laws are recovered by the algorithm, and that there are no other laws. Such computational tools pave the way to understanding desirable properties of optimization initialization in large machine learning models.

1 Introduction

State-of-the-art approaches in machine learning rely on the conjunction of gradient-based optimization with vastly "over-parameterized" architectures. A large body of empirical [27] and theoretical [4] works suggest that, despite the ability of these models to almost interpolate the input data, they are still able to generalize well. Analyzing the training dynamics of these models is thus crucial to gain a better understanding of this phenomenon. Of particular interest is to understand what properties of the initialization are preserved during the dynamics, which is often loosely referred to as being an "implicit bias" of the training algorithm. The goal of this article is to make this statement precise, by properly defining maximal sets of such "conservation laws", by linking these quantities to algebraic computations (namely a Lie algebra) associated with the model parameterization (in our framework, this parameterization is embodied by a mapping ϕ), and finally by exhibiting algorithms to implement these computations in SageMath [26].

Over-parameterized model Modern machine learning practitioners and researchers have found that over-parameterized neural networks (with more parameters than training data points), which are often trained until perfect interpolation, have impressive generalization properties [27, 4]. This performance seemingly contradicts classical learning theory [22], and a large part of the theoretical deep learning literature is aimed at explaining this puzzle. The choice of the optimization algorithm is crucial to the model generalization performance [9, 18, 12], thus inducing an *implicit bias*.

Implicit bias   The terminology "implicit bias" informally refers to properties of trained models which are induced by the optimization procedure, typically some form of regularization [19]. For gradient descent, in simple cases such as scalar linear neural networks or two-layer networks with a single neuron, it is actually possible to compute in closed form the implicit bias, which induces some approximate or exact sparsity regularization [9]. Another interesting case is logistic classification on separable data, where the implicit bias selects the max-margin classifier both for linear models [23] and for two-layer neural networks in the mean-field limit [7]. The key hypothesis to explicit the implicit bias is that the Riemannian metric associated to the over-parameterization is of Hessian type [9], which is a very strong constraint. Unfortunately, even for matrix factorization (so more than a single neuron), this is not the case, and no closed form is known for the implicit bias [10]. The work of [15] gives conditions on the over-parameterization for this to be possible (for instance the Lie brackets should vanish: they are (as could be expected

Conservation laws   Finding functions conserved during gradient flow optimization of neural networks (a continuous limit of gradient descent often used to model the optimization dynamics) is particularly useful to better understand the flow behavior. One can see conservation laws as a "weak" form of implicit bias: to explain, among a possibly infinite set of minimizers, which properties (e.g. in terms of sparsity, low-rank, etc.) are being favored by the dynamic. If there are enough conservation laws, one has an exact description of the dynamic, and in some cases, one can even determine explicitly the implicit bias. Otherwise, one can still predict what properties of the initialization are retained at convergence, and possibly leverage this knowledge. For example, in the case of linear neural networks, certain *balancedness properties* are satisfied and provide a class of conserved functions [21, 8, 1, 2, 13, 25, 16]. These conservation laws enable for instance to prove the global convergence of the gradient flow [3] under some assumptions. We detail these laws in Proposition 4.1. A subset of these "balancedness" laws still holds in the case of a ReLU activation [8], which reflects the scaling invariance of these networks (see Section 4 for more details). More generally such conservation laws are a consequence [14] of the invariances of the model: to each 1-parameter group of transformation preserving the loss, one can associate a conserved quantity, which is in some sense analogous to Noether's theorem [20]. Similar reasoning is used by [28] to show the influence of initialization on convergence and generalization performance of the neural network. Our work is somehow complementary to this line of research: instead of assuming a priori known symmetries, we directly analyze the model and give access to conservation laws using algebraic computations. For matrix factorization as well as for certain ReLU network architectures, this allows us to show that the conservation laws reported in the literature are complete (there are no other independent quantities that would be preserved by all gradient flows).

Contributions

We formalize the notion of a conservation law, a quantity preserved through all gradient flows given a model architecture (e.g. a ReLU neural network with prescribed layers) and a family of "data-fidelity functions", typically associated to the empirical loss on a training set. Our main contributions are:

- to show that for several classical losses, characterizing conservation laws for deep linear (resp. shallow ReLU) networks boils down to analyzing a finite dimensional space of vector fields;

- to propose an algorithm (coded in SageMath) identifying polynomial conservation laws on linear / ReLU network architectures; it identifies all known laws on selected examples;

- to formally define the maximum number of (not necessarily polynomial) independent conservation laws and characterize it a) theoretically via Lie algebra computations; and b) practically via an algorithm (coded in SageMath) computing this number on worked examples;

- to illustrate that in certain settings these findings allow to rewrite an over-parameterized flow as an "intrinsic" low-dimensional flow;

- to highlight that the cost function associated to the training of linear and ReLU networks, shallow or deep, with various losses (quadratic and more) fully fits the proposed framework.

A consequence of our results is to show for the first time that conservation laws commonly reported in the literature are maximal: there is no other independent preserved quantity (see Propositions 4.3, 4.2, Corollary 4.4) and Section 4.2).

2 Conservation Laws for Gradient Flows

After some reminders on gradient flows, we formalize the notion of conservation laws.

2.1 Over-parameterized models

We consider learning problems, where we denote $x_i \in \mathbb{R}^m$ the features and $y_i \in \mathcal{Y}$ the targets (for regression) or labels (for classification) in the case of supervised learning, while y_i can be considered constant for unsupervised/self-supervised learning. The prediction is performed by a parametric mapping $g_\theta : \mathbb{R}^m \to \mathbb{R}^n$ (for instance a neural network) which is trained by empirical risk minimization of a **cost** $\mathcal{E}$

$$\min_{\theta \in \mathbb{R}^D} \mathcal{E}(\theta) := \sum_i \ell(g_\theta(x_i), y_i), \tag{1}$$

where ℓ is the **loss** function (for regression, one typically has $\mathcal{Y} = \mathbb{R}^n$). The goal of this paper is to analyze what are the functions $h(\theta)$ which are preserved during the optimization by gradient descent of the cost $\mathcal{E}(\theta)$. To make the mathematical analysis tractable and provide algorithmic procedure to determine these functions, our fundamental hypothesis is that the cost $\mathcal{E}$ can be factored – at least *locally*, in a sense that will be made precise – in the form

$$\forall \theta \in \Omega, \quad \mathcal{E}(\theta) = f_{X,Y}(\phi(\theta)) \tag{2}$$

where the **data fidelity** $f_{X,Y}$ depends on the data $X := (x_i)_i$, $Y := (y_i)_i$ and the loss ℓ, while the **mapping** ϕ must be independent from these quantities. Formally, Ω is a non-empty open subset of the domain of trainable parameters, $\mathbb{R}^D$ (introduced to capture the local training dynamics) and $\phi \in \mathcal{C}^\infty(\Omega, \mathbb{R}^d)$.

Example 2.1. (Factorization for *linear* neural networks) In the two-layer case, with r neurons, denoting $\theta = (U, V) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r}$ (so that $D = (n+m)r$), we can factorize $g_\theta(x) := UV^\top x$ by the mapping $\phi(\theta) := UV^\top$ using $f_{X,Y}(\cdot) = \sum_i \ell(\cdot x_i, y_i)$. More generally for q layers, with $\theta = (U_1, \cdots, U_q)$, we can still factorize $g_\theta(x) := U_1 \cdots U_q x$ using $\phi(\theta) := U_1 \cdots U_q$ and the same $f_{X,Y}$. This factorization is *globally* valid on $\Omega = \mathbb{R}^D$ in the sense that $f_{X,Y}$ does not depend on θ.

The notion of locality of the factorization $f_{X,Y} \circ \phi$ is illustrated by the next example.

Example 2.2 (Factorization for two-layer ReLU network without bias). Consider $g_\theta(x) = \left(\sum_{j=1}^r u_{k,j} \sigma(\langle v_j, x \rangle) \right)_{k=1}^n$, with $\sigma(t) := \max(t, 0)$ the ReLU activation function and $v_j \in \mathbb{R}^m$, $u_{k,j} \in \mathbb{R}$. Then, denoting again $\theta = (U, V)$ with $U = (u_{k,j})_{k,j} =: (u_1, \cdots, u_r) \in \mathbb{R}^{n \times r}$ and $V = (v_1, \cdots, v_r) \in \mathbb{R}^{m \times r}$ (so that $D = (n+m)r$), we rewrite $g_\theta(x) = \sum_{j=1}^r u_j \varepsilon_{j,x} v_j^\top x$ where $\varepsilon_{j,x} = \mathbb{1}(v_j^\top x > 0)$ is piecewise constant with respect to θ. Thus, on any domain $\Omega \subset \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r}$ such that $\varepsilon_{j,x_i}(\theta) := \mathbb{1}(v_j^\top x_i > 0)$ is constant over $\theta \in \Omega$ for each training sample x_i, the model $g_\theta(x)$ can be factorized by the mapping $\phi(\theta) = (\phi_{jkl})_{jkl} := (u_j v_j^\top)_{j=1}^r \in \mathbb{R}^{r \times n \times m}$ (here $d = rmn$) using $f_{X,Y}(\phi) := \sum_i \ell(\sum_{j,k,l} \varepsilon_{j,x_i} \phi_{j,k,l}, y_i)$. On Ω we obtain a factorizing mapping $\phi(\theta)$ containing r matrices of size $m \times n$ (of rank at most one) associated to a "local" data-fidelity $f_{X,Y}$ valid in a neighborhood of θ. A similar factorization is possible for deeper ReLU networks, including with biases [24], as further discussed in the proof of Theorem 2.8 in Appendix B.

A priori, one can consider different "levels" of conservation, depending whether h is conserved:

    i. during the optimization of $\mathcal{E}$ for a given loss ℓ and a given data set $(x_i, y_i)_i$; i.e. during the optimization of $f_{X,Y} \circ \phi$, for a given $f_{X,Y}$;

    ii. given a loss ℓ, during the optimization of $\mathcal{E}$ for *any* data set; i.e., during the optimization of $f_{X,Y} \circ \phi$, for *every* data set (X, Y);

    iii. during the optimization of $f \circ \phi$ for *any choice* of smooth f (not necessarily associated to a data set (X, Y)).

Our analysis focuses on last two cases, and shows that under certain assumptions, being conserved *for a given loss and every dataset* is indeed equivalent to being conserved *for every smooth f*. As a consequence, our theoretical analysis then studies functions $h(\theta)$ preserved by flows of functions $f \circ \phi$ for a fixed mapping ϕ but any choice of fidelity f. We call these functions "conservation laws" associated to the mapping ϕ, and they are formally defined in Section 2.3. Theorem 2.8 shows that in the two examples given above, this is equivalent to the conservation for all cost $\mathcal{E}$ of the form (1).

2.2 Gradient dynamics

We consider training using the gradient flow (the continuous time limit of gradient descent) of $f \circ \phi$:

$$\dot{\theta}(t) = -\nabla(f \circ \phi)(\theta(t)) = -[\partial \phi(\theta(t))]^\top \nabla f(\phi(\theta(t))), \text{ with } \theta(0) = \theta_{\text{init}}, \tag{3}$$

where the "data fidelity function" f is differentiable and arises from (2) with some dataset $(x_i, y_i)_i$ and some loss function ℓ. Here $\partial \phi(\theta) \in \mathbb{R}^{d \times D}$ is the Jacobian of the factorizing mapping. Note that using stochastic optimization methods and discrete gradients would break the exact preservation of the conservation laws, and only approximate conservation would hold, as remarked in [14].

The core of our analysis is to analyze the algebraic structure of the Jacobian vector fields involved in (3). In practice, the dimensions often satisfy $\text{rank} \partial \phi(\theta) < \min(d, D)$, i.e., $\phi(\theta)$ lives in a manifold of lower dimension. This corresponds to the fact that θ is an over-parameterized variable, and ϕ is an over-parameterized model with nontrivial conserved quantities during the optimization. Our goal is to determine the "number" of independent functions conserved through *all* such flows (i.e. for *every choice* of data fidelity function f *restricted to* the form (2)). We show in Section 2.3 that, under mild assumptions on the loss ℓ, these conserved functions are exactly the functions conserved through *all flows* (3) *for every infinitely smooth* data fidelity function $f \in \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$.

Example 2.3. As a first simple example, consider a two-layer *linear* neural network in dimension 1 (both for the input and output), with a single neuron. For such – admittedly trivial – architecture, the function to minimize is factorized by the mapping $\phi : (u \in \mathbb{R}, v \in \mathbb{R}) \mapsto uv \in \mathbb{R}$ with $\theta := (u, v)$. One can directly check that the function: $h(u, v) = u^2 - v^2$ satisfies that, for all initial conditions $(u_{\text{init}}, v_{\text{init}}) \in \mathbb{R}^2$, $h(u(t), v(t)) = h(u_{\text{init}}, v_{\text{init}})$, as soon as $\theta(t) := (u(t), v(t))$ is a solution of the ODE (3) with *some* differentiable data-fidelity function f. We say in that case that h is a conservation law for ϕ. Are there other such functions? Example 3.6 explains that on this example the answer is negative. This results from algebraic computations, implemented in SageMath, see Section 3.3.

2.3 Conserved functions and Conservation laws

We define conserved functions associated with (collections of) vector fields in $\mathcal{X}(\Omega) := \mathcal{C}^\infty(\Omega, \mathbb{R}^D)$.

Definition 2.4. Let $\chi \in \mathcal{X}(\Omega)$ be an infinitely smooth vector field. By the Cauchy-Lipschitz theorem, for each initial condition θ_{init}, there exists a unique maximal solution $t \in [0, T_{\theta_{\text{init}}}) \mapsto \theta(t, \theta_{\text{init}})$ of the ODE $\dot{\theta}(t) = \chi(\theta(t))$ with $\theta(0) = \theta_{\text{init}}$. A function $h : \Omega \subseteq \mathbb{R}^D \to \mathbb{R}$ is *conserved during the flow induced by* χ if $h(\theta(t, \theta_{\text{init}})) = h(\theta_{\text{init}})$ for each choice of θ_{init} and every $t \in [0, T_{\theta_{\text{init}}})$.

It is *conserved through a subset* $V \subset \mathcal{X}(\Omega)$ if h is conserved during all flows induced by all $\chi \in V$.

A basic property of $\mathcal{C}^1$ conserved functions (which proof can be found in Appendix A) corresponds to an "orthogonality" between their gradient and the considered vector fields.

Proposition 2.5. *Given a subset $V \subset \mathcal{X}(\Omega)$, consider its* trace *at $\theta \in \Omega$, defined as the linear space*

$$V(\theta) := \text{span}\{\chi(\theta) : \chi \in V\} \subseteq \mathbb{R}^D. \tag{4}$$

A function $h \in \mathcal{C}^1(\Omega, \mathbb{R})$ is conserved through V if, and only if, $\nabla h(\theta) \perp V(\theta)$ for every $\theta \in \Omega$.

Given a family $F \subseteq \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$ of data-fidelity functions, the set of functions that are conserved during all flows defined by the ODE (3), with each $f \in F$, corresponds by definition to the functions that are conserved through the subset

$$V_\phi[F] := \{\chi : \exists f \in F, \ \chi = \nabla(f \circ \phi) \text{ on } \Omega\}. \tag{5}$$

Given a loss ℓ, our goal is to study the functions conserved through $V_\phi[F_\ell]$, where F_ℓ collects all smooth data-fidelity functions $f \in \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$ that satisfy $(f \circ \phi)(\theta) = \sum_{i=1}^N \ell(g_\theta(x_i), y_i)$ for some training set of arbitrary size, i.e.

$$F_\ell := \left\{ f \in \mathcal{C}^\infty(\phi(\Omega), \mathbb{R}) : \exists (X, Y), f \circ \phi(\theta) = f_{X,Y} \circ \phi(\theta) := \sum_{i=1}^N \ell(g_\theta(x_i), y_i) \text{ on } \Omega \right\}. \tag{6}$$

For linear and ReLU networks we show in Theorem 2.8 and Proposition 2.9 that:

1. under (mild) assumptions on the loss $\ell(\cdot, \cdot)$, being conserved through $V_\phi[F_\ell]$ is the same as being conserved through $V_\phi[\mathcal{C}^\infty] := V_\phi[\mathcal{C}^\infty(\phi(\Omega), \mathbb{R})]$, i.e. through *any infinitely smooth data-fidelity*;

2. being conserved through the (a priori infinite-dimensional) subspace $V_\phi[\mathcal{C}^\infty]$ is in turn equivalent to being conserved through the *finite-dimensional* subspace

$$V_\phi := \mathrm{span}\{\nabla\phi_1(\cdot), \cdots, \nabla\phi_d(\cdot)\} = \left\{\theta \mapsto \sum_i a_i \nabla\phi_i(\theta) : (a_1, \ldots, a_d) \in \mathbb{R}^d\right\} \qquad (7)$$

where we write $\partial\phi(\theta)^\top = (\nabla\phi_1(\theta), \cdots, \nabla\phi_d(\theta)) \in \mathbb{R}^{D\times d}$, with $\nabla\phi_i \in \mathcal{X}(\Omega)$.

The first point (that we establish below with Theorem 2.8) motivates the following definition

Definition 2.6. A real-valued function h is a *conservation law of ϕ* if it is conserved through $V_\phi[\mathcal{C}^\infty]$.

Proposition 2.5 yields the following intermediate result.

Proposition 2.7. $h \in \mathcal{C}^1(\Omega, \mathbb{R})$ *is a conservation law of ϕ iff* $\nabla h(\theta) \perp V_\phi[\mathcal{C}^\infty](\theta), \forall\, \theta \in \Omega$.

The following theorem (which proof can be found in Appendix B) establishes that in some cases, the functions conserved through $V_\phi[F_\ell]$ are exactly the conservation laws of ϕ.

Theorem 2.8. *Assume that the loss* $(z, y) \mapsto \ell(z, y)$ *satisfies the condition:*

$$\mathrm{span}_{y\in\mathcal{Y}}\{\nabla_z\ell(z, y)\} = \mathbb{R}^n, \forall z \in \mathbb{R}^n, \qquad (8)$$

*then for linear neural networks, the conservation laws of ϕ are **exactly** the conserved functions through $V_\phi[F_\ell]$, with ϕ from Example 2.1. The same result holds for two-layer ReLU networks with ϕ from Example 2.2 under an additional hypothesis on Ω: the parameter θ of the network is such that hidden neurons are associated to pairwise distinct "hyperplanes" (cf Appendix B for details).*

Condition (8) holds for classical losses ℓ (e.g. quadratic/logistic losses), as shown in Lemma B.5 in Appendix B. Note that the additional hypothesis of pairwise distinct hyperplanes for the two-layer ReLU case is a generic hypothesis and is usual (see e.g. the notion of twin neurons in [24]). The tools from Appendix B extend Theorem 2.8 beyond (deep) linear and shallow ReLU networks. An open problem is whether Theorem 2.8 still holds for deep ReLU networks.

For the second point (the link between conservation through $V_\phi[\mathcal{C}^\infty]$ and through V_ϕ), an apparent difficulty is that the space $V_\phi[\mathcal{C}^\infty]$ of all gradient fields is a priori infinite-dimensional. In contrast, the space V_ϕ defined in (7) introduces a much simpler *finite-dimensional* proxy. A cornerstone of our analysis is to show that the study of conservation laws boils down to the study of this finite-dimensional vector space. This will be crucial in Section 4.1, to provide a tractable scheme (i.e. operating in finite dimension) to analyze the algebraic relationship induced by these vector fields. By combining Proposition 2.7 with the observation that for all $\theta \in \Omega$ we have $V_\phi[\mathcal{C}^\infty](\theta) = \mathrm{span}\{\nabla\phi_1(\theta), \ldots, \nabla\phi_d(\theta)\} = \mathrm{range}(\partial\phi(\theta)^\top) = V_\phi(\theta)$, we obtain:

Proposition 2.9. $h \in \mathcal{C}^1(\Omega, \mathbb{R})$ *is a conservation law for ϕ (cf Definition 2.6) if and only if it is conserved though the finite-dimensional space V_ϕ defined in* (7)*, i.e. if*

$$\nabla h(\theta) \perp \nabla\phi_j(\theta), \ \forall\, \theta \in \Omega, \ \forall j \in \{1, \ldots, d\}.$$

Example 2.10. Revisiting Example 2.3, with $\phi : (u \in \mathbb{R}, v \in \mathbb{R}) \mapsto uv$ and $\theta := (u, v)$, we saw that $h((u, v)) := u^2 - v^2$ is conserved: and indeed $\langle\nabla h(u, v), \nabla\phi(u, v)\rangle = 2uv - 2vu = 0, \forall(u, v)$.

In this simple example, the characterization of Proposition 2.9 gives a *constructive* way to find such a conserved function: we only need to find a function h such that $\langle\nabla h(u, v), \nabla\phi(u, v)\rangle = \langle\nabla h(u, v), (v, u)^\top\rangle = 0$. The situation becomes more complex in higher dimensions, since one needs to understand the interplay between the different vector fields in V_ϕ.

2.4 Constructibility of some conservation laws

Observe that in Example 2.10 both the mapping ϕ and the conservation law h are polynomials, a property that surprisingly systematically holds in all examples of interest in the paper, making it possible to *algorithmically* construct some conservation laws as detailed now.

By Proposition 2.9, a function h is conserved if it is in the kernel of the linear operator $h \in \mathcal{C}^1(\Omega, \mathbb{R}) \mapsto (\theta \in \Omega \mapsto (\langle\nabla h(\theta), \nabla\phi_i(\theta)\rangle)_{i=1,\cdots,d})$. Thus, one could look for conservation laws in a prescribed finite-dimensional space by projecting these equations in a basis (as in finite-element

methods for PDEs). Choosing the finite-dimensional subspace could be generally tricky, but for the linear and ReLU cases all known conservation laws are actually polynomial "balancedness-type conditions" [1, 2, 8], see Section 4. In these cases, the vector fields in V_ϕ are also polynomials (because ϕ is polynomial, see Theorem B.4 and Lemma B.7 in Appendix B), hence $\theta \mapsto \langle \nabla h(\theta), \nabla \phi_i(\theta) \rangle$ is a polynomial too. This allows us to compute a basis of independent polynomial conservation laws of a given degree (to be freely chosen) for these cases, by simply focusing on the corresponding subspace of polynomials. We coded the resulting equations in SageMath, and we found back on selected examples (see Appendix I) all existing known conservation laws both for ReLU and linear networks. Open-source code is available at `https://github.com/sibyllema/Conservation_laws`.

2.5 Independent conserved functions

Having an algorithm to build conservation laws is nice, yet how can we know if we have built "all" laws? This requires first defining a notion of a "maximal" set of functions, which would in some sense be independent. This does not correspond to linear independence of the functions themselves (for instance, if h is a conservation law, then so is h^k for each $k \in \mathbb{N}$ but this does not add any other constraint), but rather to pointwise linear independence of their gradients. This notion of independence is closely related to the notion of "functional independence" studied in [6, 17]. For instance, it is shown in [17] that smooth functionally dependent functions are characterized by having dependent gradients everywhere. This motivates the following definition.

Definition 2.11. A family of N functions $(h_1, \cdots, h_N)$ conserved through $V \subset \mathcal{X}(\Omega)$ is said to be *independent* if the vectors $(\nabla h_1(\theta), \cdots, \nabla h_N(\theta))$ are linearly independent for all $\theta \in \Omega$.

The goal is thus to find the largest set of independent conserved functions. An immediate upper bound holds on the number N of functionally independent functions $h_1, \ldots, h_N$ conserved through V: for $\theta \in \Omega \subseteq \mathbb{R}^D$, the space $W(\theta) := \text{span}\{\nabla h_1(\theta), \ldots, \nabla h_N(\theta)\} \subseteq \mathbb{R}^D$ is of dimension N (by independence) and (by Proposition 2.9) orthogonal to $V_\phi(\theta)$. Thus, it is necessary to have $N \le D - \dim V(\theta)$. As we will now see, this bound can be tight *under additional assumptions on V related to Lie brackets* (corresponding to the so-called Frobenius theorem). This will in turn lead to a characterization of the maximum possible N.

3 Conservation Laws using Lie Algebra

The study of hyper-surfaces trapping the solution of ODEs is a recurring theme in control theory, since the existence of such surfaces is the basic obstruction of controllability of such systems [5]. The basic result to study these surfaces is the so-called Frobenius theorem from differential calculus (See Section 1.4 of [11] for a good reference for this theorem). It relates the existence of such surfaces, and their dimensions, to some differential condition involving so-called "Lie brackets" $[u, v]$ between pairs of vector fields (see Section 3.1 below for a more detailed exposition of this operation). However, in most cases of practical interest (such as for instance matrix factorization), the Frobenius theorem is not suitable for a direct application to the space V_ϕ because its Lie bracket condition is not satisfied. To identify the number of independent conservation laws, one needs to consider the algebraic closure of V_ϕ under Lie brackets. The fundamental object of interest is thus the Lie algebra generated by the Jacobian vector fields, that we recall next.

Notations Given a vector subspace of infinitely smooth vector fields $V \subseteq \mathcal{X}(\Omega) := \mathcal{C}^\infty(\Omega, \mathbb{R}^D)$, we recall (cf Proposition 2.5) that its trace at some θ is the subspace

$$V(\theta) := \text{span}\{\chi(\theta) : \chi \in V\} \subseteq \mathbb{R}^D. \tag{9}$$

For each open subset $\Omega' \subseteq \Omega$, we introduce the subspace of $\mathcal{X}(\Omega')$: $V_{|\Omega'} := \{\chi_{|\Omega'} : \chi \in V\}$.

3.1 Background on Lie algebra

A Lie algebra A is a vector space endowed with a bilinear map $[\cdot, \cdot]$, called a Lie bracket, that verifies for all $X, Y, Z \in A$: $[X, X] = 0$ and the Jacobi identity: $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$.

For the purpose of this article, the Lie algebra of interest is the set of infinitely smooth vector fields $\mathcal{X}(\Omega) := \mathcal{C}^\infty(\Omega, \mathbb{R}^D)$, endowed with the Lie bracket $[\cdot, \cdot]$ defined by

$$[\chi_1, \chi_2] : \quad \theta \in \Omega \mapsto [\chi_1, \chi_2](\theta) := \partial\chi_1(\theta)\chi_2(\theta) - \partial\chi_2(\theta)\chi_1(\theta), \tag{10}$$

with $\partial \chi(\theta) \in \mathbb{R}^{D \times D}$ the jacobian of χ at θ. The space $\mathbb{R}^{n \times n}$ of matrices is also a Lie algebra endowed with the Lie bracket $[A, B] := AB - BA$. This can be seen as a special case of (10) in the case of *linear* vector fields, i.e. $\chi(\theta) = A\theta$.

Generated Lie algebra Let A be a Lie algebra and let $V \subset A$ be a vector subspace of A. There exists a smallest Lie algebra that contains V. It is denoted $\mathrm{Lie}(V)$ and called the generated Lie algebra of V. The following proposition [5, Definition 20] constructively characterizes $\mathrm{Lie}(V)$, where for vector subspaces $[V, V'] := \{[\chi_1, \chi_2] : \chi_1 \in V, \chi_2 \in V'\}$, and $V + V' = \{\chi_1 + \chi_2 : \chi_1 \in V, \chi_2 \in V'\}$.

Proposition 3.1. *Given any vector subspace $V \subseteq A$ we have $\mathrm{Lie}(V) = \bigcup_k V_k$ where:*
$$\begin{cases} V_0 & := V \\ V_k & := V_{k-1} + [V_0, V_{k-1}] \ for \ k \geq 1. \end{cases}$$

We will see in Section 3.2 that the number of conservation laws is characterized by the dimension of the trace $\mathrm{Lie}(V_\phi)(\theta)$ defined in (9). The following lemma (proved in Appendix C) gives a stopping criterion to algorithmically determine this dimension (see Section 3.3 for the algorithm).

Lemma 3.2. *Given $\theta \in \mathbb{R}^D$, if for a given i, $\dim V_{i+1}(\theta') = \dim V_i(\theta)$ for every θ' in a neighborhood of θ, then there exists a neighborhood Ω of θ such that $V_k(\theta') = V_i(\theta')$ for all $\theta' \in \Omega$ and $k \geq i$, where the V_i are defined by Proposition 3.1. Thus $\mathrm{Lie}(V)(\theta') = V_i(\theta')$ for all $\theta' \in \Omega$. In particular, the dimension of the trace of $\mathrm{Lie}(V)$ is locally constant and equal to the dimension of $V_i(\theta)$.*

3.2 Number of conservation laws

The following theorem uses the Lie algebra generated by V_ϕ to characterize precisely the number of conservation laws. The proof of this result is based on two successive uses of the Frobenius theorem and can be found in Appendix D (where we also recall Frobenius theorem for the sake of completeness).

Theorem 3.3. *If $\dim\mathrm{Lie}(V_\phi)(\theta)$ is locally constant then each $\theta \in \Omega$ admits a neighborhood Ω' such that there are $D - \dim\mathrm{Lie}(V_\phi)(\theta)$ (and no more) independent conserved functions through $V_{\phi|\Omega'}$.*

Combining Proposition 2.9 and Theorem 3.3 we obtain:

Corollary 3.4. *If $\dim\mathrm{Lie}(V_\phi)(\theta)$ is locally constant then each $\theta \in \Omega$ admits a neighborhood Ω' such that there are $D - \dim\mathrm{Lie}(V_\phi)(\theta)$ (and no more) independent conservation laws of ϕ on Ω'.*

Remark 3.5. The proof of the Frobenius theorem (and therefore of our generalization Theorem 3.3) is actually constructive. From a given ϕ, conservation laws are obtained in the proof by integrating in time (*i.e.* solving an advection equation) the vector fields belonging to V_ϕ. Unfortunately, this cannot be achieved in *closed form* in general, but in small dimensions, this could be carried out numerically (to compute approximate discretized laws on a grid or approximate them using parametric functions such as Fourier expansions or neural networks).

A fundamental aspect of Corollary 3.4 is to rely only on the computation of the *dimension of the trace* of the Lie algebra associated with the finite-dimensional vector space V_ϕ. Yet, even if V_ϕ is finite-dimensional, it might be the case that $\mathrm{Lie}(V_\phi)$ itself remains infinite-dimensional. Nevertheless, what matters is not the dimension of $\mathrm{Lie}(V_\phi)$, but that of *its trace* $\mathrm{Lie}(V_\phi)(\theta)$, which is *always* finite (and potentially much smaller that $\dim\mathrm{Lie}(V_\phi)$ even when the latter is finite) and computationally tractable thanks to Lemma 3.2 as detailed in Section 3.3. In section 4.1 we work out the example of matrix factorization, a non-trivial case where the full Lie algebra $\mathrm{Lie}(V_\phi)$ itself remains finite-dimensional.

Corollary 3.4 requires that the dimension of the trace at θ of the Lie algebra is locally constant. This is a technical assumption, which typically holds outside a set of pathological points. A good example is once again matrix factorization, where we show in Section 4.1 that this condition holds generically.

3.3 Method and algorithm, with examples

Given a factorizing mapping ϕ for the architectures to train, to determine the number of independent conservation laws of ϕ, we leverage the characterization 3.1 to algorithmically compute

$\mathrm{dimLie}(V_\phi)(\theta)$ using an iterative construction of bases for the subspaces V_k starting from $V_0 := V_\phi$, and stopping as soon as the dimension stagnates thanks to Lemma 3.2. Our open-sourced code is available at https://github.com/sibyllema/Conservation_laws and uses SageMath. As we now show, this algorithmic principle allows to fully work out certain settings where the stopping criterion of Lemma 3.2 is reached at the first step ($i = 0$) or the second one ($i = 1$). Section 4.2 also discusses its numerical use for an empirical investigation of broader settings.

Example where the iterations of Lemma 3.2 stop at the first step.   This corresponds to the case where $\mathrm{Lie}V_\phi(\theta) = V_1(\theta) = V_0(\theta) := V_\phi(\theta)$ on Ω. This is the case if and only if V_ϕ satisfies that

$$[\chi_1, \chi_2](\theta) := \partial\chi_1(\theta)\chi_2(\theta) - \partial\chi_2(\theta)\chi_1(\theta) \in V_\phi(\theta), \qquad \text{for all } \chi_1, \chi_2 \in V_\phi \text{ and all } \theta \in \Omega. \quad (11)$$

i.e., when the Frobenius Theorem (see Theorem D.1 in Appendix D) applies directly. The first example is a follow-up to Example 2.2.

Example 3.6 (two-layer ReLU networks without bias). Consider $\theta = (U, V)$ with $U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}$, $n, m, r \geq 1$ (so that $D = (n + m)r$), and the mapping $\phi(\theta) := (u_i v_i^\top)_{i=1,\cdots,r} \in \mathbb{R}^{n \times m \times r}$,, where $U = (u_1; \cdots; u_r)$ and $V = (v_1; \cdots; v_r)$, As detailed in Appendix E.1, since $\phi(\theta)$ is a collection of r rank-one $n \times m$ matrices, $\dim V_\phi(\theta) = \mathrm{rank}\partial\phi(\theta) = (n + m - 1)r$ is constant on the domain Ω such that $u_i, v_j \neq 0$, and V_ϕ satisfies (11), hence by Corollary 3.4 each θ has a neighborhood Ω' such that there exists r (and no more) independent conserved function through $V_{\phi|\Omega'}$. The r known conserved functions [8] given by $h_i : (U, V) \mapsto \|u_i\|^2 - \|v_i\|^2$, $i = 1, \cdots, r$, are independent, hence they are complete.

Example where the iterations of Lemma 3.2 stop at the second step (but not the first one).   Our primary example is matrix factorization, as a follow-up to Example 2.1.

Example 3.7 (two-layer *linear* neural networks). With $\theta = (U, V)$, where $(U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r})$ as in Example 3.6, the mapping $\phi(\theta) := UV^\top \in \mathbb{R}^{n \times m}$, (here $d = nm$) factorizes the functions minimized during the training of linear neural networks with two layers (see Example 2.1). As shown in Appendix H, condition (11) is not satisfied when $r > 1$ and $\max(n, m) > 1$. Thus, the stopping criterion of Lemma 3.2 is not satisfied at the first step. However, as detailed in Proposition G.3 in Appendix G, $(V_\phi)_1 = (V_\phi)_2 = \mathrm{Lie}V_\phi$, hence the iterations of Lemma 3.2 stop at the second step.

We complete this example in the next section by showing that known conservation laws are indeed complete (see Corollary 4.4). Whether known conservation laws remain valid and/or *complete* in this settings and extended ones is further studied in Section 4 and Appendix E using the toolset that we have presented.

3.4   Application: recasting over-parameterized flows as low-dimensional Riemannian flows

One striking application of Corollary 3.4 (in simple cases where $\dim V_\phi(\theta) = \mathrm{dimLie}V_\phi(\theta)$ is constant on Ω, i.e., $\mathrm{rank}\partial\phi(\theta)$ is constant on Ω and V_ϕ satisfies (11)) is to fully rewrite the high-dimensional flow $\theta(t) \in \mathbb{R}^D$ as a low-dimensional flow $z(t) \in \mathbb{R}^d$, where this flow is associated with a Riemannian metric tensor M that is induced by ϕ and depends on the initialization θ_{init}. We insist on the fact that this is only possible in very specific cases, but this phenomenon is underlying many existing works which aim at writing in closed form the implicit bias associated with some training dynamics (see Section 1 for some relevant literature. Our analysis shed some light on cases where this is possible (see Appendix F for a proof). Note that the metric $M(z, \theta_{\mathrm{init}})$ can have a kernel, typically when $\phi(\Omega)$ is a sub-manifold. The evolution (12) should then be understood as a flow on this manifold. The kernel of $M(z, \theta_{\mathrm{init}})$ is orthogonal to the tangent space at z of this manifold.

Proposition 3.8. *Assume that* $\mathrm{rank}(\partial\phi(\theta))$ *is constant on* Ω *and that* V_ϕ *satisfies* (11). *If* $\theta(t) \in \mathbb{R}^D$ *satisfies the ODE* (3) *where* $\theta_{init} \in \Omega$, *then there is* $0 < T^\star_{\theta_{\mathrm{init}}} \leq T_{\theta_{\mathrm{init}}}$ *such that* $z(t) := \phi(\theta(t)) \in \mathbb{R}^d$ *satisfies the ODE*

$$\dot{z}(t) = -M(z(t), \theta_{init})\nabla f(z(t)) \quad \text{for all } t \in [0, T^\star_{\theta_{\mathrm{init}}}), \text{ with } z(0) = \phi(\theta_{init}), \qquad (12)$$

where $M(z(t), \theta_{init}) \in \mathbb{R}^{d \times d}$ *is a symmetric semi-definite matrix.*

Revisiting Example 3.6 leads to the following analytic example.

Example 3.9. Given the mapping $\phi : (u \in \mathbb{R}^*, v \in \mathbb{R}^d) \mapsto uv \in \mathbb{R}^d$, the variable $z := uv$ satisfies (12) with $M(z, \theta_{\text{init}}) = \|z\|_\delta I_d + \|z\|_\delta^{-1} zz^\top$, with $\|z\|_\delta := \delta + \sqrt{\delta^2 + \|z\|^2}$, $\delta := 1/2(u_{\text{init}}^2 - \|v_{\text{init}}\|^2)$.

Another analytic example is discussed in Appendix F. In light of these results, an interesting perspective is to better understand the dependance of the Riemannian metric with respect to initialization, to possibly guide the choice of initialization for better convergence dynamics.

4 Conservation Laws for Linear and ReLU Neural Networks

To showcase the impact of our results, we show how they can be used to determine whether known conservation laws for linear (resp. ReLU) neural networks are complete, and to recover these laws *algorithmically* using factorizing mappings ϕ adapted to these two settings. Concretely, we study the conservation laws for neural networks with q layers, and either a linear or ReLU activation, with an emphasis on $q = 2$. We write $\theta = (U_1, \cdots, U_q)$ with $U_i \in \mathbb{R}^{n_{i-1} \times n_i}$ the weight matrices and we assume that θ satisfies the gradient flow (3) for some data fidelity function $f \in \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$. In the linear case the mapping is $\phi_{\text{Lin}}(\theta) := U_1 \cdots U_q$. For ReLU networks, we use the (polynomial) mapping ϕ_{ReLu} of [24, Definition 6], which is defined for any (deep) feedforward ReLU network, with or without bias. In the simplified setting of networks without biases it reads explicitly as:

$$\phi_{\text{ReLu}}(U_1, \cdots, U_q) := \Big(U_1[:, j_1] U_2[j_1, j_2] \cdots U_{q-1}[j_{q-2}, j_{q-1}] U_q[j_{q-1}, :]\Big)_{j_1, \cdots, j_{q-1}} \tag{13}$$

with $U[i, j]$ the (i, j)-th entry of U. This covers $\phi(\theta) := (u_j v_j^\top)_{j=1}^r \in \mathbb{R}^{n \times m \times r}$ from Example 2.2.

Some conservation laws are known for the linear case ϕ_{Lin} [1, 2] and for the ReLu case ϕ_{ReLu} [8].

Proposition 4.1 ([1, 2, 8])**.** *If $\theta := (U_1, \cdots, U_q)$ satisfies the gradient flow (3), then for each $i = 1, \cdots, q-1$ the function $\theta \mapsto U_i^\top U_i - U_{i+1} U_{i+1}^\top$ (resp. the function $\theta \mapsto \text{diag}\left(U_i^\top U_i - U_{i+1} U_{i+1}^\top\right)$) defines $n_i \times (n_i + 1)/2$ conservation laws for ϕ_{Lin} (resp. n_i conservation laws for ϕ_{ReLu}).*

Proposition 4.1 defines $\sum_{i=1}^{q-1} n_i \times (n_i+1)/2$ conserved functions for the linear case. In general they are *not* independent, and we give below in Proposition 4.2, for the case of $q = 2$, the *exact* number of independent conservation laws among these particular laws. Establishing whether there are other (previously unknown) conservation laws is an open problem in the general case $q > 2$. We already answered negatively to this question in the two-layer ReLu case without bias (See Example 3.6). In the following Section (Corollary 4.4), we show the same result in the linear case $q = 2$. Numerical computations suggest this is still the case for deeper linear ReLU networks as detailed in Section 4.2.

4.1 The matrix factorization case ($q = 2$)

To simplify the analysis when $q = 2$, we rewrite $\theta = (U, V)$ as a vertical matrix concatenation denoted $(U; V) \in \mathbb{R}^{(n+m) \times r}$, and $\phi(\theta) = \phi_{\text{Lin}}(\theta) = UV^\top \in \mathbb{R}^{n \times m}$.

How many independent conserved functions are already known?   The following proposition refines Proposition 4.1 for $q = 2$ by detailing how many *independent* conservation laws are already known. See Appendix G.1 for a proof.

Proposition 4.2. *Consider $\Psi : \theta = (U; V) \mapsto U^\top U - V^\top V \in \mathbb{R}^{r \times r}$ and assume that $(U; V)$ has full rank noted rk. Then the function Ψ gives $\text{rk} \cdot (2r + 1 - \text{rk})/2$ independent conserved functions.*

There exist no more independent conserved functions.   We now come to the core of the analysis, which consists in actually computing $\text{Lie}(V_\phi)$ as well as its traces $\text{Lie}(V_\phi)(\theta)$ in the matrix factorization case. The crux of the analysis, which enables us to fully work out theoretically the case $q = 2$, is that V_ϕ is composed of *linear* vector fields (that are explicitly characterized in Proposition G.2 in Appendix G), the Lie bracket between two linear fields being itself linear and explicitly characterized with skew matrices, see Proposition G.3 in Appendix G. Eventually, what we need to compute is the dimension of the trace $\text{Lie}(V_\phi)(U, V)$ for any (U, V). We prove the following in Appendix G.

Proposition 4.3. *If $(U; V) \in \mathbb{R}^{(n+m) \times r}$ has full rank noted rk, then: $\dim \text{Lie}(V_\phi)(U; V) = (n + m)r - (2r + 1 - \text{rk})/2$.*

With this explicit characterization of the trace of the generated Lie algebra and Proposition 4.2, we conclude that Proposition 4.1 has indeed exhausted the list of independent conservation laws.

Corollary 4.4. *If $(U; V)$ has full rank, then all conserved functions are given by $\Psi : (U, V) \mapsto U^\top U - V^\top V$. In particular, there exist no more* independent *conserved functions.*

4.2 Numerical guarantees in the general case

The expressions derived in the previous section are specific to the linear case $q = 2$. For deeper linear networks and for ReLU networks, the vector fields in V_ϕ are non-linear polynomials, and computing Lie brackets of such fields can increase the degree, which could potentially make the generated Lie algebra infinite-dimensional. One can however use Lemma 3.2 and stop as soon as $\dim\left((V_\phi)_k(\theta)\right)$ stagnates. Numerically comparing this dimension with the number N of independent conserved functions known in the literature (predicted by Proposition 4.1) on a sample of depths/widths of small size, we empirically confirmed that there are no more conservation laws than the ones already known for deeper linear networks and for ReLU networks too (see Appendix I for details). Our code is open-sourced and is available at `https://github.com/sibyllema/Conservation_laws`. It is worth mentioning again that in all tested cases ϕ is polynomial, and there is a maximum set of conservation laws that are also polynomial, which are found algorithmically (as detailed in Section 2.4).

Conclusion

In this article, we proposed a constructive program for determining the number of conservation laws. An important avenue for future work is the consideration of more general classes of architectures, such as deep convolutional networks, normalization, and attention layers. Note that while we focus in this article on gradient flows, our theory can be applied to any space of displacements in place of V_ϕ. This could be used to study conservation laws for flows with higher order time derivatives, for instance gradient descent with momentum, by lifting the flow to a higher dimensional phase space. A limitation that warrants further study is that our theory is restricted to continuous time gradient flow. Gradient descent with finite step size, as opposed to continuous flows, disrupts exact conservation. The study of approximate conservation presents an interesting avenue for future work.

Acknowledgement

References

[1] S. ARORA, N. COHEN, N. GOLOWICH, AND W. HU, *A convergence analysis of gradient descent for deep linear neural networks*, arXiv preprint arXiv:1810.02281, (2018).

[2] S. ARORA, N. COHEN, AND E. HAZAN, *On the optimization of deep networks: Implicit acceleration by overparameterization*, in Int. Conf. on Machine Learning, PMLR, 2018, pp. 244–253.

[3] B. BAH, H. RAUHUT, U. TERSTIEGE, AND M. WESTDICKENBERG, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, Information and Inference: A Journal of the IMA, 11 (2022), pp. 307–353.

[4] M. BELKIN, D. HSU, S. MA, AND S. MANDAL, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proc. of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[5] B. BONNARD, M. CHYBA, AND J. ROUOT, *Geometric and Numerical Optimal Control - Application to Swimming at Low Reynolds Number and Magnetic Resonance Imaging*, Springer-Briefs in Mathematics, Springer Int. Publishing, 2018.

[6] A. B. BROWN, *Functional dependence*, Transactions of the American Mathematical Society, 38 (1935), pp. 379–394.

[7] L. CHIZAT AND F. BACH, *Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss*, in Conf. on Learning Theory, PMLR, 2020, pp. 1305–1338.

[8] S. S. DU, W. HU, AND J. D. LEE, *Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced*, Adv. in Neural Inf. Proc. Systems, 31 (2018).

[9] S. GUNASEKAR, J. LEE, D. SOUDRY, AND N. SREBRO, *Characterizing implicit bias in terms of optimization geometry*, in Int. Conf. on Machine Learning, PMLR, 2018, pp. 1832–1841.

[10] S. GUNASEKAR, B. E. WOODWORTH, S. BHOJANAPALLI, B. NEYSHABUR, AND N. SRE-BRO, *Implicit regularization in matrix factorization*, Adv. in Neural Inf. Proc. Systems, 30 (2017).

[11] A. ISIDORI, *Nonlinear system control*, New York: Springer Verlag, 61 (1995), pp. 225–236.

[12] Z. JI, M. DUDÍK, R. E. SCHAPIRE, AND M. TELGARSKY, *Gradient descent follows the regularization path for general losses*, in Conf. on Learning Theory, PMLR, 2020, pp. 2109–2136.

[13] Z. JI AND M. TELGARSKY, *Gradient descent aligns the layers of deep linear networks*, arXiv preprint arXiv:1810.02032, (2018).

[14] D. KUNIN, J. SAGASTUY-BRENA, S. GANGULI, D. L. YAMINS, AND H. TANAKA, *Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics*, arXiv preprint arXiv:2012.04728, (2020).

[15] Z. LI, T. WANG, J. LEE, AND S. ARORA, *Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent*, arXiv preprint arXiv:2207.04036, (2022).

[16] H. MIN, S. TARMOUN, R. VIDAL, AND E. MALLADA, *On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks*, in Int. Conf. on Machine Learning, PMLR, 2021, pp. 7760–7768.

[17] W. F. NEWNS, *Functional dependence*, The American Mathematical Monthly, 74 (1967), pp. 911–920.

[18] B. NEYSHABUR, *Implicit regularization in deep learning*, arXiv preprint arXiv:1709.01953, (2017).

[19] B. NEYSHABUR, R. TOMIOKA, AND N. SREBRO, *In search of the real inductive bias: On the role of implicit regularization in deep learning*, arXiv preprint arXiv:1412.6614, (2014).

[20] E. NOETHER, *Invariante variationsprobleme*, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, 1918 (1918), pp. 235–257.

[21] A. M. SAXE, J. L. MCCLELLAND, AND S. GANGULI, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, arXiv preprint arXiv:1312.6120, (2013).

[22] S. SHALEV-SHWARTZ AND S. BEN-DAVID, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[23] D. SOUDRY, E. HOFFER, M. S. NACSON, S. GUNASEKAR, AND N. SREBRO, *The implicit bias of gradient descent on separable data*, The Journal of Machine Learning Research, 19 (2018), pp. 2822–2878.

[24] P. STOCK AND R. GRIBONVAL, *An Embedding of ReLU Networks and an Analysis of their Identifiability*, Constructive Approximation, (2022). Publisher: Springer Verlag.

[25] S. TARMOUN, G. FRANCA, B. D. HAEFFELE, AND R. VIDAL, *Understanding the dynamics of gradient flow in overparameterized linear models*, in Int. Conf. on Machine Learning, PMLR, 2021, pp. 10153–10161.

[26] THE SAGE DEVELOPERS, *SageMath, the Sage Mathematics Software System (Version 9.7)*, 2022. `https://www.sagemath.org`.

[27] C. ZHANG, S. BENGIO, M. HARDT, B. RECHT, AND O. VINYALS, *Understanding deep learning requires rethinking generalization*, in Int. Conf. on Learning Representations, 2017.

[28] B. ZHAO, I. GANEV, R. WALTERS, R. YU, AND N. DEHMAMY, *Symmetries, flat minima, and the conserved quantities of gradient flow*, arXiv preprint arXiv:2210.17216, (2022).

A Proof of Proposition 2.5

Proposition 2.5 is a direct consequence of the following lemma (remember that $\nabla h(\theta) = [\partial h(\theta)]^\top$).

Lemma A.1 (Smooth functions conserved through a given flow.). *Given $\chi \in \mathcal{X}(\Omega)$, a function $h \in \mathcal{C}^1(\Omega, \mathbb{R})$ is conserved through the flow induced by χ if and only if $\partial h(\theta)\chi(\theta) = 0$ for all $\theta \in \Omega$.*

Proof. Assume that $\partial h(\theta)\chi(\theta) = 0$ for all $\theta \in \Omega$. Then for all $\theta_{\text{init}} \in \Omega$ and for all $t \in (0, T_{\theta_{\text{init}}})$:

$$\frac{\mathrm{d}}{\mathrm{d}t} h(\theta(t, \theta_{\text{init}})) = \partial h(\theta(t, \theta_{\text{init}}))\dot{\theta}(t, \theta_{\text{init}}) = \partial h(\theta(t, \theta_{\text{init}}))\chi(\theta(t, \theta_{\text{init}})) = 0.$$

Thus: $h(\theta(t, \theta_{\text{init}})) = h(\theta_{\text{init}})$, *i.e.*, h is conserved through χ. Conversely, assume that there exists $\theta_0 \in \Omega$ such that $\partial h(\theta_0)\chi(\theta_0) \neq 0$. Then by continuity of $\theta \in \Omega \mapsto \partial h(\theta)\chi(\theta)$, there exists $r > 0$ such that $\partial h(\theta)\chi(\theta) \neq 0$ on $B(\theta_0, r)$. With $\theta_{\text{init}} = \theta_0$ by continuity of $t \mapsto \theta(t, \theta_{\text{init}})$, there exists $\varepsilon > 0$, such that for all $t < \varepsilon$, $\theta(t, \theta_{\text{init}}) \in B(\theta_0, r)$. Then for all $t \in (0, \varepsilon)$:$\frac{\mathrm{d}}{\mathrm{d}t} h(\theta(t, \theta_{\text{init}})) = \partial h(\theta(t, \theta_{\text{init}}))\chi(\theta(t, \theta_{\text{init}})) \neq 0$, hence h is not conserved through the flow induced by χ. $\square$

B Proof of Theorem 2.8

Considering a family of functions $F \subset \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$ we recall the notation (cf (5))

$$V_\phi[F] := \{\chi : \exists f \in F, \ \chi = \nabla(f \circ \phi) \text{ on } \Omega\}. \tag{14}$$

Given a loss function ℓ we also recall our primary interest in the family

$$F_\ell := \left\{ f \in \mathcal{C}^\infty(\phi(\Omega), \mathbb{R}) : \exists N \in \mathbb{N}, \exists (x_i, y_i)_{i=1}^N, \forall \theta \in \Omega, (f \circ \phi)(\theta) = \sum_{i=1}^N \ell(g_\theta(x_i), y_i) \right\}. \tag{15}$$

We note $\mathrm{CL}(V_\phi[\mathcal{C}^\infty])$ (resp $\mathrm{CL}(V_\phi[F_\ell])$ the set of all functions conserved through $V_\phi[\mathcal{C}^\infty]$ (resp. $V_\phi[F_\ell]$)). Our goal is to show that $\mathrm{CL}(V_\phi[\mathcal{C}^\infty]) = \mathrm{CL}(V_\phi[F_\ell])$ under the assumptions of the theorem. For this we will show below that linear and ReLU networks satisfy the following intermediate assumption:

Assumption B.1. The model $(x, \theta) \in X \times \Omega \mapsto g_\theta(x) \in \mathbb{R}^n$, where $X \subseteq \mathbb{R}^m$, can be factorized via ϕ: for every $x \in X$, there is $L^x \in \mathcal{C}^\infty(\mathbb{R}^{d'}, \mathbb{R}^n)$ such that $g_\theta(x) = L^x(\phi(\theta))$ for every $\theta \in \Omega$.

For linear networks, with $\theta := (U_1, \ldots, U_q)$ and $\phi(\theta) := U_1 \ldots U_q$, since $g_\theta(x) = U_1 \ldots U_q x$ the assumption indeed holds with $X = \mathbb{R}^m$, any Ω, and $L^x(\phi(\theta)) := \phi(\theta)x$. The assumption also holds for shallow ReLU networks without bias when Ω is a sufficiently small neighborhood of some reference parameter θ and X is chosen adequately, cf Example 2.2. As a consequence of the results of [24] this extends to deeper ReLU networks, including with biases, as will be soon detailed.

As a first step, we establish the following consequence of Assumption B.1.

Proposition B.2. *Consider a loss function $\ell(z, y)$ that is differentiable with respect to $z \in \mathbb{R}^n$. Under Assumption B.1 (and with the corresponding notations), if for all $\theta \in \Omega$ we have*

$$\operatorname*{span}_{x \in X, y} \left\{ [\partial L^x(\phi(\theta))]^\top \nabla_z \ell(g_\theta(x), y)) \right\} = \mathbb{R}^d \tag{16}$$

then $\mathrm{CL}(V_\phi[\mathcal{C}^\infty]) = \mathrm{CL}(V_\phi[F_\ell])$.

Proof. We will show that $V_\phi[F_\ell](\theta) = V_\phi[\mathcal{C}^\infty](\theta)$ for all $\theta \in \Omega$. By Proposition 2.5, it will follow that $\mathrm{CL}(V_\phi[\mathcal{C}^\infty]) = \mathrm{CL}(V_\phi[F_\ell])$ as claimed.

By definition, $F_\ell \subseteq \mathcal{C}^\infty(\phi(\Omega), \mathbb{R})$, we have $V_\phi[F_\ell] \subseteq V_\phi[\mathcal{C}^\infty(\phi(\Omega), \mathbb{R})] =: V_\phi[\mathcal{C}^\infty]$ (cf (5)) hence $V_\phi[F_\ell](\theta) \subseteq V_\phi[\mathcal{C}^\infty](\theta)$ for every $\theta \in \Omega$. There only remains to prove the converse inclusion. For this, consider an arbitrary $\theta \in \Omega$. By the definition (5) of $V_\phi[\mathcal{C}^\infty]$ and of its trace (cf Proposition 2.5), one can check that $V_\phi[\mathcal{C}^\infty](\theta) = \mathrm{range}\left(\partial\phi(\theta)^\top\right)$. Thus we simply need to show that $\mathrm{range}\left(\partial\phi(\theta)^\top\right) \subseteq V_\phi[F_\ell](\theta)$. By Assumption B.1 for each $x \in X$ there is a $\mathcal{C}^\infty$ function L^x such

that $g_\theta(x) = L^x(\phi(\theta))$ for every $\theta \in \Omega$. For each "label" y, defining $f^{x,y}(\cdot) := \ell(L^x(\cdot), y)$, we obtain that $(f^{x,y} \circ \phi)(\theta) = \ell(L^x(\phi(\theta)), y) = \ell(g_\theta(x), y)$ is C^∞ on $\phi(\Omega)$, so that $f^{x,y} \in F_\ell$, and the vector field $\chi^{x,y} : \theta \mapsto \nabla(f^{x,y} \circ \phi) = [\partial\phi(\theta)]^\top [\partial L^x(\phi(\theta))]^\top \nabla_z \ell(g_\theta(x), y))$ belongs to $V_\phi[F_\ell]$. As a consequence for all $\theta \in \Omega$:

$$V_\phi[F_\ell](\theta) \supseteq \partial\phi(\theta)^\top \operatorname*{span}_{x \in X, y} \{[\partial L^x(\phi(\theta))]^\top \nabla_z \ell(g_\theta(x), y))\} \stackrel{(16)}{=} \operatorname{range}\left(\partial\phi(\theta)^\top\right) = V_\phi[C^\infty](\theta). \ \square$$

As a corollary, we establish a result that decouples an assumption on the loss function ℓ and an assumption on the model g_θ, expressed via properties of L^x (under Assumption B.1).

Corollary B.3. *Consider a loss $\ell(z, y)$ that is differentiable with respect to $z \in \mathbb{R}^n$ and such that*

$$\operatorname{span}_y \nabla_z \ell(z, y) = \mathbb{R}^n, \qquad \forall z \in \mathbb{R}^n. \tag{17}$$

Under Assumption B.1, if for all $\theta \in \Omega$,

$$\operatorname{span}_{x \in X, w \in \mathbb{R}^n} \{[\partial L^x(\phi(\theta))]^\top w\} = \mathbb{R}^d, \tag{18}$$

then $\mathtt{CL}(V_\phi[C^\infty]) = \mathtt{CL}(V_\phi(F_\ell))$.

Proof. Altogether the two conditions imply (16) so that we can apply Proposition B.2. $\square$

We are now equipped to prove the theorem, beginning by the most easy setting of linear networks.

Theorem B.4 (linear networks)**.** *Consider a linear network parameterized by q matrices, $\theta = (U_1, \ldots, U_q)$ and defined via $g_\theta(x) := U_1 \ldots U_q x$. With $\phi(\theta) := U_1 \ldots U_q \in \mathbb{R}^{n \times m}$ (identified with $\mathbb{R}^d$ with $d = nm$), and for any loss ℓ satisfying (17), we have $\mathtt{CL}(V_\phi[C^\infty]) = \mathtt{CL}(V_\phi[F_\ell])$.*

Proof. Write $g_\theta(x) = \phi(\theta)x =: L^x(\phi(\theta)) \in \mathbb{R}^n$ for $\theta \in \Omega = \mathbb{R}^D$ and $x \in X = \mathbb{R}^m$. Since $\operatorname{span}_{x \in \mathbb{R}^m, w \in \mathbb{R}^n} \{[\partial L^x(\phi(\theta))]^\top w\} = \operatorname{span}_{x,w} \{wx^\top\} = \mathbb{R}^d$ we can apply Corollary B.3. $\square$

Before proceeding to the case of ReLU networks, let us show that (17) holds for standard ML losses.

Lemma B.5. *The mean-squared error loss $(z, y) \mapsto \ell_2(z, y) := \|y - z\|^2$ and the logistic loss $(z \in \mathbb{R}, y \in \{-1, 1\}) \mapsto \ell_{\mathtt{logis}}(z, y) := \log(1 + \exp(-zy))$ satisfy condition (17).*

Proof. To show that ℓ_2 satisfies (17) we observe that, with e_i the i-th canonical vector, we have

$$\mathbb{R}^n = \operatorname{span}\{e_i : 1 \le i \le n\} = \operatorname*{span}_{y \in \{z - e_i/2\}_{i=1}^n} 2(z - y) \subseteq \operatorname*{span}_{y \in \mathbb{R}^n} 2(z - y) = \operatorname*{span}_{y \in \mathbb{R}^n} \nabla_z \ell_2(z, y) \subseteq \mathbb{R}^n.$$

For the logistic loss, $\nabla_z \ell_{\mathtt{logis}}(z, y) = \frac{-y \exp(-zy)}{1 + \exp(-zy)} \ne 0$ hence $\operatorname{span}_y \nabla_z \ell_{\mathtt{logis}}(z, y) = \mathbb{R}$. $\square$

Remark B.6. In the case of the cross-entropy loss $(z \in \mathbb{R}^n, y \in \{1, \cdots, n\}) \mapsto \ell_{\mathtt{cross}}(z, y) := -z_y + \log\left(\sum_{i=1}^n \exp z_i\right)$, $\ell_{\mathtt{cross}}$ *does not* satisfy (17) as $\nabla_z \ell_{\mathtt{cross}}(z, y) = -e_y + \begin{pmatrix} \exp(z_1)/(\sum_i \exp z_i) \\ \cdots \\ \exp(z_n)/(\sum_i \exp z_i) \end{pmatrix}$ satisfies for all $z \in \mathbb{R}^n$:

$$\operatorname{span}_y \nabla_z \ell_{\mathtt{cross}}(z, y) = \{w := (w_1, \cdots, w_n) \in \mathbb{R}^n : \sum w_i = 0\} =: W_{\mathtt{cross}}.$$

An interesting challenge is to investigate variants of Corollary B.3 under weaker assumptions that would cover the cross-entropy loss.

We now treat the case of ReLU networks, using notations and concepts from [24] that generalize observations from Example 2.2 to deep ReLU networks with biases. Given a feedforward network architecture of arbitrary depth, denote θ the collection of all parameters (weights and biases) of a ReLU network on this architecture, and consider $\theta \mapsto \phi_{\mathtt{ReLU}}(\theta)$ the rescaling-invariant polynomial function of [24, Definition 6] and $C_{\theta,x}$, the matrices of [24, Corollary 3] such that the output of the network with parameters θ, when fed with an input vector $x \in \mathbb{R}^m$, can be written $g_\theta(x) = C_{\theta,x}\phi(\theta)$. From its definition in [24, Corollary 3], given x, the matrix $C_{\theta,x}$ only depends on θ via the so-called *activation status* of the neurons in the network (cf [24, Section 4.1]). By [24, Lemma 11 and

Definition 10], for each θ, there exists[1] an open set $\mathcal{X}_\theta \subseteq \mathbb{R}^m$ such that, for every $x \in \mathcal{X}_\theta$, the activation status is locally constant on a neighborhood of (θ, x). The specifics of how $\phi_{\texttt{ReLU}}(\theta)$ and $C_{\theta,x}$ are built will be provided soon, but they are unimportant at this stage of the analysis. With these notations we can proceed.

Lemma B.7. *Consider a feedforward ReLU network architecture and the rescaling-invariant polynomial function $\theta \mapsto \phi_{\texttt{ReLU}}(\theta)$ from [24, Definition 6]. Given a parameter θ_0 and $x_0 \in \mathcal{X}_{\theta_0} \subseteq \mathbb{R}^m$, there exists an open neighborhood Ω of θ_0 and an open neighborhood $X \subseteq \mathbb{R}^m$ of x_0 such that Assumption B.1 holds with the model $(x, \theta) \in X \times \Omega \mapsto g_\theta(x) \in \mathbb{R}^n$: $g_\theta(x) = L^x(\phi_{\texttt{ReLU}}(\theta)) := C_{\theta_0,x}\phi_{\texttt{ReLU}}(\theta)$ for all $x \in X$, $\theta \in \Omega$, with $C_{\theta,x}$ the matrices of [24, Corollary 3].*

Proof. As described above, by definition of $\mathcal{X}_{\theta_0}$ the activation status of the neurons is locally constant in a neighborhood of (θ_0, x_0), hence there exists an open neighborhood Ω of θ_0 and an open neighborhood X of x_0 such that $C_{\theta,x} = C_{\theta_0,x}$ for every $\theta \in \Omega$ and $x \in X$. The conclusion follows from the fact that $g_\theta(x) = C_{\theta,x}(\phi_{\texttt{ReLU}}(\theta))$ for every θ, x. □

Lemma B.8. *Consider a feedforward ReLU network architecture and the rescaling-invariant polynomial function $\theta \mapsto \phi_{\texttt{ReLU}}(\theta)$ from [24, Definition 6]. If θ_0 is a parameter such that:*

$$\mathrm{span}_{x \in \mathcal{X}_{\theta_0}, w \in \mathbb{R}^n}\{C_{\theta_0,x}^\top w\} = \mathbb{R}^d, \tag{19}$$

then θ_0 admits a neighborhood Ω such that, for any loss ℓ satisfying (17), $\texttt{CL}(V_{\phi_{\texttt{ReLU}}}[\mathcal{C}^\infty]) = \texttt{CL}(V_{\phi_{\texttt{ReLU}}}[F_\ell])$.

Proof. Let us assume that θ_0 verifies (19) and that the loss ℓ satisfies (17). By definition of a generated vector space in $\mathbb{R}^d$, there exists a finite set $\mathcal{X}_0 := \{x_i\}_{i=1}^d \subset \mathcal{X}_{\theta_0}$ such that:

$$\mathrm{span}_{x \in \mathcal{X}_0, w \in \mathbb{R}^n}\{C_{\theta_0,x}^\top w\} = \mathrm{span}_{x \in \mathcal{X}_{\theta_0}, w \in \mathbb{R}^n}\{C_{\theta_0,x}^\top w\} \overset{(19)}{=} \mathbb{R}^d. \tag{20}$$

Then, for $1 \leq i \leq d$ by applying Lemma B.7 there exists an open neighborhood Ω_i of θ_0 and an open neighborhood $X_i \subseteq \mathcal{X}_{\theta_0} \subseteq \mathbb{R}^m$ of x_i such that Assumption B.1 holds with the model $(x, \theta) \in X_i \times \Omega_i \mapsto g_\theta(x) \in \mathbb{R}^n$. Thus, by taking $\Omega := \cap_i \Omega_i$ and $X := \cup_i X_i$, Assumption B.1 holds with the model $(x, \theta) \in X \times \Omega \mapsto g_\theta(x) \in \mathbb{R}^n$: $g_\theta(x) = L^x(\phi_{\texttt{ReLU}}(\theta)) := C_{\theta_0,x}\phi_{\texttt{ReLU}}(\theta)$ for all $x \in X$, $\theta \in \Omega$, with $C_{\theta,x}$ the matrices of [24, Corollary 3]. Finally, for each $\theta \in \Omega$, we have

$$\mathrm{span}_{x \in X, w \in \mathbb{R}^n}\{[\partial L^x(\phi(\theta))]^\top w\} = \mathrm{span}_{x \in X, w \in \mathbb{R}^n}\{C_{\theta_0,x}^\top w\} \supseteq \mathrm{span}_{x \in \mathcal{X}_0, w \in \mathbb{R}^n}\{C_{\theta_0,x}^\top w\} \overset{(20)}{=} \mathbb{R}^d.$$

The conclusion follows using Corollary B.3. □

In the specific case of two-layer (ReLU) networks with r neurons, one can write $\theta = (U, V, b, c) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r} \times \mathbb{R}^r \times \mathbb{R}^n$, and denote u_j (resp. v_j, b_j) the columns of U (resp. columns of V, entries of b), so that $g_\theta(x) = \sum_{j=1}^r u_j \sigma(v_j^\top x + b_j) + c$. As soon $v_j \neq 0$ for every neuron j, the set $\mathcal{X}_\theta$ is [24, Definition 10] simply the complement in the input domain $\mathbb{R}^m$ of the union of the hyperplanes

$$\mathcal{H}_j := \{x \in \mathbb{R}^m : v_j^\top x + b_j = 0\}. \tag{21}$$

Theorem B.9 (two-layer ReLU networks). *On a two-layer ReLU network architecture, let θ be a parameter such that all hyperplanes $\mathcal{H}_j$ defined in (21) are pairwise distincts. Then θ admits a neighborhood such that, for any loss ℓ satisfying (17), we have $\texttt{CL}(V_{\phi_{\texttt{ReLU}}}[\mathcal{C}^\infty]) = \texttt{CL}(V_{\phi_{\texttt{ReLU}}}[F_\ell])$.*

Proof. Consider θ a parameter. By Lemma B.8, we only need to show that:

$$\mathrm{span}_{x \in \mathcal{X}_\theta, w \in \mathbb{R}^n}\{C_{\theta,x}^\top w\} = \mathbb{R}^d.$$

1st case: We consider first the case without bias ($b_j, c = 0$).

In that case we write: $\theta = (U, V) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r}$, and denote u_j (resp. v_j) the columns of U (resp. columns of V), as in Example 2.2.

---

[1]The set $\mathcal{X}_\theta$ essentially corresponds [24, Lemma 11] to all inputs for which each neuron is either strictly activated or strictly non-activated, cf [24, Definition 10].

Here $d = rnm$ and the matrix $C_{\theta,x}$ is defined by:

$$C_{\theta,x}^\top := \begin{pmatrix} \varepsilon_1(x,\theta)A(x) \\ \cdots \\ \varepsilon_r(x,\theta)A(x) \end{pmatrix} \in \mathbb{R}^{(rnm)\times n}$$

where:

$$A : x \in \mathbb{R}^m \mapsto A(x) := \begin{pmatrix} x & 0 & \cdots & 0 \\ 0 & x & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & x \end{pmatrix} \in \mathbb{R}^{(nm)\times n},$$

and where $\varepsilon_i(x,\theta) = \mathbb{1}(v_i^\top x > 0)$. Indeed one can easily check that the matrix $C_{\theta,x}$ satisfies the properties of Lemma B.7 where $\phi_{\texttt{ReLU}}$ coincides with the mapping of Example 2.2, $\phi(\theta) := (u_j v_j^\top)_{j=1}^r$. For $j = 1, \cdots, r$ we denote:

$$\mathcal{A}_j^+ := \{x \in \mathbb{R}^m : v_j^\top x > 0\}, \quad \text{and} \quad \mathcal{A}_j^- := \{x \in \mathbb{R}^m : v_j^\top x < 0\}.$$

The open Euclidean ball of radius $r > 0$ centered at $c \in \mathbb{R}^m$ is denoted $B(c, r)$.

Consider a hidden neuron $i \in \{1, \cdots, r\}$ and denote $\mathcal{H}_i' := \mathcal{H}_i - \left(\bigcup_{j \neq i} \mathcal{H}_j\right)$. Since the hyperplanes are pairwise distinct, $\mathcal{H}_i' \neq \emptyset$ so we can consider an arbitrary $x' \in \mathcal{H}_i'$. Given any $\eta > 0$, by continuity of $x \in \mathbb{R}^m \mapsto (v_1^\top x, \cdots, v_r^\top x) \in \mathbb{R}^r$, there exists $x_\eta^+ \in B(x', \eta) \cap \mathcal{A}_i^+$ and $x_\eta^- \in B(x', \eta) \cap \mathcal{A}_i^-$ such that for all $j \neq i$, $\text{sign}(v_j^\top x_\eta^\pm) = \text{sign}(v_j^\top x')$. It follows that $x_\eta^\pm \in \mathcal{X}_\theta$ (remember that $\mathcal{X}_\theta$ is the complement of $\cup_j \mathcal{H}_j$). As a consequence:

$$\begin{pmatrix} 0 \\ \cdots \\ 0 \\ A(x') \\ 0 \\ \cdots \\ 0 \end{pmatrix} = \lim_{\eta \to 0} \left(C_{\theta,x_\eta^+}^\top - C_{\theta,x_\eta^-}^\top \right) \in \overline{\underset{x \in \mathcal{X}_\theta}{\text{span}}\{C_{\theta,x}^\top\}} = \underset{x \in \mathcal{X}_\theta}{\text{span}}\{C_{\theta,x}^\top\},$$

where the nonzero line in the left-hand-side is the i-th, and we used that every finite-dimensional space is closed.

Moreover still by continuity of $x \in \mathbb{R}^m \mapsto (v_1^\top x, \cdots, v_r^\top x) \in \mathbb{R}^r$, there exists $\gamma > 0$, such that for $k = \{-2, -1, 1, 2\}$, the vectors defined as:

$$x_k := x' + \gamma k v_i,$$

satisfy for all $j \neq i$, $\text{sign}(v_j^\top x_k) = \text{sign}(v_j^\top x')$ and $v_i^\top x_k \neq 0$, so that $x_k \in \mathcal{X}_\theta$ and we similarly obtain

$$\begin{pmatrix} 0 \\ \cdots \\ 0 \\ \gamma A(v_i) \\ 0 \\ \cdots \\ 0 \end{pmatrix} = C_{\theta,x_2}^\top - C_{\theta,x_1}^\top - \left(C_{\theta,x_{-1}}^\top - C_{\theta,x_{-2}}^\top \right) \in \underset{x \in \mathcal{X}_\theta}{\text{span}}\{C_{\theta,x}^\top\}.$$

As this holds for every $x' \in \mathcal{H}_i'$, and since $\text{span}\{v_i, \mathcal{H}_i'\} = \mathbb{R}^m$, we deduce that for any $x \in \mathbb{R}^m$

$$\begin{pmatrix} 0 \\ \cdots \\ 0 \\ A(x) \\ 0 \\ \cdots \\ 0 \end{pmatrix} \in \underset{x \in \mathcal{X}_\theta}{\text{span}}\{C_{\theta,x}^\top\}.$$

As this holds for every hidden neuron $i = 1, \cdots, r$ it follows that for every $x^1, \cdots, x^r \in \mathbb{R}^m$

$$\begin{pmatrix} A(x^1) \\ \cdots \\ A(x^r) \end{pmatrix} \in \underset{x \in \mathcal{X}_\theta}{\text{span}}\{C_{\theta,x}^\top\}.$$

Moreover, by definition of $A(\cdot)$, for each $x \in \mathbb{R}^m$ and each $w = (w_1, \cdots, w_n) \in \mathbb{R}^n$, we have

$$A(x)w = \begin{pmatrix} w_1 x \\ \cdots \\ w_n x \end{pmatrix} \in \mathbb{R}^{nm}.$$

Identifying $\mathbb{R}^{nm}$ with $\mathbb{R}^{m \times n}$ and the above expression with $xw^\top$, we deduce that

$$\operatorname*{span}_{x \in \mathbb{R}^m, w \in \mathbb{R}^n} A(x)w = \mathbb{R}^{nm}$$

and we let the reader check that this implies

$$\operatorname*{span}_{x^1, \cdots, x^r \in \mathbb{R}^m, w \in \mathbb{R}^n} \begin{pmatrix} A(x^1) \\ \cdots \\ A(x^r) \end{pmatrix} w = \operatorname*{span}_{x^1, \cdots, x^r \in \mathbb{R}^m, w \in \mathbb{R}^n} \begin{pmatrix} A(x^1)w \\ \cdots \\ A(x^r)w \end{pmatrix} = \mathbb{R}^{rnm}.$$

Thus, as claimed, we have

$$\operatorname*{span}_{x \in \mathcal{X}_\theta, w \in \mathbb{R}^n} \{C_{\theta,x}^\top w\} = \mathbb{R}^{rnm} = \mathbb{R}^d.$$

2d case: General case with biases. The parameter is $\theta = (U, V, b, c) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r} \times \mathbb{R}^r \times \mathbb{R}^n$ with $b = (b_i)_{i=1}^r$, where $b_i \in \mathbb{R}$ the bias of the i-th hidden neuron, and c the output bias.

In that case, $d = rn(m+1)$ and one can check that the conditions of Lemma B.8 hold with $\phi_{\mathtt{ReLU}}(\theta) := ((u_i(v_i, b_i)^\top)_{i=1}^r, c)$ and the matrix $C_{\theta,x}$ defined by:

$$C_{\theta,x}^\top := \begin{pmatrix} \varepsilon_1(x, \theta)A'(x) \\ \cdots \\ \varepsilon_r(x, \theta)A'(x) \\ I_n \end{pmatrix} \in \mathbb{R}^{(rn(m+1)+n) \times n}$$

where, denoting $\bar{x} = (x^\top, 1)^\top \in \mathbb{R}^{m+1}$, we defined

$$A' : x \in \mathbb{R}^m \mapsto A'(x) := \begin{pmatrix} \bar{x} & 0 & \cdots & 0 \\ 0 & \bar{x} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \bar{x} \end{pmatrix} \in \mathbb{R}^{n(m+1) \times n},$$

and $\varepsilon_i(x, \theta) := \mathbb{1}(v_i^\top x + b_i > 0)$.

Using the sets

$$\mathcal{A}_j^+ := \{x \in \mathbb{R}^m : v_j^\top x + b_j > 0\}, \quad \text{and} \quad \mathcal{A}_j^- := \{x \in \mathbb{R}^m : v_j^\top x + b_j < 0\},$$

a reasoning analog to the case without bias allows to show that for each $i = 1, \cdots, r$:

$$\operatorname*{span}_{x \in \mathbb{R}^m} \begin{pmatrix} 0 \\ \cdots \\ 0 \\ A'(x) \\ 0 \\ \cdots \\ 0 \end{pmatrix} \in \operatorname*{span}_{x \in \mathcal{X}_\theta} \{C_{\theta,x}^\top\}$$

so that, again, for every $x^1, \ldots, x^r \in \mathbb{R}^m$ we have

$$\begin{pmatrix} A'(x^1) \\ \cdots \\ A'(x^r) \\ 0 \end{pmatrix} \in \operatorname*{span}_{x \in \mathcal{X}_\theta} \{C_{\theta,x}^\top\}.$$

As $\begin{pmatrix} 0 \\ \cdots \\ 0 \\ I_n \end{pmatrix} \in \operatorname*{span}_{x \in \mathcal{X}_\theta} \{C_{\theta,x}^\top\}$ too, we obtain that $\begin{pmatrix} A'(x^1) \\ \cdots \\ A'(x^r) \\ I_n \end{pmatrix} \in \operatorname*{span}_{x \in \mathcal{X}_\theta} \{C_{\theta,x}^\top\}.$

17

Now, for each $x \in \mathbb{R}^m$ and $w = (w_1, \cdots, w_n) \in \mathbb{R}^n$, we have

$$A'(x)w = \begin{pmatrix} w_1 \bar{x} \\ \cdots \\ w_n \bar{x} \end{pmatrix} = \begin{pmatrix} w_1 x \\ w_1 \\ \cdots \\ w_n x \\ w_n \end{pmatrix} \in \mathbb{R}^{n(m+1)}.$$

Again, identifying the above expression with $w(x^\top, 1) \in \mathbb{R}^{n \times (m+1)}$ it is not difficult to check that

$$\operatorname*{span}_{x \in \mathbb{R}^m, w \in \mathbb{R}^n} A'(x)w = \mathbb{R}^{n(m+1)},$$

and we conclude as before. $\qquad\square$

Combining Theorem B.4 and Theorem B.9 establishes the proof of Theorem 2.8 as claimed.

C Proof of Lemma 3.2

Lemma C.1. *Given $\theta \in \mathbb{R}^D$, if for a given i, $\dim V_{i+1}(\theta') = \dim V_i(\theta)$ for every θ' in a neighborhood of θ, then for all $k \geq i$, we have $V_k(\theta') = V_i(\theta')$ for all θ' in a neighborhood Ω of θ, where the V_i are defined by Proposition 3.1. Thus $\operatorname{Lie}(V)(\theta') = V_i(\theta')$ for all $\theta' \in \Omega$. In particular, the dimension of the trace of $\operatorname{Lie}(V)$ is locally constant and equal to the dimension of $V_i(\theta)$.*

Proof. The result is obvious for $k = i$. The proof is by induction on k starting from $k = i + 1$. We denote $m := \dim V_i(\theta)$.

1st step: Initialization $k = i + 1$. By definition of the spaces V_i (cf Proposition 3.1) we have $V_i \subset V_{i+1}$ hence $V_i(\theta) \subseteq V_{i+1}(\theta)$. Since $\dim V_{i+1}(\theta) = \dim V_i(\theta) = m$, it follows that there exists $\chi_1, \cdots, \chi_m \in V_i$ such that $\operatorname{span}_j \chi_j(\theta) = V_i(\theta) = V_{i+1}(\theta)$ (hence the m vectors $(\chi_1(\theta), \cdots, \chi_m(\theta))$ are linearly independent). Since each χ_j is smooth, it follows that $(\chi_1(\theta'), \cdots \chi_m(\theta'))$ remain linearly independent on some neighborhood Ω of θ, which we assume to be small enough to ensure $\dim V_{i+1}(\theta') = m$ for all $\theta' \in \Omega$. As $\chi_j \in V_i \subset V_{i+1}$, we obtain that for each $\theta' \in \Omega$, the family $\{\chi_j(\theta')\}_{j=1}^m$ is a basis of the m-dimensional subspace $V_{i+1}(\theta')$, hence:

$$V_i(\theta') \subset V_{i+1}(\theta') = \operatorname{span}_j \chi_j(\theta') \subset V_i(\theta'), \quad \forall \theta' \in \Omega \tag{22}$$

2nd step: Induction. We assume $V_k(\theta') = V_i(\theta')$ on Ω. Let us show that $V_{k+1}(\theta') = V_i(\theta')$ on Ω. Since $V_{k+1} := V_k + [V_0, V_k]$ it is enough to show that $[V_0, V_k](\theta') \subseteq V_i(\theta')$ on Ω. For this, considering two vector fields, $f \in V_0$ and $\chi \in V_k$, we will show that $[f, \chi](\theta') \in V_{i+1}(\theta')$ for each $\theta' \in \Omega$. In light of (22), this will allow us to conclude.

Indeed, from the induction hypothesis we know that $V_k(\theta') = \operatorname{span}_j \chi_j(\theta') = V_i(\theta')$ on Ω, hence for each $\theta' \in \Omega$ there are coefficients $a_j(\theta')$ such that $\chi(\theta') = \sum_{j=1}^m a_j(\theta') \chi_j(\theta')$. Standard linear algebra shows that these coefficients depend smoothly on $\chi(\theta')$ and $\chi_j(\theta')$, which are smooth functions of θ', hence the functions $a_j(\cdot)$ are smooth. By linearity of the Lie bracket and of $V_{i+1}(\theta')$ it is enough to show that $[f, a_j \chi_j](\theta') \in V_{i+1}(\theta')$ on Ω for each j. Standard calculus yields

$$[f, a_j \chi_j] = (\partial f)(a_j \chi_j) - \underbrace{\partial(a_j \chi_j)}_{= \chi_j \partial a_j + a_j \partial \chi_j} f = a_j[(\partial f)\chi_j - (\partial \chi_j)f] - \chi_j(\partial a_j)f$$

$$= a_j[f, \chi_j] - [(\partial a_j)f]\chi_j$$

since $(\partial a_j)f$ is scalar-valued (consider the corresponding dimensions). Since $f \in V_0$ and $\chi_j \in V_i$, by definition of V_{i+1} (cf Proposition 3.1) we have $[f, \chi_j], \chi_j \in V_{i+1}$ hence by linearity we conclude that $[f, a_j \chi_j](\theta') \in V_{i+1}(\theta')$. As this holds for all j, we obtain $[f, \chi](\theta') \in V_{i+1}(\theta')$. As this is valid for any $f \in V_0$, $\chi \in V_k$ this establishes $[V_0, V_k](\theta') \subseteq V_{i+1}(\theta') \overset{(22)}{=} V_i(\theta')$ and we conclude as claimed that $V_i(\theta') \subseteq V_{k+1}(\theta') = V_k(\theta') + [V_0, V_k](\theta') \subseteq V_i(\theta')$ on Ω. $\qquad\square$

D Proof of Theorem 3.3

We recall first the fundamental result of Frobenius using our notations (See Section 1.4 of [11]).

Theorem D.1 (Frobenius theorem). *Consider $V \subseteq \mathcal{X}(\Omega)$, and assume that the dimension of $V(\theta)$ is constant on $\Omega \subseteq \mathbb{R}^D$. Then the two following assertions are equivalent:*

1. *each $\theta \in \Omega$ admits a neighborhood Ω' such that there exists $D - \dim V(\theta)$ independent conserved functions through $V_{|\Omega'}$;*

2. *the following property holds:*

$$[u, v](\theta) \in V(\theta), \quad \text{for each } u, v \in V, \ \theta \in \Omega \tag{23}$$

Proposition D.2. *Under the assumption that $\dim V(\theta)$ is locally constant on Ω, Condition (23) of Frobenius Theorem holds if, and only if, the linear space $V' := \{\chi \in \mathcal{X}(\Omega), \forall \theta \in \Omega : \chi(\theta) \in V(\theta)\}$ (which is a priori infinite-dimensional) is a Lie algebra.*

Proof. $\Leftarrow$ If V' is a Lie algebra, then as $V \subset V'$ we get: for all $u, v \in V \subset V', [u, v] \in V'$. Given the definition of V' this means that (23) is satisfied.

$\Rightarrow$ Assuming now that (23) holds, we prove that V' is a Lie algebra. For this, given $X, Y \in V'$ we wish to show that $[X, Y](\theta) \in V(\theta)$ for every $\theta \in \Omega$.

Given $\theta \in \Omega$, we first reason as in the first step of the proof of Lemma 3.2 to obtain the existence of a neighborhood Ω' of θ and of $m := \dim V(\theta)'$ vector fields $\chi_1, \cdots, \chi_m \in V$ such that $(\chi_1(\theta'), \cdots, \chi_m(\theta'))$ is a basis of $V(\theta')$ for each $\theta' \in \Omega$. By definition of V' we have $X(\theta') \in V(\theta')$ and $Y(\theta') \in V(\theta')$ for every $\theta' \in \Omega'$. Thus, there are smooth functions a_j, b_j such that $X(\cdot) = \sum_1^m a_i(\cdot)\chi_i(\cdot)$ and $Y(\cdot) = \sum_1^m b_i(\cdot)\chi_i(\cdot)$ on Ω', and we deduce by bilinearity of the Lie brackets that $[X, Y](\theta') = \sum_{i,j}[a_i\chi_i, b_j\chi_j](\theta')$ on Ω'. Since $V(\theta)$ is a linear space, we will conclude that $[X, Y](\theta) \in V(\theta)$ if we can show that $[a_i\chi_i, b_j\chi_j](\theta) \in V(\theta)$. Indeed, we can compute

$$[a_i\chi_i, b_j\chi_j] = a_ib_j[\chi_i, \chi_j] + b_j[(\partial a_i)\chi_j]\chi_j - a_i[(\partial b_j)\chi_i]\chi_j$$

where, due to dimensions, both $(\partial a_i)\chi_j$ and $(\partial b_j)\chi_i$ are smooth scalar-valued functions. By construction of the basis $\{\chi_j\}_j$ we have $\chi_i(\theta), \chi_j(\theta) \in V(\theta)$, and by assumption (23) we have $[\chi_i, \chi_j](\theta) \in V(\theta)$, hence we conclude that $[X, Y](\theta) \in V(\theta)$. Since this holds for any choice of $X, Y \in V'$, this establishes that V' is a Lie algebra. $\square$

Theorem D.3. *If $\dim \mathrm{Lie}(V_\phi)(\theta)$ is locally constant then each $\theta \in \Omega$ has a neighborhood Ω' such that there are $D - \dim \mathrm{Lie}(V_\phi)(\theta)$ (and no more) independent conserved functions through $V_{\phi|\Omega'}$.*

Proof. 1st step: Existence of Ω' and of $D - \dim \mathrm{Lie}(V_\phi)(\theta)$ independent conserved functions. Let $\theta \in \Omega$. Since $\dim \mathrm{Lie}(V_\phi)(\theta)$ is locally constant there is a neighborhood Ω'' of θ on which it is constant. Since $V := (\mathrm{Lie}(V_\phi))_{|\Omega''} \subseteq \mathcal{X}(\Omega'')$ is a Lie Algebra, by Proposition D.2 and Frobenius theorem (Theorem D.1) there exists a neighborhood $\Omega' \subseteq \Omega''$ of θ and $D - \dim V(\theta)$ independent conserved functions through $V_{|\Omega'}$. As $V_\phi \subset \mathrm{Lie}(V_\phi)$, these functions are (locally) conserved through V_ϕ too. We only need to show that there are no more conserved functions.

2nd step: There are no more conserved functions. By contradiction, assume there exists $\theta_0 \in \Omega$, an open neighborhood Ω' of θ_0, a dimension $k < \dim \mathrm{Lie}(V_\phi)(\theta_0)$, and a collection of $D - k$ independent conserved functions through V_ϕ, gathered as the coordinates of a vector-valued function $h \in \mathcal{C}^1(\Omega', \mathbb{R}^{D-k})$. Consider $W := \{X \in \mathcal{X}(\Omega'), \forall \theta \in \Omega', X(\theta) \in \ker \partial h(\theta)\}$. By the definition of independent conserved functions, the rows of the $(D - k) \times D$ Jacobian matrix

$\partial h(\theta)$ are linearly independent on Ω', and the dimension of $W(\theta) = \ker \partial h(\theta)$ is constant and equal to k on Ω'. By construction of W and Proposition 2.5, the $D - k$ coordinate functions of h are independent conserved functions through W. Thus, by Frobenius Theorem (Theorem D.1) and Proposition D.2, W is a Lie algebra. By Proposition 2.5 we have $V_\phi(\theta) = \mathrm{range}\,\partial\phi(\theta)^\top \subset \ker \partial h(\theta)$ on Ω', hence $V_{\phi|\Omega'} \subset W$, and therefore $\mathrm{Lie}(V_\phi)_{|\Omega'} = \mathrm{Lie}(V_{\phi|\Omega'}) \subset W$. In particular: $\mathrm{Lie}(V_\phi)(\theta_0) \subset W(\theta_0)$, which leads to the claimed contradiction that $\dim \mathrm{Lie}(V_\phi)(\theta_0) \leq \dim W(\theta_0) = k$.

$\square$

E Proofs of the Examples of Section 3.3 and additional example

E.1 Proof of the result given in Example 3.6

Proposition E.1. *Consider $\theta = (U, V) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r}$, ϕ, and $\Omega \subseteq \mathbb{R}^D$, $D = (n + m)r$, as in Example 3.6. The dimension of $V_\phi(\theta)$ is constant and equal to $(n + m - 1)r$ and V_ϕ verifies condition* (11) *of Frobenius Theorem (i.e. condition* (23) *of Theorem D.1).*

Proof. Denoting u_i (resp. v_i) the columns of U (resp. of V), for $\theta \in \Omega$ we can write $\phi(\theta) = (\psi(u_i, v_i))_{i=1,\cdots r}$ with $\psi : (u \in \mathbb{R}^n - \{0\}, v \in \mathbb{R}^m - \{0\}) \mapsto uv^\top \in \mathbb{R}^{n \times m}$. As this decouples ϕ into r functions each depending on a separate block of coordinates, Jacobian matrices and Hessian matrices are block-diagonal. Establishing condition (23) of Frobenius theorem is thus equivalent to showing it for each block, which can be done by dealing with the case $r = 1$. Similarly, $V_\phi(\theta)$ is a direct sum of the spaces associated to each block, hence it is enough to treat the case $r = 1$ (by proving that the dimension is $n + m - 1$) to obtain that for any $r \geq 1$ the dimension is $r(n + m - 1)$.

1st step: We show that V_ϕ satisfies condition (23) *of Frobenius Theorem.* For $u \in \mathbb{R}^n - \{0\}$, $v \in \mathbb{R}^m - \{0\}$ we write $\theta = (u; v) \in \mathbb{R}^D = \mathbb{R}^{n+m}$ and $\phi_{i,j}(\theta) := u_i v_j$ for $i = 1, \cdots, n$ and $j = 1, \cdots, m$. Now u_i and v_j are *scalars* (and no longer columns of U and V). Denoting $e_i \in \mathbb{R}^D = \mathbb{R}^{n+m}$ the vector such that all its coordinates are null except the i-th one, we have:

$$\nabla \phi_{i,j}(\theta) = v_j e_i + u_i e_{n+j} \in \mathbb{R}^D,$$
$$\partial^2 \phi_{i,j}(\theta) = E_{j+n,i} + E_{i,j+n} \in \mathbb{R}^{D \times D},$$

with $E_{i,j} \in \mathbb{R}^{D \times D}$ the one-hot matrix with the (i, j)-th entry being 1. Let $i, k \in \{1, \cdots, n\}$ and $j, l \in \{1, \cdots, m\}$.

1st case: $(i, j) = (k, l)$ Then trivially $\partial^2 \phi_{i,j}(\theta) \nabla \phi_{k,l}(\theta) - \partial^2 \phi_{k,l}(\theta) \nabla \phi_{i,j}(\theta) = 0$.
2nd case: $((i \neq k)$ *and* $(j \neq l))$ Then

$$[\nabla \phi_{i,j}, \nabla \phi_{k,l}](\theta) = (E_{j+n,i} + E_{i,j+n})(v_l e_k + u_k e_{n+l}) - (E_{l+n,k} + E_{k,l+n})(v_j e_i + u_i e_{n+j}) = 0 - 0.$$

3d case: $i = k$ *and* $j \neq l$. Then as $u \neq 0$, there exists $l' \in \{1, \cdots, n\}$ such that $u_{l'} \neq 0$.

$$\partial^2 \phi_{i,j}(\theta) \nabla \phi_{k,l}(\theta) - \partial^2 \phi_{k,l}(\theta) \nabla \phi_{i,j}(\theta) = v_l e_{n+j} - v_j e_{n+l}$$
$$= \frac{v_l}{u_{l'}} \nabla \phi_{l',j}(\theta) - \frac{v_j}{u_{l'}} \nabla \phi_{l',l}(\theta),$$
$$\in \text{span}\{\nabla \phi_{i,j}(\theta)\} = V_\phi(\theta).$$

4d case: $((i \neq k)$ *and* $(j = l))$ We treat this case in the exact same way than the 3d case.

Thus V_ϕ verifies condition (11) of Frobenius Theorem.

2d step: We show that $\dim V_\phi(\theta) = (n + m - 1)$. As $u, v \neq 0$ each of these vectors has at least one nonzero entry. For simplicity of notation, and without loss of generality, we assume that $u_1 \neq 0$ and $v_1 \neq 0$. It is straightforward to check that $(\nabla \phi_{1,1}(\theta), (\nabla \phi_{1,j}(\theta))_{j=2,\cdots,m}, (\nabla \phi_{i,1}(\theta))_{i=2,\cdots,n})$ are $n + m - 1$ linearly independent vectors. To show that $\dim V_\phi(\theta) = (n + m - 1)$ is it thus sufficient to show that they span $V_\phi(\theta)$. This is a direct consequence of the fact that, for any i, j, we have

$$\nabla \phi_{i,j}(\theta) = v_j e_i + u_i e_{n+j} = \frac{v_j}{v_1}(v_1 e_i + u_i e_{n+1}) + \frac{u_i}{u_1}(u_1 e_{n+j} + v_j e_1) - \frac{v_j u_i}{u_1 v_1}(u_1 e_{n+1} + v_1 e_1),$$
$$= \frac{v_j}{v_1} \nabla \phi_{i,1}(\theta) + \frac{u_i}{u_1} \nabla \phi_{1,j}(\theta) + \frac{v_j u_i}{u_1 v_1} \nabla \phi_{1,1}(\theta). \qquad \square$$

E.2 An additional example beyond ReLU

In complement to Example 3.6, we give a simple example studying a two-layer network with a positively homogeneous activation function, which include the ReLU but also variants such as the leaky ReLU or linear networks.

Example E.2 (Beyond ReLU: Neural network with one hidden neuron with a positively homogeneous activation function of degree one). Let σ be a positively one-homogeneous activation function. In (2), this corresponds to setting $g_\theta(x) = \sum_{i=1}^r u_i \sigma(\langle v_i, x \rangle) \in \mathbb{R}$. Assuming $\langle v_i, x \rangle \neq 0$

for all i to avoid the issue of potential non-differentiability at 0 of σ (for instance for the ReLU), and in particular assuming $v_i \neq 0$, the function minimized during training can be factored via $\phi(\theta) = (\psi(u_i, v_i))_{i=1}^r$ where

$$\theta := (u \in \mathbb{R}, v \in \mathbb{R}^{d-1} - \{0\}) \overset{\psi}{\mapsto} (u\|v\|, v/\|v\|) \in \mathbb{R} \times \mathcal{S}_{d-1} \subset \mathbb{R}^d. \tag{24}$$

Proposition E.3. Consider $d \geq 2$ and $\phi(\theta) = (\psi(u_i, v_i))_{i=1}^r$ where ψ is given by (24) on $\Omega := \{\theta = (u \in \mathbb{R}^r, V = (v_1, \ldots, v_r) \in \mathbb{R}^{m \times r}) : v_i \neq 0\}$. We have $\dim V_\phi(y) = r(d-1)$ and V_ϕ verifies condition (23) of Frobenius Theorem (Theorem D.1), so each $\theta = (u, V) \in \Omega$ admits a neighborhood Ω' such that there exists r (and no more) conserved function through $V_{\phi|\Omega'}$.

As in Example 3.6, such candidate functions are given by $h_i : (u_i, v_i) \mapsto u_i^2 - \|v_i\|^2$. A posteriori, these functions are in fact conserved through all V_ϕ.

Proof of Proposition E.3. As in the proof of Proposition E.1 it is enough to prove the result for $r = 1$ hidden neuron. Note that here $D = d$. To simplify notations, we define $\phi_0, \ldots, \phi_{d-1}$ for $\theta = (u, v)$ as:

$$\phi_0(\theta) = u\|v\|,$$

and for $i = 1, \ldots, d-1$:

$$\phi_i(\theta) = v_i/\|v\|.$$

1st step: explicitation of $\mathrm{span}\{\nabla\phi_0, \ldots, \nabla\phi_{d-1}\}$. We have

$$\partial\phi(\theta) = \left(\begin{array}{c|c} \|v\| & uv^\top/\|v\| \\ \hline 0_{(d-1)\times 1} & \frac{1}{\|v\|}P_v \end{array} \right),$$

where: $P_v := \mathrm{I}_{d-1} - vv^\top/\|v\|^2$ is the orthogonal projector on $(\mathbb{R}v)^\perp$ (seen here as a subset of $\mathbb{R}^{d-1}$) and its rank is $d-2$. Thus $\dim V_\phi(\theta) = \mathrm{rank}(\partial\phi(\theta)) = d-1$ and $\mathrm{span}\{\nabla\phi_0, \ldots, \nabla\phi_{d-1}\} = \mathbb{R}\nabla\phi_0 + (\mathbb{R}v)^\perp$.

2d step: calculation of the Hessians.

1st case: The Hessian of ϕ_i for $i \geq 1$. In this case, ϕ_i does not depend on the first coordinate u so we proceed as if the ambient space here was $\mathbb{R}^{d-1}$. We have already that for $i \geq 1$:

$$\nabla\phi_i(\theta) = e_i/\|v\| - v_i v/\|v\|^3$$

hence

$$\partial^2\phi_i = 3v_i vv^\top/\|v\|^5 - 1/\|v\|^3 \left(v_i \mathrm{I}_{d-1} + V_i + V_i^\top\right),$$

where all columns of matrix $V_i := (0, \ldots, v, 0, \ldots, 0)$ are zero except the i-th one, which is set to v.

2d case: The Hessian of ϕ_0. Since

$$\nabla\phi_0(\theta) = \left(\|v\|, uv^\top/\|v\|\right)^\top.$$

we have

$$\partial^2\phi_0(\theta) = \left(\begin{array}{c|c} 0 & v^\top/\|v\| \\ \hline v/\|v\| & \frac{u}{\|v\|}P_v \end{array} \right).$$

3rd step: Conclusion.

1st case: $i, j \geq 1$ and $i \neq j$. We have:

$$\partial^2\phi_i(\theta)\nabla\phi_j(\theta) - \partial^2\phi_j(\theta)\nabla\phi_i(\theta),$$
$$= v_j/\|v\|^4 e_i - v_i/\|v\|^4 e_j \in (\mathbb{R}v)^\perp,$$
$$\subset \mathrm{span}\{\nabla\phi_0(\theta), \ldots, \nabla\phi_{d-1}(\theta)\}.$$

2d case: $i \geq 1$ and $j = 0$. We have:
$$\partial^2 \phi_i(\theta) \nabla \phi_0(\theta) - \partial^2 \phi_0(\theta) \nabla \phi_i(\theta),$$
$$= -2u/\|v\| \nabla \phi_i(\theta),$$
$$\in \text{span}\{\nabla \phi_0(\theta), ..., \nabla \phi_{d-1}(\theta)\}. \qquad \square$$

In both cases, we obtain as claimed that the condition (23) of Frobenius Theorem is satisfied, and we conclude using the latter.

F Proof of Proposition 3.8 and additional example

Proposition F.1. *Assume that* $\text{rank}(\partial \phi(\theta))$ *is constant on* Ω *and that* V_ϕ *satisfies* (11)*. If* $t \mapsto \theta(t)$ *satisfies the ODE* (3) *then there is* $0 < T^\star_{\theta_{\text{init}}} < T_{\theta_{\text{init}}}$ *such that* $z(t) := \phi(\theta(t)) \in \mathbb{R}^d$ *satisfies the ODE*
$$\begin{cases} \dot{z}(t) & = -M(z(t), \theta_{init}) \nabla f(z(t)) \quad \text{for all } 0 \leq t < T^\star_{\theta_{\text{init}}}, \\ z(0) & = \phi(\theta_{init}), \end{cases} \qquad (25)$$
where $M(z(t), \theta_{init}) \in \mathbb{R}^{d \times d}$ *is a symmetric positive semi-definite matrix.*

Proof. As $z = \phi(\theta)$ and as θ satisfies (3), we have:
$$\dot{z} = \partial \phi(\theta) \dot{\theta} = -\partial \phi(\theta) \nabla (f \circ \phi)(\theta) = -\partial \phi(\theta)[\partial \phi(\theta)]^\top \nabla f(z).$$

Thus, we only need to show $M(t) := \partial \phi(\theta(t))[\partial \phi(\theta(t))]^\top$, which is a symmetric, positive semi-definite $d \times d$ matrix, only depends on $z(t)$ and θ_{init}. Since $\dim V_\phi(\theta) = \text{rank}(\partial \phi(\theta))$ is constant on Ω and V_ϕ satisfies (11), by Frobenius Theorem (Theorem D.1), for each $\theta \in \Omega$, there exists a neighborhood Ω_1 of θ and $D - d'$ independent conserved functions $h_{d'+1}, \cdots, h_D$ through $(V_\phi)_{|\Omega'}$, with $d' := \dim V_\phi(\theta) = \text{rank}(\partial \phi(\theta))$. Moreover, by definition of the rank, for the considered θ, there exists a set $I \subset \{1, \ldots, d\}$ of d' indices such that the gradient vectors $\nabla \phi_i(\theta)$, $i \in I$ are linearly independent. By continuity, they stay linearly independent on a neighborhood Ω_2 of θ. Let us denote P_I the restriction to the selected indices and
$$\theta' \in \mathbb{R}^D \longmapsto \Phi_I(\theta') := (P_I \phi(\theta'), h_{d'+1}(\theta'), ..., h_D(\theta')) \in \mathbb{R}^D$$

As the functions h_i are *independent* conserved functions, for each $\theta' \in \Omega' := \Omega_1 \cap \Omega_2$ their gradients $\nabla h_i(\theta')$, $d' + 1 \leq i \leq D$ are both linearly independent and (by Proposition 2.5 and (7)) orthogonal to $V_\phi(\theta') = \text{range}[\partial \phi(\theta')]^\top = \text{span}\{\nabla \phi_i(\theta) : i \in I\}$. Hence, on Ω', the Jacobian $\partial \Phi_I$ is an invertible $D \times D$ matrix. By the implicit function theorem, the function Φ_I is thus locally invertible. Applying this analysis to $\theta = \theta(0)$ and using that h_i are conserved functions, we obtain that in an interval $[0, T^\star_{\theta_{\text{init}}})$ we have
$$\Phi_I(\theta(t)) = (P_I z(t), h_{d+1}(\theta_{\text{init}}), ..., h_D(\theta_{\text{init}})) \qquad (26)$$
By local inversion of Φ_I this allows to express $\theta(t)$ (and therefore also $M(t) = \partial \phi(\theta(t))[\partial \phi(\theta(t))]^\top$) as a function of $z(t)$ and of the initialization. $\square$

In complement to Example 3.9 we provide another example related to Example E.2.
Example F.2. Given the mapping $\phi : (u \in \mathbb{R}, v \in \mathbb{R}^{d-1} - \{0\}) \mapsto (u\|v\|, v/\|v\|) \in \mathbb{R} \times \mathcal{S}_{d-1} \subset \mathbb{R}^d$ (cf (24)), the variable $z := (r, h) = (u\|v\|, v/\|v\|)$ satisfies (25) with:
$$M(z, \theta_{\text{init}}) = \begin{pmatrix} \sqrt{r^2 + \delta^2} & 0_{1 \times k} \\ \hline 0_{(d-1) \times 1} & \frac{1}{\delta + \sqrt{r^2 + \delta^2}} P_h \end{pmatrix}, \text{ where } P_h := \text{I}_{d-1} - hh^T/\|h\|^2 \text{ and } \delta :=$$
$u_{\text{init}}^2 - \|v_{\text{init}}\|^2$.

G Proofs of results of Section 4

G.1 Proof of Proposition 4.2

Proposition G.1. *Consider* $\Psi : (U, V) \mapsto U^\top U - V^\top V \in \mathbb{R}^{r \times r}$ *and assume that* $(U; V)$ *has full rank. Then:*

1. if $n + m \leq r$, the function Ψ gives $(n+m)(r - 1/2(n+m-1))$ independent conserved functions,

2. if $n + m > r$, the function Ψ gives $r(r+1)/2$ independent conserved functions.

Proof. Let write $U = (U_1; \cdots; U_r)$ and $V = (V_1; \cdots; V_r)$ then: $\Psi_{i,j}(U,V) = \langle U_i, U_j \rangle - \langle V_i, V_j \rangle$ for $i,j = 1, \cdots, r$. Then $f_{i,j} := \nabla \Psi_{i,j}(U,V) = (0; \cdots; 0; \underset{(i)}{U_j}; \cdots; \underset{(j)}{U_i}; 0; \cdots; \underset{(i+r)}{-V_j}; \cdots; \underset{(j+r)}{V_i}; \cdots; 0)^\top \in \mathbb{R}^{(n+m)r \times 1}$.

1st case: $n + m \leq r$. As $(U;V)$ has full rank, its rank is $n + m$. In particular, U and V have a full rank too. Without loss of generality we can assume that $(U_1, \cdots, U_{n+m})$ are linearly independent, and $(V_1, \cdots, V_{n+m})$ too. Then for all $i > n + m$, $U_i \in \mathcal{F}_U := \mathrm{span}(U_1, \cdots, U_{n+m})$ and $V_i \in \mathcal{F}_V := \mathrm{span}(V_1, \cdots, V_{n+m})$. We want to count the number of $f_{i,j}$ that are linearly independent.

1. if $i \leq j \in [\![1, n+m]\!]$, then all the associated $f_{i,j}$ are linearly independent together. There are $(n+m)(n+m+1)/2$ such functions. Moreover, these functions generate vectors of the form:
$$(A_1; \cdots; A_{n+m}; 0; \cdots; 0; B_1; \cdots; B_{n+m}; 0; \cdots; 0)$$
where $A_i \in \mathcal{F}_U$ and $B_i \in \mathcal{F}_V$.

2. if $i \in [\![1, n+m]\!]$ and $j \in [\![n+m+1, r]\!]$, then all of the associated $f_{i,j}$ are linearly independent and the last ones are linearly independent together. We obtain $(n+m)(r - (n+m))$ more functions. Moreover, these functions generate vectors of the form:
$$(0 \cdots; 0; A_{n+m+1}; \cdots; A_r;$$
$$0; \cdots; 0; B_{n+m+1}; \cdots; B_r)$$
where $A_i \in \mathcal{F}_U$ and $B_i \in \mathcal{F}_V$.

3. if $i \leq j \in [\![n+m+1, r]\!]$, the associated $f_{i,j}$ are linearly dependent of thus already obtained.

Finally there are exactly $(n+m)(r - 1/2(n+m-1)$ independent conserved functions given by Ψ.

2d case: $n + m > r$. Then all $(U_i; -V_i)$ for $i = 1, \cdots r$ are linearly independent. Then there are $r(r+1)/2$ independent conserved functions given by Ψ. $\square$

G.2 Proofs of other results

Proposition G.2. *For every* $\Delta \in \mathbb{R}^{n \times m}$ *denote* $S_\Delta := \begin{pmatrix} 0 & \Delta \\ \Delta^\top & 0 \end{pmatrix}$, *one has* $\partial \phi(U,V)^\top : \Delta \in \mathbb{R}^{n \times m} \mapsto S_\Delta \cdot (U;V)$. *Hence* $V_\phi = \mathrm{span}\{A_\Delta, \forall \Delta \in \mathbb{R}^{n \times m}\}$, *where* $A_\Delta : (U;V) \mapsto S_\Delta \cdot (U;V)$ *is a linear endomorphism. Moreover one has* $[A_\Delta, A_{\Delta'}] : (U,V) \mapsto [S_\Delta, S_{\Delta'}] \times (U;V)$.

This proposition enables the computation of the Lie brackets of V_ϕ by computing the Lie bracket of matrices. In particular, $\mathrm{Lie}(V_\phi)$ is necessarily of finite dimension.

Proposition G.3. *The Lie algebra* $\mathrm{Lie}(V_\phi)$ *is equal to*
$$\left\{ (U;V) \mapsto \begin{pmatrix} \mathrm{I}_n & 0 \\ 0 & -\mathrm{I}_m \end{pmatrix} \times M \times \begin{pmatrix} U \\ V \end{pmatrix} : M \in \mathcal{A}_{n+m} \right\}$$
where $\mathcal{A}_{n+m} \subset \mathbb{R}^{(n+m) \times (n+m)}$ *is the space of skew symmetric matrices.*

Remark G.4. By the characterization of $\mathrm{Lie}(V_\phi)$ in Proposition G.3 we have that the dimension of $\mathrm{Lie}(V_\phi)$ is equal to $(n+m) \times (n+m-1)/2$.

Proof. 1st step: Let us characterize $V_1 = \mathrm{span}\{V_\phi + [V_\phi, V_\phi]\}$. Let $\Delta, \Delta' \in \mathbb{R}^{n \times m}$, then:
$$[A_\Delta, A_{\Delta'}]((U,V)) = [S_\Delta, S_{\Delta'}] \times (U;V) = \begin{pmatrix} Y, 0 \\ 0, Z \end{pmatrix} \times \begin{pmatrix} U \\ V \end{pmatrix}, \tag{27}$$

23

with $Y := \Delta\Delta'^\top - \Delta'\Delta^\top \in \mathcal{A}_n$ and $Z := \Delta^\top\Delta' - \Delta'^\top\Delta \in \mathcal{A}_m$. Then:

$$V_1 = \left\{ (U;V) \mapsto \begin{pmatrix} Y, X \\ X^\top, Z \end{pmatrix} \times \begin{pmatrix} U \\ V \end{pmatrix} : X \in \mathbb{R}^{n\times m}, Y \in \mathcal{A}_n, Z \in \mathcal{A}_m \right\},$$

$$= \left\{ u_M := (U;V) \mapsto \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} \times M \times \begin{pmatrix} U \\ V \end{pmatrix} : M \in \mathcal{A}_{n+m} \right\}.$$

2d step: Let us show that $V_2 = V_1$. Let $M, M' \in \mathcal{A}_{n+m}$. Then:

$$[u_M, u_{M'}] = \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} \left(M \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} M' - M' \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} M \right) = \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} \tilde{M},$$

with $\tilde{M} := M \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} M' - M' \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} M \in \mathcal{A}_{n+m}$.

Finally: $\mathrm{Lie}(V_\phi) = V_1 = \left\{ (U;V) \mapsto \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} \times M \times \begin{pmatrix} U \\ V \end{pmatrix} : M \in \mathcal{A}_{n+m} \right\}.$ $\qquad\square$

Eventually, what we need to compute is the dimension of the trace $\mathrm{Lie}(V_\phi)(U,V)$ for any (U,V).

Proposition G.5. *Let us assume that $(U;V) \in \mathbb{R}^{(n+m)\times r}$ has full rank. Then:*

1. *if $n + m \leq r$, then $\dim\mathrm{Lie}(V_\phi)(U;V) = (n+m)(n+m-1)/2$;*

2. *if $n + m > r$, then $\dim\mathrm{Lie}(V_\phi)(U;V) = (n+m)r - r(r+1)/2$.*

Proof. Let us consider the linear application:

$$\Gamma : M \in \mathcal{A}_{n+m} \mapsto \begin{pmatrix} I_n, 0 \\ 0, -I_m \end{pmatrix} \times M \times \begin{pmatrix} U \\ V \end{pmatrix},$$

where $\mathcal{A}_{n+m} \subset \mathbb{R}^{(n+m)^2}$ is the space of skew symmetric matrices. As $\mathrm{range}\Gamma(\mathcal{A}_{n+m}) = \mathrm{Lie}(V_\phi)(U;V)$, we only want to calculate $\mathrm{rank}\Gamma(\mathcal{A}_{n+m})$. But by rank–nullity theorem, we have:

$$\dim \ker \Gamma + \mathrm{rank}\ \Gamma = (n+m)(n+m-1)/2.$$

1st case: $n + m \leq r$. Then as $(U;V)$ has full rank $n + m$, Γ is injective and then $\mathrm{rank}\Gamma(\mathcal{A}_{n+m}) = (n+m)(n+m-1)/2$.

2d case: $n + m > r$. We write $(U;V) = (C_1; \cdots ; C_r)$ with $(C_1, \cdots, C_r)$ that are linearly independent as $(U;V)$ has full rank r. Let $M \in \mathcal{A}_{n+m}$ such that $\Gamma(M) = 0$. Then $M \cdot (U;V) = 0$. Then we write $M^\top = (M_1; \cdots ; M_{n+m})$. Then as $M \times (U;V) = 0$, we have that $\langle M_i, C_j \rangle = 0$ for all $i = 1, \cdots, n+m$ and for all $j = 1, \cdots, r$. We note $C := \mathrm{span}_{i=1,\cdots,r} C_i$ that is of dimension r as $(U;V)$ has full rank r.

M_1 must be in $C^\perp$ and its first coordinate must be zero as M must be a skew matrix. Then M_1 lies in a space of dimension $n + m - r - 1$. Then M_2 must be in $C^\perp$ too, and its first coordinate is determined by M_1 and its second is null as M is a skew matrix. Then M_2 lies in a space of dimension $n + m - r - 2$. By recursion, after building $M_1, \cdots, M_i$, M_{i+1} must be in $C^\perp$ too, and its i first coordinates are determined by $M_1, \cdots, M_i$ and its $i+1$-th one is null as M is a skew matrix. Then M_{i+1} lies in a space of dimension $\max(0, n+m-r-(i+1))$. Finally the dimension of $\ker\Gamma$ is equal to:

$$\sum_{i=1}^{n+m-r} (n+m-r-i) = (n+m-r-1)(n+m-r)/2.$$

Then: $\mathrm{rank}\Gamma(\mathcal{A}_{n+m}) = (n+m)r - r(r+1)/2.$ $\qquad\square$

Thanks to this explicit characterization of the trace of the generated Lie algebra, combined with Proposition 4.2, we conclude that Proposition 4.1 has indeed exhausted the list of independent conservation laws.

Corollary G.6. *If $(U; V)$ has full rank, then all conserved functions are given by $\Psi : (U, V) \mapsto U^\top U - V^\top V$. In particular, there exist no more conserved functions.*

Proof. As $(U; V)$ has full rank, this remains locally the case. By Proposition 4.3 the dimension of $\mathrm{Lie}(V_\phi)(U; V)$ is locally constant, denoted $m(U, V)$. By Theorem 3.4, the exact number of independent conserved functions is equal to $(n + m)r - m(U, V)$ and that number corresponds to the one given in Proposition 4.2. $\qquad\square$

H   About Example 3.7

Proposition H.1. *Let us assume that $(U; V) \in \mathbb{R}^{(n+m)\times r}$ has full rank. If $\max(n, m) > 1$ and $r > 1$, then V_ϕ does not satisfy the condition (11).*

Proof. Let us consider the linear application:

$$\Gamma' : \Delta \in \mathbb{R}^{n\times m} \mapsto \begin{pmatrix} 0, \Delta \\ \Delta^\top, 0 \end{pmatrix} \times \begin{pmatrix} U \\ V \end{pmatrix}.$$

By Proposition G.2, $\mathrm{range}\Gamma'(\mathbb{R}^{n\times m}) = V_\phi(U; V)$. Thus, as by definition $V_\phi(U; V) \subseteq \mathrm{Lie}(V_\phi(U; V))$, V_ϕ does not satisfy the condition (11) if and only if $\dim V_\phi(U; V) < \dim\mathrm{Lie}V_\phi(U; V)$.

1st case: $n + m \leq r$. Then as $(U; V)$ has full rank $n + m$, Γ' is injective and then $\mathrm{rank}\Gamma'(\mathbb{R}^{n\times m}) = n \times m$.

Thus by Proposition G.5, we only need to verify that: $n \times m < (n + m)(n + m - 1)/2 =:$ $\mathrm{Lie}V_\phi(U; V)$. It is the case as $\max(n, m) > 1$.

2d case: $n + m > r$. We write $(U; V) = (C_1; \cdots; C_r)$ with $(C_1, \cdots, C_r)$ that are linearly independent as $(U; V)$ has full rank r. Let $\Delta \in \mathbb{R}^{n\times m}$ such that $\Gamma'(\Delta) = 0$. Let us define the symmetric matrix M by:

$$M := \begin{pmatrix} 0, \Delta \\ \Delta^\top, 0 \end{pmatrix}. \tag{28}$$

Then $M \cdot (U; V) = 0$. Then we write $M^\top = (M_1; \cdots; M_{n+m})$. Then as $M \times (U; V) = 0$, we have that $\langle M_i, C_j \rangle = 0$ for all $i = 1, \cdots, n + m$ and for all $j = 1, \cdots, r$. We note $C := \underset{i=1,\cdots,r}{\mathrm{span}} C_i$ that is of dimension r as $(U; V)$ has full rank r.

For all $i = 1, \cdots, n$, M_i must be in $C^\perp$ and its n first coordinate must be zero by definition (28). Then M_i lies in a space of dimension $\max(0, n + m - r - n)$. For all $j > n$, M_j are entirely determined by $\{M_i\}_{i\leq n}$ by definition (28). Finally the dimension of $\ker\Gamma'$ is equal to: $n \times \max(0, m - r)$. Then: $\dim V_\phi(U; V) = \mathrm{rank}\Gamma'(\mathbb{R}^{n\times m}) = nm - n \times \max(0, m - r)$.

Thus by Proposition G.5, we only need to verify that: $nm - n\max(0, m - r) < (n + m)r - r(r + 1)/2 =: \mathrm{Lie}V_\phi(U; V)$.

Let us assume $m < r$. Then by looking at $f(r) := (n+m)r - r(r+1)/2 - nm = \dim\mathrm{Lie}V_\phi(U; V) - \dim V_\phi(U; V)$ for $r \in \{m+1, \cdots, n+m-1\} =: I_{n,m}$, we have: $f'(r) = (n + m) - 1/2 - r > 0$ (as $n + m > r$ is an integer), so f is increasing, so on $I_{n,m}$, we have (as $r > m$): $f(r) > f(m) = (n + m)m - m(m + 1)/2 - nm = m^2 - m(m + 1)/2 \geq 0$ as $m \geq 1$.

Let us assume $m \geq r$. Then

$$\begin{aligned}
\dim\mathrm{Lie}V_\phi(U; V) - \dim V_\phi(U; V) &= (n + m)r - r(r + 1)/2 - (nm - n(m - r)), \\
&= mr - r(r + 1)/2, \\
&\geq r^2 - r(r + 1)/2 \quad \text{as } m \geq r, \\
&> 0 \quad \text{as } r > 1.
\end{aligned}$$

Thus $\dim\mathrm{Lie}V_\phi(U; V) - \dim V_\phi(U; V) > 0$. $\qquad\square$

I Details about experiments

We used the software SageMath [26] that relies on a Python interface. Computations were run in parallel using 64 cores on an academic HPC platform.

First we compared the dimension of the generated Lie algebra $\mathrm{Lie}(V_\phi)(\theta)$ (computed using the algorithm presented in Section 3.3) with $D - N$, where N is the number of independent conserved functions known by the literature (predicted by Proposition 4.1 for ReLU and linear neural networks). We tested both linear and ReLU architectures (with and without biases) of various depths and widths, and observed that the two numbers matched in all our examples.

For this, we draw 50 random ReLU (resp. linear) neural network architectures, with depth drawn uniformly at random between 2 to 5 and i.i.d. layer widths drawn uniformly at random between 2 to 10 (resp. between 2 to 6). For ReLU architectures, the probability to include biases was $1/2$.

Then we checked that all conservation laws can be explicitly computed using the algorithm presented in Section 2.4 and looking for polynomial solutions of degree 2 (as conservation laws already known by the literature are polynomials of degree 2). As expected we found back all known conservation laws by choosing 10 random ReLU (resp. linear) neural network architectures with depth drawn uniformly at random between 2 to 4 and i.i.d. layer widths drawn uniformly at random between 2 to 5.