ON PARAMETER SYMMETRIES AND CONSERVATION LAWS IN GRADIENT FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter space symmetries and conservation laws play an important role in understanding the loss landscapes and implicit biases of neural networks. Inspired by Noether's theorem in physics, prior works have sought to derive conservation laws under gradient flow from parameter symmetries, but the scope and limitations of this connection remain unclear. We develop a unified geometric framework that clarifies the precise relationship between the two notions, including the conditions under which symmetries correspond to conservation laws. We introduce a notion of compositional identifiability and use it to establish a general inheritance principle for complete characterizations of symmetries and conservation laws in multilayer networks. We apply the framework to multi-head and grouped-query attention, polynomial neural networks, and square deep linear networks.

1 INTRODUCTION

Modern deep learning models give rise to highly complex, non-convex training objectives. Optimizing such models requires careful consideration of both the model architecture and the optimization procedure to obtain successful training (Glorot & Bengio, 2010), fast and stable convergence (Santurkar et al., 2018), and to induce favourable implicit biases for generalization (Wang et al., 2021). Understanding these optimization characteristics is thus a fundamental problem in deep learning, and has been the subject of extensive and sustained research (see, e.g., Sun, 2019; Vardi, 2023).

Parameter space symmetries offer one avenue to study these problems, as they characterize the redundancy in the loss landscape (Simsek et al., 2021; Zhao et al., 2025). Alternatively, conservation laws constrain the optimization trajectory from initialization to convergence, providing an explicit characterization of the resulting optimization bias. Inspired by Noether's theorem (Noether, 1918), which roughly states that every continuous symmetry of a physical system induces a conservation law, prior works have developed theoretical frameworks to derive conservation laws from symmetries in neural network training dynamics (Kunin et al., 2021; Głuch & Urbanke, 2021).

However, for gradient flow, the relationship is more nuanced than these frameworks suggest. Recent work raises the question of whether all conserved quantities in neural networks arise from known symmetries (Zhao et al., 2025). Other works point at a more fundamental connection between symmetries and conservation laws through their underlying vector fields (Marcotte et al., 2025; Gegenfurtner et al., 2026). Based on these observations, we develop a geometric framework that precisely characterizes the relationship between symmetries and conservation laws in gradient flow.

1.1 MAIN CONTRIBUTIONS

- We develop a unified vector-field framework for loss symmetries and conservation laws under gradient flow and formalize the gap between them, showing that independent conservation laws induce independent loss symmetries, but not vice versa.
- We connect this framework to functional symmetries and model identifiability, and derive general results on the inheritance and completeness of symmetries and conservation laws in compositional models which can be applied to multilayer networks.
- We apply our framework to establish the complete set of conservation laws for multi-head self-attention, grouped-query attention and deep polynomial neural networks, and provide a bound for the number of independent conservation laws in square deep linear networks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.2 Related works

Parameter symmetry and identifiability. Neural networks often admit transformations of the parameters that leave the represented function unchanged (Sussmann, 1992; Fefferman & Markel, 1993). Some of these parameter symmetries are well known, while others depend more subtly on the architecture and activation function. They arise, e.g., in linear networks (Baldi & Hornik, 1989; Trager et al., 2020), polynomial neural networks (Kileel et al., 2019; Usevich et al., 2025), ReLU networks (Grigsby et al., 2023; Gegenfurtner et al., 2026), networks with batch normalization (Kunin et al., 2021), multi-head self-attention (Tran et al., 2025b), and symmetries induced by particular activation functions (Vlačić & Bölcskei, 2021; Zhao et al., 2023). This has led to new theoretical understandings of optimization landscapes (Ziyin, 2024; Tran et al., 2025a) and practical techniques for improving the convergence and generalization of neural networks (Zhao et al., 2024). For a survey of parameter symmetries in neural networks and their applications, we refer to Zhao et al. (2025).

Conservation laws in gradient flow. Conservation laws first appeared as balancedness conditions in deep linear networks (Saxe et al., 2014) and have been useful in understanding the convergence behavior (Arora et al., 2019a; Du et al., 2018), acceleration characteristics (Arora et al., 2018), and implicit bias of training dynamics (Arora et al., 2019b). Le & Jegelka (2022) derived conservation laws for nonlinear layers, including ResNet and convolutional layers, and used them to explain the low-rank phenomenon in the training of deep feedforward networks. Recent work by Marcotte et al. (2023) introduced a quantitative notion of completeness for conservation laws of a network architecture, with applications to shallow linear and ReLU networks. Subsequent work by Marcotte et al. (2025); Tran et al. (2025b) derived the complete set of conservation laws for further layer types, including single and multi-head self-attention, ResNet, modern activations, and Mixture-of-Experts. These works also highlight the difficulty of extending such results to more complex architectures, such as multilayer networks. Other active topics include conservation laws in other training dynamics (Marcotte et al., 2024), data-dependent conservation laws (Galley et al., 2026), and topology and geometry of the learning space (Nurisso et al., 2024; 2026).

Connection between symmetries and conservation laws. Noether's theorem establishes that every continuous symmetry in a Lagrangian system corresponds to a conservation law (Noether, 1918). Prior works by Kunin et al. (2021); Tanaka & Kunin (2021); Głuch & Urbanke (2021); Zhao et al. (2023) have sought to establish a similar framework in neural network training dynamics, but the constructions have remained limited to specific types of symmetries and do not give a complete characterization of the relationship between symmetries and conservation laws for gradient flows. Several recent works have observed that symmetries of the loss function and conservation laws of the gradient flow are related through their underlying vector fields (Josz, 2025; White, 2025; Marcotte et al., 2025; Gegenfurtner et al., 2026). Our framework builds upon this geometric observation. We summarize and contrast prior works with ours in Appendix B.

2 Preliminaries

Let $\Theta \subset \mathbb{R}^D$ be the parameter space, $\mathcal{X}$ be the input space, and $\mathcal{Z}$ be the output space of a network

$$G : \Theta \times \mathcal{X} \to \mathcal{Z}, \quad (\theta, x) \mapsto G(\theta, x).$$

Let ℓ be a loss function that compares the model output $z := G(\theta, x)$ against the label $y \in \mathcal{Y}$,

$$\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}, \quad (z, y) \mapsto \ell(z, y).$$

Given a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the single-sample loss is $\ell_{(x,y)}(\theta) := \ell(G(\theta, x), y)$. For a dataset $\mathcal{D} = (x^{(i)}, y^{(i)})_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^*$, training seeks to minimize the empirical loss

$$L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(G(\theta, x^{(i)}), y^{(i)}\right) = \frac{1}{n} \sum_{i=1}^{n} \ell_{(x^{(i)}, y^{(i)})}(\theta).$$

In this paper, we specifically consider gradient flow as the training mechanism:

$$\dot{\theta}(t) = -\nabla_\theta L_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell_{(x^{(i)}, y^{(i)})}(\theta), \quad \theta(0) = \theta_0. \tag{1}$$

For this ODE to be well-defined by the Picard-Lindelöf theorem, we assume throughout that Ω is open and, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell_{(x,y)}(\theta)$ is continuously differentiable with locally Lipschitz gradient $\nabla_\theta \ell_{(x,y)}(\theta)$ on Θ. This assumption is satisfied by neural network architectures that use smooth activation functions, such as sigmoid, tanh, GELU, Swish, SwiGLU, etc., and that are trained with smooth loss functions, such as the mean squared error loss and the logistic loss.

Moreover, for common loss functions whose value is bounded from below, the following proposition shows that the gradient flow is well-defined globally, i.e., it exists and is unique for all time $t \in [0, \infty)$. This permits us to talk about gradient flow corresponding to any finite dataset $\mathcal{D}$, any initialization $\theta_0 \in \Theta$, and for all time $t \in [0, \infty)$.

Proposition 1. *Suppose $\Theta = \mathbb{R}^D$, the loss function ℓ is bounded from below, and for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell_{(x,y)}(\theta) \in C^1(\Theta, \mathbb{R})$ with locally lipschitz gradient $\nabla_\theta \ell_{(x,y)}(\theta)$. Then for every finite nonempty dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^*$ and every initialization $\theta_0 \in \Theta$, the gradient flow (1) admits a unique solution $\theta(t)$ defined for all $t \in [0, \infty)$.*

The proof of this result, along with all other proofs in the paper, is deferred to the appendix.

2.1 Conservation laws

As defined by Marcotte et al. (2023), a function $h : \Theta \to \mathbb{R}$ is called a conservation law if, for any solution $\theta(t)$ of the gradient flow equation (1) with respect to any dataset $\mathcal{D}$ and initialization θ_0, the value of h remains constant along the trajectory, i.e., $h(\theta(t)) = h(\theta_0)$ for all t. Corollary 2.6 and Proposition 2.7 from the same paper provide the following equivalent condition, which says that the conservativeness of any function h is completely characterized by its infinitesimal orthogonal alignment with respect to the gradient of the loss function.

Proposition 2. *A continuously differentiable function $h : \Theta \to \mathbb{R}$ is a conservation law if and only if $\nabla_\theta h(\theta) \perp W_\theta$ for all $\theta \in \Theta$, where*

$$W_\theta := \operatorname{span}_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ [D_\theta G(\theta, x)]^\top \nabla_z \ell(G(\theta, x), y) \right\} = \operatorname{span}_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ \nabla \ell_{(x,y)}(\theta) \right\}. \quad (2)$$

We call W the *gradient flow distribution*, which captures the directions in which the parameters can evolve under gradient flow for all possible data pairs (x, y), and hence also the directions in which the parameters can evolve under gradient flow for any dataset $\mathcal{D}$. The terminology *distribution* is used in the sense of differential geometry, where a distribution is an assignment of a subspace of the tangent space at each point on a manifold. In this case, the manifold is the parameter space Θ, and the tangent space at each point θ is simply $\mathbb{R}^D$.

A fundamental question is to characterize all conservation laws of a given network architecture and loss function. To this end, Marcotte et al. (2023) gives a definition of independent conservation laws based on the concept of functional independence (Newns, 1967). We call a set of conservation laws complete if they are functionally independent and there are no other functionally independent conservation laws. We recall the definition of functional independence:

Definition 3. *A family of N functions $\{h_1, \ldots, h_N\}$ in $C^1(\Theta, \mathbb{R})$ is said to be functionally independent if the vectors $\{\nabla_\theta h_1(\theta), \ldots, \nabla_\theta h_N(\theta)\}$ are linearly independent for all $\theta \in \Theta$.*

The definition is justified by Proposition 2.12 of Marcotte et al. (2025), which states that, given a complete set of conservation laws, any other conservation law can locally be expressed as a function of the existing conservation laws. In verifying this result, however, we found that the proof is insufficient for the stated conclusion. Specifically, the proof invokes Theorem 1 of Newns (1967), which does not by itself show that an additional dependent function can be expressed as a function of the prior independent functions. To address this gap, we provide a general local factorization result that captures the intended conclusion.

Proposition 4. *Let $\Theta \subset \mathbb{R}^D$ be open, and let $H = (h_1, \ldots, h_k) : \Theta \to \mathbb{R}^k$, $h : \Theta \to \mathbb{R}$ be C^r functions for $r \geq 1$. Assume $\nabla_\theta h_1, \ldots, \nabla_\theta h_k$ are linearly independent at every $\theta \in \Theta$. Then the following two conditions are equivalent:*

(i) For every $\theta \in \Theta$, $\nabla_\theta h(\theta) \in \operatorname{span}\{\nabla_\theta h_1(\theta), \ldots, \nabla_\theta h_k(\theta)\}$.

(ii) For every $\theta_0 \in \Theta$, there exist neighborhoods $V \ni \theta_0, W \ni H(\theta_0)$ and $f \in C^r(W)$ such that

$$h(\theta) = f(H(\theta)) = f(h_1(\theta), \ldots, h_k(\theta)), \quad \forall \theta \in V. \quad (3)$$

2.2 LOSS SYMMETRIES AND FUNCTIONAL SYMMETRIES

A symmetry of an object is an invertible transformation that leaves the structure of the object unchanged. As defined in the survey by Zhao et al. (2025), a bijective parameter transformation $T : \Theta \to \Theta$ is a loss symmetry if it preserves the value of the loss function L. It is a functional symmetry if it preserves the function represented by the model G. Every functional symmetry is a loss symmetry, but the converse is not true in general.

Since symmetries are invertible transformations, they form a group under composition. Hence, prior works have utilized the language of group actions to study symmetries. Let Γ be a group that acts on Θ via a group action $\psi : \Gamma \times \Theta \to \Theta$ satisfying the group action axioms:

$$\psi(e, \theta) = \theta, \qquad \forall \theta \in \Theta, \text{ where } e \text{ is the identity element in } \Gamma,$$
$$\psi(g, \psi(h, \theta)) = \psi(gh, \theta), \quad \forall g, h \in \Gamma, \forall \theta \in \Theta.$$

For convenience, we also write the group action as $g \cdot \theta := \psi(g, \theta)$.

For a given dataset $\mathcal{D}$, we say that (Γ, ψ) is a symmetry of the loss function $L_{\mathcal{D}}$ if, for all $g \in \Gamma, \theta \in \Theta$,

$$L_{\mathcal{D}}(g, \theta) := L_{\mathcal{D}}(\psi(g, \theta)) = L_{\mathcal{D}}(g \cdot \theta) = L_{\mathcal{D}}(\theta).$$

If the symmetry holds for all datasets $\mathcal{D}$, we simply say that (Γ, ψ) is a symmetry of the loss function L. Similarly, we say that (Γ, ψ) is a symmetry of the model G if, for all $g \in \Gamma, \theta \in \Theta, x \in \mathcal{X}$,

$$\mathcal{G}(g, \theta, x) := G(\psi(g, \theta), x) = G(g \cdot \theta, x) = G(\theta, x).$$

Of particular interest are differentiable symmetries, where Γ is a Lie group and the group action ψ is smooth. If Γ is a one-dimensional connected Lie group, it is well-known that Γ is isomorphic to the real line $\mathbb{R}$ or the circle S^1. But since $\mathbb{R}$ is a universal cover of both groups, we can always lift the group action of Γ to one of $\mathbb{R}$. Thus, it suffices to consider $\Gamma = \mathbb{R}$. In this case, Kunin et al. (2021) derive the following necessary condition for $(\mathbb{R}, \psi)$ to be a differentiable symmetry of $L_{\mathcal{D}}$:

$$\langle \nabla L_{\mathcal{D}}(\theta), D_g \psi(0, \theta) \rangle = 0, \quad \forall \theta \in \Theta. \tag{4}$$

2.3 A GAP BETWEEN DIFFERENTIABLE SYMMETRIES AND CONSERVATION LAWS

The following example illustrates the gap between the number of independent differentiable symmetries and independent conservation laws. Consider a two-layer linear network

$$G : (\mathbb{R}^{2\times 2})^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \quad G((U, V), x) = UV^\top x.$$

By Corollary 4.4 of Marcotte et al. (2023), a complete set of conservation laws for this network is

$$h_1(U, V) = U_{11}^2 + U_{21}^2 - V_{11}^2 - V_{21}^2,$$
$$h_2(U, V) = U_{11}U_{12} + U_{21}U_{22} - V_{11}V_{12} - V_{21}V_{22},$$
$$h_3(U, V) = U_{12}^2 + U_{22}^2 - V_{12}^2 - V_{22}^2.$$

It is also known that for any invertible matrix $S \in \mathbb{R}^{2\times 2}$, the transformation $(U, V) \mapsto (US, V(S^{-1})^\top)$ is a functional symmetry of the network. In particular, the following maps $\psi_i : \mathbb{R} \times (\mathbb{R}^{2\times 2})^2 \to (\mathbb{R}^{2\times 2})^2$, $i = \overline{1, 4}$, give four one-parameter symmetries of the network, whose infinitesimal generators $D_t\psi_i(0, \theta), i = \overline{1, 4}$, are linearly independent vector fields:

$$\psi_1 = \left(U \begin{pmatrix} e^{2t} & 0 \\ 0 & 1 \end{pmatrix}, V \begin{pmatrix} e^{-2t} & 0 \\ 0 & 1 \end{pmatrix}\right), \quad \psi_2 = \left(U \begin{pmatrix} \cosh(t) & \sinh(t) \\ \sinh(t) & \cosh(t) \end{pmatrix}, V \begin{pmatrix} \cosh(t) & -\sinh(t) \\ -\sinh(t) & \cosh(t) \end{pmatrix}\right),$$
$$\psi_3 = \left(U \begin{pmatrix} 1 & 0 \\ 0 & e^{2t} \end{pmatrix}, V \begin{pmatrix} 1 & 0 \\ 0 & e^{-2t} \end{pmatrix}\right), \quad \psi_4 = \left(U \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix}, V \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix}\right).$$

Observe that $\nabla_\theta h_i(\theta) = D_t\psi_i(0, \theta)$ for each $i = 1, 2, 3$. Since these conservation laws are complete, ψ_4 yields no new independent conservation law. This runs counter to a naive analogy with Noether's theorem, which might suggest that every differentiable symmetry gives rise to a corresponding conservation law.

Several works have alluded to this gap in different contexts, including those of Zhao et al. (2023); Gegenfurtner et al. (2026). The recent survey by Zhao et al. (2025) explicitly identifies as an open question whether all conserved quantities originate from known symmetries. We formalize this gap between differentiable symmetries and conservation laws in Section 3.

3 Loss Symmetries, Conservation Laws, and Vector Fields

In this section, we build on standard results from differential geometry to develop a framework to characterize loss symmetries and conservation laws in neural networks under gradient flow via their underlying vector fields. We provide a short review of the relevant concepts in Appendix A for completeness, and refer to standard texts (Olver, 1993; Isidori, 1995; Lee, 2003) for further details.

Our first result establishes a necessary and sufficient condition for a Lie group action to be a differentiable symmetry of a given loss function, thus both extending and generalizing the necessary condition (4) derived by Kunin et al. (2021).

Proposition 5. *Let Γ be a connected Lie group that acts on Θ via a smooth group action ψ. A necessary and sufficient condition for Γ to be a differentiable symmetry of $L_{\mathcal{D}}$ is*

$$D_g\psi(e,\theta)^\top \nabla_\theta L_{\mathcal{D}}(\theta) = 0, \quad \forall \theta \in \Theta. \tag{5}$$

A direct application of Proposition 5 yields Corollary 6 in the case where $\Gamma = \mathbb{R}$. This then further yields Corollary 7 by imposing that the symmetry holds for all datasets $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^*$.

Corollary 6. *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a smooth group action $\psi(t,\theta)$. Let $\mathcal{D}$ be a dataset in $(\mathcal{X} \times \mathcal{Y})^*$. A necessary and sufficient condition for Γ to be a differentiable symmetry of $L_{\mathcal{D}}$ is*

$$\langle D_t\psi(0,\theta), \nabla_\theta L_{\mathcal{D}}(\theta)\rangle = 0, \quad \forall \theta \in \Theta. \tag{6}$$

Corollary 7. *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a smooth group action $\psi(t,\theta)$. A necessary and sufficient condition for Γ to be a differentiable symmetry of L is*

$$D_t\psi(0,\theta) \perp W_\theta, \quad \forall \theta \in \Theta. \tag{7}$$

We remark that the smooth action of any connected Lie group Γ is uniquely determined by the smooth actions of its 1-parameter Lie subgroups (see Appendix C). Thus, 1-parameter symmetries provide the building blocks for higher dimensional symmetries, motivating us to focus on those.

Since Proposition 2 and Corollary 7 are both equivalent conditions of the same form, they together suggest that conservation laws and symmetries can be completely characterized by smooth vector fields that are perpendicular to the gradient flow distribution W, i.e., those in $W^\perp$. Indeed, given enough regularity, we fully describe the connection between vector fields in $W^\perp$ and the corresponding conservation laws and symmetries via Proposition 8 and Proposition 9. Before doing so, we recall several notions from differential geometry. A vector field $\chi = (\chi_1, \ldots, \chi_m) : \Theta \to \mathbb{R}^D$ is said to be closed if its cross partial derivatives are equal, i.e.,

$$\frac{\partial \chi_i}{\partial \theta_j} = \frac{\partial \chi_j}{\partial \theta_i}, \quad \forall 1 \le i, j \le D. \tag{8}$$

It is conservative, or exact, if it is the gradient of some scalar potential function, i.e., $\chi = \nabla h$ for some $h : \Theta \to \mathbb{R}$. A vector field is complete if the flow generated by it exists for all time.

Proposition 8. *If h is a conservation law, then $\chi := \nabla h$ is a closed vector field in $W^\perp$. On the other hand, if χ is a closed vector field in $W^\perp$, then locally around every $\theta \in \Theta$, there exists a unique scalar potential function h such that $\chi = \nabla h$, up to an additive constant. Furthermore, if Θ is a star-shaped domain, then the scalar potential function h exists globally on Θ.*

Proposition 9. *If $(\mathbb{R}, \psi)$ is a symmetry of the loss function L, then $\chi := D_t\psi \mid_{t=0}$ is a smooth vector field in $W^\perp$. On the other hand, if $\chi \in W^\perp$ is a smooth vector field, then there exists a unique maximal local flow $\psi : \mathcal{O} \to \Theta$, where $\mathcal{O} \subset \mathbb{R} \times \Theta$ is the maximal open set containing $\{0\} \times \Theta$, for which the solution of*

$$\frac{d}{dt}\psi(t,\theta) = \chi(\psi(t,\theta)), \quad \psi(0,\theta) = \theta$$

is well-defined. This unique maximal local flow satisfies the following four equations for all $\theta \in \Theta$ and t, s whenever the terms are well-defined:

$$\psi(0,\theta) = \theta, \qquad \psi(t+s,\theta) = \psi(t,\psi(s,\theta)), \qquad \textit{(group action axioms)}$$
$$D_t\psi(0,\theta) = \chi(\theta), \qquad \textit{(infinitesimal generator)}$$
$$L(\psi(t,\theta)) = L(\theta). \qquad \textit{(loss invariance condition)}$$

Furthermore, if χ is complete, then ψ is a symmetry of the loss function L.

The maximal local flow ψ as defined in Proposition 9 is known as a partial group action (Cheeger & Gromov, 1986; 1990; Batista, 2017), meaning that its domain is a maximal neighborhood of $\{0\} \times \Theta$, and the group action axioms are satisfied wherever it is defined. Hence, we call $(\mathbb{R}, \psi)$ a partial symmetry of the loss function L. Since the infinitesimal symmetry condition 7 is a local condition at $t = 0$, Corollary 6, Corollary 7 and Proposition 9 still hold verbatim when applied to partial symmetries (see Appendix E.6 for details). The results can be summarized as follows:

- There exists a one-to-one correspondence between vector fields in $W^\perp$ and 1-parameter partial symmetries of the loss function L. The partial symmetry is a global symmetry precisely when its generating vector field is complete.
- There exists a one-to-one correspondence between closed vector fields in $W^\perp$ and local conservation laws of the gradient flow dynamics, up to an additive constant. The local conservation law is a global one when the parameter space Θ is star-shaped.

We term $W^\perp$ the *symmetry distribution*, as it captures the directions in which the parameters can be transformed without changing the loss value. The following result gives the exact correspondence between symmetries and conservation laws.

Corollary 10. *Let $h : \Theta \to \mathbb{R}$ be a conservation law of the gradient flow dynamics. There exists a unique partial symmetry $(\mathbb{R}, \psi)$ of the loss function L such that*

$$D_t\psi(0, \theta) = \nabla h(\theta), \quad \forall \theta.$$

Conversely, let $(\mathbb{R}, \psi)$ be a partial symmetry of the loss function L. If $D_t\psi(0, \theta)$ is closed, then locally, there exists a unique conservation law h, up to an additive constant, such that

$$\nabla h(\theta) = D_t\psi(0, \theta), \quad \forall \theta.$$

The additional closedness assumption above suggests that there are more symmetries than conservation laws. To formalize this intuition, we introduce the following notions of infinitesimally independent and complete partial symmetries. The idea is to define a set of partial symmetries such that their infinitesimal actions span the symmetry distribution $W^\perp$.

Definition 11. *A family of r partial symmetries $\{\psi_1, \ldots, \psi_r\}$ of $\mathbb{R}$ onto Θ is said to be infinitesimally independent if the vectors $\{D_t\psi_1(0, \xi), \ldots, D_t\psi_r(0, \xi)\}$ are linearly independent for all $\theta \in \Theta$. They are called infinitesimally complete if they are independent and their infinitesimal actions span the symmetry directions near θ.*

We recall Theorem 3.3 of Marcotte et al. (2023), which states that if the dimension of the Lie completion of the reparametrized gradient flow distribution $\dim(\mathrm{Lie}(W_\phi)(\theta))$ is locally constant, then there are exactly $D - \dim(\mathrm{Lie}(W_\phi)(\theta))$ independent conservation laws in a neighborhood of θ. The following theorem generalizes this result by removing the need for a nice reparametrization of the parameters, as well as extending it to account for symmetries of the loss function on the regular rank strata of the gradient flow distribution.

Theorem 12. *Assume that the gradient flow distribution W is a smooth distribution of constant rank r, and the Lie completion $\mathrm{Lie}(W)$ is a smooth distribution of constant rank $\overline{r}$ in Θ. Then, for every $\theta \in \Theta$, there exists a neighborhood $U \ni \theta$ such that the following statements hold:*

(i) There exists $s := D - r$ partial symmetries of the loss function L that are independent in U, and they are complete.

(ii) There exist $c := D - \overline{r}$ independent conservation laws of the gradient flow dynamics that are independent in U, and they are complete.

(iii) The symmetry-conservation law gap is $s - c := \overline{r} - r \geq 0$. Equality holds if and only if W is involutive, i.e., $\mathrm{Lie}(W) = W$.

One consequence of Theorem 12 is that, if W is not involutive, then the conservation laws at inital-ization can not determine the specific local minimum that the gradient flow dynamics might converge to. Nevertheless, they still provide an initialization-dependent implicit bias that can be used to determine important properties of this minimum. We demonstrate these ideas for a two-layer linear network in the simple regression task of learning a scaling function in Appendix H, where we show that the singular values and condition number of the learned weight matrices are determined by the conservation laws at initialization.

4 INHERITANCE PRINCIPLE FOR SYMMETRIES AND CONSERVATION LAWS

While prior work has made significant progress in understanding the symmetries and conservation laws in individual layers, a fundamental challenge has been to understand how these are inherited by a larger network that contains the subnetwork as a component, and whether the inherited symmetries and conservation laws are complete. In this section, we connect the framework developed in Section 3 to functional symmetry and parameter identifiability, and obtain general inheritance and completeness results for symmetries and conservation laws of multilayer networks.

Recall that every functional symmetry of the model G is also a symmetry of the loss function L. If the loss function L is sufficiently well-behaved, then the converse is also true, i.e., every loss symmetry of L is also a functional symmetry of G. One way to formalize this notion is through the following Assumption 13, which we term the *separability of predictions* property. This is a natural condition that is satisfied by many loss functions (see Appendix F.1). Based on this, we obtain Proposition 14, which allows us to utilize the framework developed in the previous section to obtain our inheritance and completeness results.

Assumption 13 (Separability of predictions)**.** *The loss function $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ satisfies*

$$\ell(z, y) = \ell(z', y), \forall y \in \mathcal{Y} \implies z = z'.$$

Proposition 14. *If the loss function ℓ satisfies Assumption 13, then any partial symmetry of the loss function L is also a partial functional symmetry of the model G.*

Our first result, Proposition 15, establishes a general inheritance result for symmetries and conservation laws of subnetworks embedded in a larger network. Proposition J.3 by Marcotte et al. (2025) formulated an inheritance result for conservation laws arising from invariant transformations for serial compositions. Building on the same underlying principle, we formulate our result directly in terms of preservation of functional equivalence. Our result naturally covers a broad class of compositional structures, including serial composition, sum, concatenation, Hadamard product, outer product, graph aggregation, etc. of subnetworks (see Table 1 in Appendix F.3 for further details).

Proposition 15 (Inheritance from subnetworks I)**.** *Let $g : \Theta \times \mathcal{X}^g \to \mathcal{Z}^g$ and $G : (\Theta \times \Omega) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$g(\theta, \cdot) = g(\theta', \cdot) \quad \implies \quad G((\theta, \omega), \cdot) = G((\theta', \omega), \cdot) \qquad \text{for every } \omega \in \Omega. \tag{9}$$

Every partial symmetry $(\mathbb{R}, \psi)$ or conservation law h of g for a loss satisfying Assumption 13 extends to a partial symmetry $(\mathbb{R}, \tilde{\psi})$ or conservation law $\tilde{h}$ of G for any loss function via

$$\tilde{\psi}\,(t, (\theta, \omega)) = (\psi(t, \theta), \omega), \qquad \text{or} \qquad \tilde{h}(\theta, \omega) = h(\theta). \tag{10}$$

Our second result, Proposition 16, shows that if the functional equivalence of the larger network forces the functional equivalence of its subnetworks, then the symmetries and conservation laws of the larger network must be infinitesimally generated by the natural extensions of the symmetries and conservation laws of its subnetworks.

Proposition 16 (Inheritance from subnetworks II)**.** *Let $g^j : \Theta^j \times \mathcal{X}^{g^j} \to \mathcal{Z}^{g^j}, j = 1, 2$ and $G : (\Theta^1 \times \Theta^2) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$[G((\theta_1, \theta_2), \cdot) = G((\theta_1', \theta_2'), \cdot)] \implies [g^j(\theta_j, \cdot) = g^j(\theta_j', \cdot), \forall j = 1, 2]. \tag{11}$$

For $j = 1, 2$, let $(\mathbb{R}, \psi_i^j)_{i=1}^{s_j}$ be a complete set of partial functional symmetries of g^j, and let $h_1^j, \ldots, h_{c_j}^j$ be a complete set of conservation laws of g^j w.r.t. a loss satisfying Assumption 13.

Every partial symmetry $(\mathbb{R}, \psi)$ or conservation law h of G with respect to a loss satisfying Assumption 13 is infinitesimally generated by the extensions

$$\tilde{\psi}_i^1(t, (\theta_1, \theta_2)) = (\psi_i^1(t, \theta_1), \theta_2), \quad \tilde{\psi}_i^2(t, (\theta_1, \theta_2)) = (\theta_1, \psi_i^2(t, \theta_2)), \quad \text{or} \quad \tilde{h}_i^j(\theta_1, \theta_2) = h_i^j(\theta_j),$$

respectively. The extensions themselves need not be partial symmetries nor conservation laws of G.

Finally, by combining the previous two results, we obtain a completeness result for symmetries and conservation laws of multilayer networks.

Theorem 17 (Inheritance from subnetworks III). *Let $g^j : \Theta^j \times \mathcal{X}^{g^j} \to \mathcal{Z}^{g^j}, j = 1, \ldots, k$ and $G : (\Theta^1 \times \ldots \times \Theta^k) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$[G((\theta_1, \ldots, \theta_k), \cdot) = G((\theta'_1, \ldots, \theta'_k), \cdot)] \iff [g^j(\theta_j, \cdot) = g^j(\theta'_j, \cdot), \forall j = 1, \ldots, k]. \tag{12}$$

For $j = \overline{1,k}$, let $(\mathbb{R}, \psi_i^j)_{i=1}^{s_j}$ be a complete set of partial functional symmetries of g^j, and let $h_1^j, \ldots, h_{c_j}^j$ be a complete set of conservation laws of g^j w.r.t. a loss satisfying Assumption 13.

Assume that the loss used for the larger network G also satisfies Assumption 13. A complete set of partial functional symmetries of G is given by the extensions

$$\tilde{\psi}_i^j(t, (\theta_1, \ldots, \theta_k)) = (\theta_1, \ldots, \psi_i^j(t, \theta_j), \ldots, \theta_k), \qquad j = \overline{1,k},\ i = \overline{1, s_j},$$

and a complete set of conservation laws of G is given by

$$\tilde{h}_i^j(\theta_1, \ldots, \theta_k) = h_i^j(\theta_j), \qquad j = \overline{1,k},\ i = \overline{1, c_j}.$$

We call equation 12 the *compositional identifiability* condition. The implication is that if the larger network is compositionally identifiable, then all of its symmetries and conservation laws are inherited from its component subnetworks.

5 Application to Neural Networks

We demonstrate the utility of our results by applying them to several neural network architectures.

The first is grouped query attention (GQA) (Ainslie et al., 2023), which is widely used in modern transformer architectures for fast inference with KV caching. Marcotte et al. (2025) posed completeness of the conservation laws of the standard multi-head self-attention (MHSA) mechanism (Vaswani et al., 2023) as an open problem. Later, Tran et al. (2026) provided a formulation and proof of this result. However, the proof does not provide an explicit justification that, near regular points, every solution of the derived PDE system is locally a function of the known conservation laws. Thus, the argument as written leaves the local completeness step unestablished.

Using our framework, we provide a proof for the complete characterization of the symmetries as well as conservation laws of the GQA mechanism in Proposition 18, which implies completeness for the MHSA mechanism as a special case. The proof is short and instructive, relying on results on completeness for two-layer linear networks (Marcotte et al., 2023) and identifiability of attention heads (Tran et al., 2025b).

Proposition 18 (Grouped query attention). *Let $n_H = kn_G$ and $D_h \leq D$ be positive integers. Consider the grouped query attention mechanism G with n_H heads and n_G groups with parameters*

$$\theta = (Q_1^j, \ldots, Q_k^j, K^j, V^j, O_1^j, \ldots, O_k^j)_{j=1}^{n_G} \in \left(\left(\mathbb{R}^{D \times D_h}\right)^k \times \mathbb{R}^{D \times D_h} \times \mathbb{R}^{D \times D_h} \times \left(\mathbb{R}^{D \times D_h}\right)^k \right)^{n_G},$$

where for any $X \in \cup_{L=1}^{\infty} \mathbb{R}^{L \times D}$,

$$G(\theta, X) = \sum_{j=1}^{n_G} \sum_{i=1}^{k} \mathrm{softmax}\left(\frac{X Q_i^j (K^j)^\top X^\top}{\sqrt{D_h}} \right) X V^j (O_i^j)^\top.$$

Let θ_0 be a point in the parameter space so that Q_i^j, K^j, V^j, O_i^j are full rank and the matrices $Q_i^j (K^j)^\top$ are pairwise distinct for $j = \overline{1, n_G}, i = \overline{1,k}$. Locally around θ_0, the following hold:

- *All functionally equivalent parameters of the grouped query attention mechanism G are given by the following partial group action of $(\mathrm{GL}_{D_h}(\mathbb{R}) \times \mathrm{GL}_{D_h}(\mathbb{R}))^{n_G}$:*

  $$(S^j, T^j)_{j=1}^{n_G} \cdot (Q_1^j, \ldots, Q_k^j, K^j, V^j, O_1^j, \ldots, O_k^j)_{j=1}^{n_G} = (\tilde{Q}_1^j, \ldots, \tilde{Q}_k^j, \tilde{K}^j, \tilde{V}^j, \tilde{O}_1^j, \ldots, \tilde{O}_k^j)_{j=1}^{n_G},$$

  where

  $$\tilde{Q}_i^j = Q_i^j S^j, \quad \tilde{K}^j = K^j (S^j)^{-\top}, \quad \tilde{V}^j = V^j T^j, \quad \tilde{O}_i^j = O_i^j (T^j)^{-\top}.$$

- *The upper triangular elements of the following matrix functions form a complete set of conservation laws with respect to any loss satisfying Assumption 13:*

$$\sum_{i=1}^{k}(Q_i^j)^\top Q_i^j - (K^j)^\top K^j, \quad (V^j)^\top V^j - \sum_{i=1}^{k}(O_i^j)^\top O_i^j, \quad j \in \overline{1, n_G}.$$

The case $n_H = n_G$ corresponds to standard multi-head self-attention, and all results still hold.

In Appendix I, we examine the conservation laws derived in Proposition 18 and their drift during prolonged constant step size SGD in a language model. We observe that approximate conservation can coexist with learning, as reflected in reductions in training and validation loss. However, we also observe that conservation law preservation for certain parameter blocks may deteriorate during prolonged SGD.

Next, we consider polynomial neural networks (PNNs). Recently, Usevich et al. (2025) showed that a wide variety of polynomial networks satisfy a finite-to-one identifiability condition. Our result relies on showing that for these networks, the symmetry-conservation law gap is zero, allowing for a direct transfer of symmetries to conservation laws. This is the first deep network architecture that we know of for which a complete characterization of the conservation laws has been obtained.

Proposition 19 (Polynomial neural network)**.** *Consider the polynomial neural network (PNN) mechanism G with parameters $\theta = (W_j, b_j)_{j=1}^L \in \prod_{j=1}^L \left(\mathbb{R}^{D_j \times D_{j-1}} \times \mathbb{R}^{D_j}\right)$ defined by*

$$G(\theta, x) = f_L \circ \rho_{r_{L-1}} \circ f_{L-1} \circ \cdots \circ \rho_{r_1} \circ f_1(x),$$

where $\rho_{r_j}(z) = (z_1^{r_j}, \ldots, z_{D_j}^{r_j})$ is a monomial activation function of degree r_j and $f_j(z) = W_j z + b_j$ for $j = 1, \ldots, L$. If the architecture is finite-to-one at a generic θ_0, then locally around θ_0, a complete set of conservation laws of G with respect to any loss satisfying Assumption 13 is given by

$$h_i^j(\theta) = \|(W_j)_{i,:}\|_2^2 + (b_j)_i^2 - r_j\|(W_{j+1})_{:,i}\|_2^2, \quad i = \overline{1, D_j}, j = \overline{1, L-1}.$$

In the deep square linear network case, Lindsey & Menon (2026) recently characterized the symmetries of the model on the full-rank strata of the parameter space. Theorem 12 allows us to obtain an upper bound of $(L-1)D^2$ independent conservation laws of the model, which improves slightly upon the obvious bound of LD^2. More generally, this illustrates that a complete characterization of the symmetries of a model automatically yields an upper bound on the number of independent conservation laws, which may be useful for future work on other architectures.

Proposition 20 (Deep square linear network)**.** *Consider the deep square linear network mechanism G with parameters $\theta = (W_j)_{j=1}^L \in \prod_{j=1}^L \mathbb{R}^{D \times D}$ defined by*

$$G(\theta, x) = W_L W_{L-1} \ldots W_1 x.$$

If $W_L \ldots W_1$ is full rank, then locally around $(W_1, \ldots, W_L)$, there are at most $(L-1)D^2$ independent conservation laws of G with respect to any loss satisfying Assumption 13.

Finally, we observe that Proposition 15 readily recovers known conservation laws derived by prior symmetry-based frameworks. For example, if trainable weight matrices A, B enter a multilayer network only through their product AB, as in the popular low-rank adaptation (LoRA) scheme (Hu et al., 2022), then it automatically inherits conservation laws given by the entries of $A^\top A - BB^\top$. We refer the reader to Appendix B.2 for more details.

6 Conclusion

We develop a unified geometric framework for studying symmetries and conservation laws in neural network gradient flow. It clarifies the gap between symmetries and conservation laws, establishes results on their inheritance from subnetworks and their completeness in larger networks, and characterizes the complete set of conservation laws for several architectures. A limitation is that compositional identifiability does not hold for matrix multiplication and thus our inheritance result cannot be applied to deep linear networks. Future work could pursue inheritance under weaker compositional identifiability conditions. Other promising directions include extending our results to non-gradient-flow and data dependent dynamics, and investigation of conservation-laws informed initialization and training.

AI USE STATEMENT

The theoretical framework and essential proof arguments were developed by the authors without generative AI assistance.

We used generative AI tools to help formulate mathematical claims, propose or refine hypotheses, provide feedback on research methodology and experiments, assist with implementation and interpretation of experimental results, and suggest experimental parameters. We also used these tools to create or edit code and figures, identify and summarize relevant literature, brainstorm research directions, and improve the readability of the manuscript.

We did not use generative AI to develop theoretical models or conceptual frameworks, supply critical ingredients for proofs, generate synthetic datasets, assist in the writing of proofs, clean or reformat datasets, or translate text. The authors reviewed all AI-assisted material, independently checked the mathematical content and cited literature, and verified and tested AI-assisted code. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work is theoretical in nature and does not involve human subjects or the collection or release of datasets. We are not aware of any specific ethical concerns arising from the results presented here.

REPRODUCIBILITY STATEMENT

The detailed assumptions and proofs for all theoretical results are provided in the main text and appendix. Code for reproducing the experiments is included in the supplementary material.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=hmOwOZWzYE.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/arora18a.html.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=SkMQg3C5K7.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit regularization in deep matrix factorization*. Curran Associates Inc., Red Hook, NY, USA, 2019b.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989. ISSN 0893-6080. URL https://www.sciencedirect.com/science/article/pii/0893608089900142.

Eliezer Batista. Partial actions: what they are and why we care. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, 24(1):35 – 71, 2017. doi: 10.36045/bbms/1489888814. URL https://doi.org/10.36045/bbms/1489888814.

Jeff Cheeger and Mikhael Gromov. Collapsing Riemannian manifolds while keeping their curvature bounded. I. *Journal of differential geometry*, 1986. ISSN 0022-040X.

Jeff Cheeger and Mikhael Gromov. Collapsing Riemannian manifolds while keeping their curvature bounded. II. *Journal of Differential Geometry*, 32(1):269 – 298, 1990. doi: 10.4310/jdg/1214445047. URL https://doi.org/10.4310/jdg/1214445047.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: layers are automatically balanced. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 382–393, Red Hook, NY, USA, 2018. Curran Associates Inc.

Charles Fefferman and Scott Markel. Recovering a feed-forward net from its output. In J. Cowan, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/e49b8b4053df9505e1f48c3a701c0682-Paper.pdf.

William Fulton and Joe Harris. *Representation Theory: A First Course*, volume 129 of *Graduate Texts in Mathematics*. Springer New York, New York, 1991.

Jakob Galley, Vahid Shahverdi, and Axel Flinth. Conservation laws from data symmetry in neural networks, 2026. URL https://arxiv.org/abs/2606.10913.

Johanna Marie Gegenfurtner, Moritz Grillo, and Guido Montúfar. The symmetries of three-layer relu networks, 2026. URL https://arxiv.org/abs/2605.18319.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272. JMLR.org, 2017.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of ReLU networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11734–11760. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

Grzegorz Głuch and Rüdiger Urbanke. Noether: The more things change, the more stay the same, 2021. URL https://arxiv.org/abs/2104.05508.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 11 1997. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alberto Isidori. *Nonlinear Control Systems*. Communications and Control Engineering. Springer London, London, 3rd edition, 1995. ISBN 978-3-540-19916-8. doi: 10.1007/978-1-84628-615-5.

Cédric Josz. Subdifferentiation with symmetry, 2025. URL https://arxiv.org/abs/2509.11422.

Andrej Karpathy. Tiny Shakespeare dataset. Dataset in the char-rnn GitHub repository, 2015. URL https://github.com/karpathy/char-rnn/blob/master/data/tinyshakespeare/input.txt.

Andrej Karpathy. nanoGPT. GitHub repository, 2022. URL https://github.com/karpathy/nanoGPT.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a0dc078ca0d99b5ebb465a9f1cad54ba-Paper.pdf.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=q8qLAbQBupm.

Thien Le and Stefanie Jegelka. Training invariances and the low-rank phenomenon: beyond linear networks, 2022. URL https://arxiv.org/abs/2201.11968.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003. ISBN 9780387954486. URL https://books.google.com/books?id=eqfgZtjQceYC.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Kathryn Lindsey and Govind Menon. Regularization implies balancedness in the deep linear network, 2026. URL https://arxiv.org/abs/2511.01137.

Sibylle Marcotte, Remi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63210–63221. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c7bee9b76be21146fd592fc2b46614d5-Paper-Conference.pdf.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Keep the momentum: Conservation laws beyond Euclidean gradient flows. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34790–34821. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/marcotte24a.html.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or conservative? Conservation laws for ResNets and transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43140–43176. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/marcotte25a.html.

W. F. Newns. Functional dependence. *The American Mathematical Monthly*, 74(8):911–920, 1967. URL http://www.jstor.org/stable/2315264.

E. Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918:235–257, 1918. URL http://eudml.org/doc/59024.

Marco Nurisso, Pierrick Leroy, and Francesco Vaccarino. Topological obstruction to the training of shallow ReLU neural networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35358–35387. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3e356b84727460df86c8d0eeff1ff85d-Paper-Conference.pdf.

Marco Nurisso, Pierrick Leroy, Giovanni Petri, and Francesco Vaccarino. Topology and geometry of the learning space of reLU networks: connectivity and singularities. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O4Oy7NsSwG.

Peter Olver. *Applications of Lie Groups to Differential Equations*, volume 107 of *Graduate Texts in Mathematics*. Springer New York, 2 edition, 1993. doi: 10.1007/978-1-4612-4350-2.

R. Piziak and P. L. Odell. Full rank factorization of matrices. *Mathematics Magazine*, 72(3):193–201, 1999. URL http://www.jstor.org/stable/2690882.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. URL https://api.semanticscholar.org/CorpusID:205001834.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Mądry. How does batch normalization help optimization? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2488–2498, Red Hook, NY, USA, 2018. Curran Associates Inc.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://openreview.net/forum?id=_wzZwKpTDF_9C.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/simsek21a.html.

Ruoyu Sun. Optimization for deep learning: theory and algorithms, 2019. URL https://arxiv.org/abs/1912.08957.

Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992. URL https://www.sciencedirect.com/science/article/pii/S0893608005800371.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. URL http://arxiv.org/abs/1409.4842.

Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25646–25660. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/d76d8deea9c19cc9aaf2237d2bf2f785-Paper.pdf.

Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and spurious critical points: a geometric study of linear networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgOlCVYvB.

Viet-Hoang Tran, Van-Hoan Trinh, Khanh Vinh Bui, and Tan Minh Nguyen. On linear mode connectivity of mixture-of-experts architectures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=RF3miSqdXa.

Viet-Hoang Tran, Thieu Vo, An Nguyen, Tho-Huu Tran, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Nguyen. Equivariant neural functional networks for transformers. In *International Conference on Learning Representations*, volume 2025, pp. 16270–16341, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29962c2c9daf1fbd92530a7c958dfc2b-Paper-Conference.pdf.

Viet-Hoang Tran, Vinh Khanh Bui, Tan Lai Ngoc, Nam Nguyen, Tuan Quang Dam, and Tan Minh Nguyen. Conservation laws for modern neural architectures. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ay4Q69fAJL.

L.W. Tu. *An Introduction to Manifolds*. Universitext. Springer Science+Business Media, LLC, 2010. ISBN 9781441974006.

Konstantin Usevich, Ricardo Augusto Borsoi, Clara Dérand, and Marianne Clausel. Identifiability of deep polynomial neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MrUsZfQ9pC.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93, May 2023. URL https://doi.org/10.1145/3571070.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021. URL https://www.sciencedirect.com/science/article/pii/S0001870820305132.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

Nicholas C. White. Beyond Lagrangians: Noether's theorem in gradient flow PDEs. *Phys. Rev. E*, 112:064214, Dec 2025. URL https://link.aps.org/doi/10.1103/5jbb-df5h.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113 (47):E7351–E7358, 2016. URL https://www.pnas.org/doi/abs/10.1073/pnas.1614734113.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9ZpciCOunFb.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L0r0GphlIL.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=jLpWq5QY6I.

Liu Ziyin. Symmetry induces structure and constraint of learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62847–62866. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ziyin24a.html.

A Background on Vector Fields

We present a brief overview of the necessary background on vector fields and differential geometry based on Chapter 1 of Isidori (1995). For a more comprehensive treatment, we refer the reader to Isidori (1995); Olver (1993); Lee (2003).

We consider functions on an open subset $U \subset \mathbb{R}^D$, where U can be taken as $\mathbb{R}^D$ itself in the globally smooth regime. A *vector field* is a smooth function $f : U \to \mathbb{R}^D$ that associates to each $\theta \in U$ a direction in $\mathbb{R}^D$. A *covector field* is a smooth function $w^* : U \to (\mathbb{R}^D)^*$, where we identify the dual space $(\mathbb{R}^D)^* = \{T : \mathbb{R}^D \to \mathbb{R}, T \text{ linear}\}$ with the space of row vectors $\mathbb{R}^{1\times D}$. This makes it convenient to write

$$w^*(f) = w^* f = \langle w, f\rangle, \quad w = (w^*)^\top : U \to \mathbb{R}^D.$$

Let $\lambda : U \to \mathbb{R}$ be a smooth function, then its derivative is the covector field

$$D\lambda(\theta) = \left(\frac{\partial\lambda}{\partial\theta_1}, \ldots, \frac{\partial\lambda}{\partial\theta_n}\right).$$

Given a vector field f, the *derivative of λ along f* is written as $L_f\lambda(\theta)$, and is defined to be the inner product between $\nabla\lambda = (D\lambda)^\top$ with f

$$L_f\lambda(\theta) = D\lambda(\theta)f(\theta) = \langle\nabla\lambda(\theta), f(\theta)\rangle = \sum_{i=1}^{n}\frac{\partial\lambda}{\partial\theta_i}f_i(\theta).$$

Given two prior vector fields f, g. The *Lie bracket* $[f, g]$ is a vector field satisfying

$$[f, g](\theta) = Dg(\theta)\, f(\theta) - Df(\theta)\, g(\theta).$$

The Lie bracket is bilinear over $\mathbb{R}$, skew commutative, and satisfies the Jacobi identity.

Let $f_1, \ldots, f_d$ be smooth vector fields. These identify a unique *smooth distribution* given by

$$\Delta(\theta) = \operatorname{span}\{f_1(\theta), \ldots, f_d(\theta)\},$$

which associates each point $\theta \in U$ with a subspace of $\mathbb{R}^n$ spanned by all f_i at θ. The term 'smooth' here refers to the fact that it can be spanned by smooth vector fields. A vector field f belongs to a distribution Δ if this is true pointwise. The dimension of a distribution Δ at $\theta \in U$ is $\dim\Delta(\theta)$. If F is a smooth matrix with n rows then the distribution spanned by its columns is

$$\Delta(\theta) = \operatorname{Im}(F(\theta)), \quad \dim\Delta(\theta) = \operatorname{rank} F(\theta).$$

A distribution Δ is called *nonsingular* if its dimension $\dim\Delta(\theta) \equiv d$ is constant on U. A point θ^0 is said to be a *regular point* if there exists a neighborhood U^0 on which Δ is nonsingular. If Δ is a smooth, nonsingular distribution of dimension d then it has a spanning set of vector fields $f_1, \ldots, f_d$ such that $\Delta(\theta) = \operatorname{span}\{f_i(\theta)\}$ and any smooth vector field $\tau \in \Delta$ can be expressed as $\tau(\theta) = \sum c_i(\theta)f_i(\theta)$ for smooth real-valued functions c_i. A distribution Δ is *involutive* if it is closed under taking the Lie bracket, i.e., $[\tau_1, \tau_2] \in \Delta, \forall\tau_1, \tau_2 \in \Delta$.

Similar to a distribution, given smooth covector fields $w_1^*, \ldots, w_d^*$, we can define the *smooth codistritbution* $\Theta = \operatorname{span}(w_1^*, \ldots, w_d^*)$. Similar constructions to a distribution applies.

There is a natural way to construct a codistribution starting from a distribution Δ. For each θ, let

$$\Delta^\perp(\theta) = \{w^* \in (\mathbb{R}^n)^* : w^* v = 0, \forall v \in \Delta(\theta)\}$$

be the annihilator of $\Delta(\theta)$. This defines a codistribution called the *annihilator* of Δ. Conversely, given a codistribution Θ, one can similarly construct a distribution $\Theta^\perp$, called the annihilator of Θ. By construction, the pointwise sum of dimensions of Δ and $\Delta^\perp$ is n.

If a distribution Δ is spanned by the columns of a smooth matrix F, then its annihilator is identified by

$$\Delta^\perp(\theta) = \{w^* \in (\mathbb{R}^n)^* : w^* F(\theta) = 0\}.$$

Conversely, if a codistribution Θ is spanned by the rows of a smooth matrix W then its annihilator is identified by

$$\Theta^\perp(\theta) = \{v \in \mathbb{R}^n : W(\theta)v = 0.\}$$

Thus, $\Theta^{\perp}(\theta)$ is the kernel of the matrix $W(\theta)$.

If θ is a regular point of a smooth distribution Δ, then it is also a regular point of $\Delta^{\perp}$. Consider a nonsingular distribution Δ with dimension d and basis vector fields $f_1, \ldots, f_d$. We know that its codistribution $\Delta^{\perp}$ is smooth, nonsingular, and has a basis $w_1^*, \ldots, w_{n-d}^*$. By construction, these solve the equation

$$w_j^*(\theta)F(\theta) = 0, \quad F(\theta) = (f_1(\theta), \ldots, f_d(\theta))$$

for all θ. We call Δ *completely integrable* if we can choose $w_1^*, \ldots, w_{n-d}^*$ such that there exists real-valued smooth functions $\ell_1, \ldots, \ell_{n-d}$ such that $D\ell_i = w_i^*, \forall i$.

Theorem 21 (Frobenius)**.** *A nonsingular distribution is completely integrable if and only if it is involutive.*

B Comparison With Prior Works

B.1 A typical derviation of Noether's theorem

Classical Noether theory establishs a connection between symmetries of the Lagrangian and conseration laws of the dynamical system induced by the Lagrangian. To illustrate this, we briefly review a typical derivation of Noether's theorem in the context of classical mechanics. Consider the space $\mathbb{R}^n$ equipped with the Lagrangian

$$L : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \quad (x, v, t) \mapsto L(x, v, t),$$

which defines the action functional

$$S[\gamma] = \int_a^b L(\gamma(t), \gamma'(t), t)dt$$

for any path $\gamma : [a, b] \to \mathbb{R}^n$. Let q be an optimal path that minimizes $S[\gamma]$. Then, via calculus of variations, the optimal path q satisfies the Euler–Lagrange equations

$$\nabla_x L - \frac{d}{dt}\nabla_v L = 0.$$

Let $\psi : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ be a symmetry of the Lagrangian L and let $\chi : \mathbb{R}^n \to \mathbb{R}^n$ be the corresponding infinitesimal generator, i.e.,

$$L(\psi(g, q), D_q\psi(g, q)\,\dot{q}, t) = L(q, \dot{q}, t), \quad \chi(q) = \frac{d}{dg}\psi(g, q)\big|_{g=0},$$

then taking the derivative with respect to g and evaluating at $g = 0$ gives the infinitesimal symmetry condition

$$\langle \nabla_x L, \chi\rangle + \langle \nabla_v L, D_q\chi\,\dot{q}\rangle = 0.$$

If we take the derivative of

$$\lambda = \langle \nabla_v L, \chi\rangle.$$

with respect to time t then we obtain

$$\begin{aligned}\frac{d}{dt}\lambda(t) &= \left\langle \frac{d}{dt}\nabla_v L, \chi \right\rangle + \langle \nabla_v L, D_q\chi\,\dot{q}\rangle \\ &= \underbrace{\left\langle \frac{d}{dt}\nabla_v L - \nabla_x L, \chi \right\rangle}_{=0 \text{ by Euler–Lagrange equations}} + \underbrace{\langle \nabla_x L, \chi\rangle + \langle \nabla_v L, D_q\chi\,\dot{q}\rangle}_{=0 \text{ by the infinitesimal symmetry condition}} = 0\end{aligned}$$

Thus, we posit that λ is a conserved quantity along the optimal path q. We call this term the Noether charge associated with the symmetry ψ. Noether's theorem states that every symmetry of the Lagrangian induces a conservation law taking the form of the Noether charge $\lambda(t)$.

In the context of gradient flow, it is important to note that the dynamics of the system are not derived from a Lagrangian. The gradient flow equation $\dot{\theta} = -\nabla f(\theta)$ directly specifies the dynamics of the system, and thus Noether's theorem does not directly apply.

B.2 Comparison with other frameworks connecting symmetries and conservation laws under gradient flow

Inspired by Noether's theorem, prior frameworks that connect symmetries and conservation laws have tried to adapt the derivation of Noether's theorem to the context of gradient flow. Their common approach is to try to integrate the infinitesimal symmetry condition (4) to obtain a constant function with respect to time t along the gradient flow. We recall this condition here for convenience:

$$\langle \nabla_\theta L_{\mathcal{D}}(\theta), D_g\psi(0,\theta)\rangle = 0.$$

If this can be factored into a function of θ independent of t, then that function is a conserved quantity along the gradient flow. We now discuss each of these prior works in turn.

Kunin et al. (2021) notice that if ψ is a symmetry of the gradient flow $\dot\theta = -\nabla_\theta L_{\mathcal{D}}(\theta)$ and $\chi = D_g\psi(0,\theta)$ is its underlying vector field, then

$$\begin{aligned}\frac{d}{dt}\langle\theta,\chi(\theta)\rangle &= \frac{d}{dt}\langle\theta, D_g\psi(0,\theta)\rangle \\ &= \langle\dot\theta, D_g\psi(0,\theta)\rangle + \langle\theta, D_tD_g\psi(0,\theta)\,\dot\theta\rangle \\ &= \underbrace{-\langle\nabla_\theta L_{\mathcal{D}}(\theta), D_g\psi(0,\theta)\rangle}_{=0 \text{ by the infinitesimal invariance condition 4}} - \langle\theta, D_tD_g\psi(0,\theta)\,\nabla_\theta L_{\mathcal{D}}(\theta)\rangle.\end{aligned}$$

Hence, if the additional condition

$$\langle\theta, D_tD_g\psi(0,\theta)\,\nabla_\theta L_{\mathcal{D}}(\theta)\rangle = 0 \tag{13}$$

is satisfied, then they posit that

$$\langle\theta, D_g\psi(0,\theta)\rangle$$

is a conserved quantity along the gradient flow. Thus, not all symmetries can induce conservation laws based on their framework, but only those that satisfy the additional condition 13. This is reflected in the additional assumption in Theorem 1 of Kunin et al. (2021).

Zhao et al. (2023) show that if the infinitesimal generator of a symmetry can be linearly represented by a matrix M, i.e.,

$$\psi(g,\theta) = e^{gM}\theta, \quad \text{and} \quad \chi(\theta) = D_g\psi(0,\theta) = M\theta,$$

then the derivative of the same quantity in Kunin et al. (2021) can be rewritten as

$$\frac{d}{dt}\langle\theta,\chi(\theta)\rangle = \frac{d}{dt}\langle\theta, M\theta\rangle = \langle\dot\theta, M\theta\rangle + \langle\theta, M\dot\theta\rangle = \underbrace{\langle\nabla_\theta L_{\mathcal{D}}(\theta), D_g\psi(0,\theta)\rangle}_{=0 \text{ by the infinitesimal invariance condition 4}} + \langle M^\top\theta, \dot\theta\rangle.$$

Thus, if the additional condition

$$\langle M^\top\theta, \dot\theta\rangle = 0$$

is satisfied, then $\langle\theta, M\theta\rangle$ is a conserved quantity along the gradient flow. In particular, if M is symmetric or skew-symmetric, then this condition is automatically satisfied as we can convert $M^\top$ to either M or $-M$, yielding the infinitesimal symmetry condition again. This is reflected in the additional assumption in Proposition 5.1 of Zhao et al. (2023).

To give a concrete example of a network with a symmetry that is associated to a conservation law through our framework, but that does not give rise to a conservation law under the frameworks of Kunin et al. (2021); Zhao et al. (2023), consider the following scalar network:

$$G(w_1, w_2) = w_1 e^{2w_2} x.$$

It has a symmetry given by the group $\mathbb{R}$ acting on the parameters as

$$g\cdot(w_1,w_2) \mapsto (w_1e^{-2g}, w_2+g).$$

Its infinitesimal generator is

$$\chi = D_g(0,(w_1,w_2)) = (-2w_1, 1),$$

which automatically falls outside the scope of Zhao et al. (2023) since the infinitesimal generator of this symmetry is not linear in the parameters. The additional condition (13) in the work of Kunin et al. (2021) is not satisfied, as we have

$$\langle (w_1, w_2), D_t D_g(0, (w_1, w_2))\, \nabla_{(w_1,w_2)} L_{\mathcal{D}}(w_1, w_2) \rangle = \langle (w_1, w_2), D_t(-2w_1, 1) \rangle = -2w_1 \dot{w}_1 \neq 0.$$

In contrast, from our framework, we observe that χ is closed, and thus it is locally the gradient of a conservation law h. This conservation law is easy to find by inspection, and is given by

$$h(w_1, w_2) = -w_1^2 + w_2$$

An alternative framework to connect symmetries and conservation laws by Tanaka & Kunin (2021) considers approximating gradient flow by other dynamics that can be induced from The Bregman Lagrangian, introduced by Wibisono et al. (2016). The authors found that not all symmetries of the loss function are also symmetries of the Bregman Lagrangian, a phenomenon they term as *symmetry breaking*. Nevertheless, by continuing with the derivations, they found that the usual Noether charge satisfies an evolution equation, which they term the *Noether's Learning Dynamics*. By taking the mass-less limit of this equation, they recover an equation equivalent to the infinitesimal symmetry condition 4. Their limiting identity recovers gradient-flow conservation properties, but does not provide a general criterion for a prescribed generator to admit a conserved potential.

Finally, Głuch & Urbanke (2021) study symmetries and conservation laws of dynamics of the form

$$\kappa_2(t)\ddot{\theta} + \kappa_1(t)\dot{\theta} + \nabla L(\theta) = 0,$$

induced by a Lagrangian similar to Tanaka & Kunin (2021). Although the authors also discuss an interpretation of gradient flow as a massless limit of this dynamic, their derivation does not require any construction of a Lagrangian. To derive conservation laws from symmetries, they again start from the infinitesimal symmetry condition 4. The main observation is that if the term in the infinitesimal symmetry condition $\langle \dot{\theta}, D_g(0, \theta) \rangle$ is exactly the time derivative of some potential function, then that function is a conserved quantity along the dynamic. This approach generalizes those presented by Kunin et al. (2021); Zhao et al. (2023), in the sense that it does not prescribe any particular form of the conserved quantity. Rather, in cases where they can find a potential function for the infinitesimal symmetry condition 4, they deduce that such function is a conserved quantity, while in cases where they cannot find such a potential function, they retain the identity as a constraint.

For the purpose of finding conservation laws of gradient flow, our framework refines the conservation law construction of Głuch & Urbanke (2021). In particular, if $h(\theta)$ is a conservation law then

$$\frac{d}{dt} h(\theta) = \langle \dot{\theta}, \nabla h(\theta) \rangle = 0.$$

Thus, if we define a partial symmetry ψ with infinitesimal generator $\chi = \nabla h$, then the infinitesimal symmetry condition 4 is satisfied, and h is a potential function for the infinitesimal symmetry condition. Through Proposition 8 and Proposition 9, our framework further makes explicit the integrability criterion underlying their construction. That is, the infinitesimal symmetry condition

$$\langle \dot{\theta}, D_g \psi(0, \theta) \rangle$$

is integrable locally if and only if $D_g \psi(0, \theta)$ is a closed vector field.

As such, our framework characterizes the fundamental relation underlying previous frameworks by Kunin et al. (2021); Głuch & Urbanke (2021); Zhao et al. (2023). Any sufficiently smooth conservation law of gradient flow derived by these frameworks could be recovered by ours as well.

B.3 Recent works connecting conservation laws and symmetries via vector fields

The observation that invariant transformations of the loss function and conservation laws of the gradient flow are related through their infinitesimal generators has been made in several recent works.

Marcotte et al. (2025) briefly discuss this in their Appendix J. They define a conservation law induced by a loss invariant transformation as a function h that has the same gradient as the infinitesimal generator of the transformation itself.

In the case where the symmetry is a linear group action of some Lie group $G \leq GL(n, \mathbb{R})$, Corollary 1 of Josz (2025) defines a candidate conservation law for gradient flow with non-smooth objective functions of the form

$$h(\theta) = \mathrm{proj}_{s(\mathfrak{g})}(\theta\theta^\top),$$

where $s(\mathfrak{g})$ is the symmetric part of the Lie algebra $\mathfrak{g} = T_{I_n}(G)$ of G. The main ingredient in the proof is the observation that such a candidate annihilates the gradient flow vector field.

Gegenfurtner et al. (2026) observe in their Appendix F that if the infitesimal generator of a particular family of symmetries is exact, then its potential function must be conserved. They also exhibit an infinitesimal generator of a symmetry that is not closed, and conclude that it does not give rise to a conservation law through this construction.

Finally, White (2025) states similar observations under a variational framework for gradient flow PDEs arising in physics.

In contrast to prior works, our framework systematically characterizes the precise relationship between local one-parameter loss symmetries and conservation laws through their infinitesimal generators. Here, closedness determine the existence of local conserved potentials, while each sufficiently smooth conservation law determines a unique partial symmetry generated by its gradient, as discussed in Section 3.

C Justification for 1D Symmetries

We motivates the study of 1 dimensional symmetries by showing how multidimensional symmetries can be recovered by understanding symmetries induced by the trajectories of the Lie algebra $\mathfrak{g}$.

Let Γ be a connected Lie group of dimension s and $\psi : \Gamma \times \mathbb{R}^D \to \mathbb{R}^D$ be a smooth symmetry of the loss function L. Let $\mathfrak{g} = \mathrm{span}(X_1, \ldots, X_s)$ be its Lie algebra. For each i, the exponential trajectories $\{\exp(tX_i) : t \in \mathbb{R}\}$ form a 1-parameter Lie subgroup. Thus, we can define s 1-dimensional symmetries of L by taking

$$\begin{aligned} \varphi_i : \mathbb{R} \times \mathbb{R}^D &\to \mathbb{R}^D \\ (t, \theta) &\mapsto \psi(\exp(tX_i), \theta) = \exp(tX_i) \cdot \theta. \end{aligned}$$

On the other hand, if we know all the 1-parameter Lie subgroups φ_i for $i = 1, \ldots, d$, then we can recover the multidimensional symmetry ψ if the Lie group Γ is connected, as can be seen from the following theorem.

Theorem 22. *If Γ is a connected Lie group and $\mathfrak{g} = \mathrm{span}(X_1, \ldots, X_s)$ then the product of exponentials of the form $\exp(t_k X_{i_k}) \ldots \exp(t_1 X_{i_1})$ generates Γ.*

For a standard reference, see Exercise 8.1 and Proposition 8.3 from Fulton & Harris (1991).

Thus, a corollary of these facts is that if the 1-parameter Lie subgroups $\varphi_i(t, \theta) = \exp(tX_i) \cdot \theta$ are known for all i, then we can reconstruct $\psi : \Gamma \times \mathbb{R}^D \to \mathbb{R}^D$. This justifies our focus on characterizing 1-dimensional symmetries for the time being.

D Proofs for Section 2

D.1 Proof for Proposition 1

Proposition 1. *Suppose $\Theta = \mathbb{R}^D$, the loss function ℓ is bounded from below, and for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell_{(x,y)}(\theta) \in C^1(\Theta, \mathbb{R})$ with locally lipschitz gradient $\nabla_\theta \ell_{(x,y)}(\theta)$. Then for every finite nonempty dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^*$ and every initialization $\theta_0 \in \Theta$, the gradient flow (1) admits a unique solution $\theta(t)$ defined for all $t \in [0, \infty)$.*

Proof. Let $\mathcal{D}$ be any finite nonempty dataset and θ_0 be any initialization. Since these are now fixed, we suppress the dataset $\mathcal{D}$ in the notation and write $L(\theta) := L_{\mathcal{D}}(\theta)$. Since summations of continuously differentiable functions are continuously differentiable, L is continuously differentiable.

Similarly, since summations of locally Lipschitz functions are locally Lipschitz, $\nabla_\theta L$ is locally Lipschitz. By the Picard–Lindelöf theorem, for every $\theta_0 \in \mathbb{R}^D$, there exists a unique local solution

$$\theta : [0, T) \longrightarrow \mathbb{R}^D$$

of

$$\dot{\theta}(t) = -\nabla_\theta L(\theta(t)), \qquad \theta(0) = \theta_0,$$

for some $T \in (0, \infty]$. By a standard continuation argument, there exists a maximal $T_{\max} \in (0, \infty]$ such that the solution can be extended to $[0, T_{\max})$. It remains to prove that $T_{\max} = \infty$.

Note that $T_{\max} < \infty$ if and only if there is a finite time blow-up of the solution, i.e.,

$$\lim_{t \to T_{\max}^-} \|\theta(t)\| = \infty.$$

However, along the solution, the chain rule gives the identity

$$\frac{\mathrm{d}}{\mathrm{d}t} L(\theta(t)) = \Big\langle \nabla_\theta L(\theta(t)), \dot{\theta}(t) \Big\rangle = - \|\nabla_\theta L(\theta(t))\|^2 = -\|\dot{\theta}(t)\|^2.$$

Integrating from 0 to $t < T_{\max}$, and noting that the loss function L is bounded from below since ℓ itself is bounded from below, we obtain

$$\int_0^t \|\dot{\theta}(s)\|^2 \, \mathrm{d}s = L(\theta_0) - L(\theta(t)) \leq L(\theta_0) - \inf_{\theta \in \mathbb{R}^D} L(\theta).$$

By the Cauchy–Schwarz inequality,

$$\|\theta(t) - \theta_0\| = \left\| \int_0^t \dot{\theta}(s) \, \mathrm{d}s \right\| \leq \int_0^t \|\dot{\theta}(s)\| \, \mathrm{d}s \leq \sqrt{t} \left(\int_0^t \|\dot{\theta}(s)\|^2 \, \mathrm{d}s \right)^{1/2} \leq \sqrt{t(L(\theta_0) - \inf_{\theta \in \mathbb{R}^D} L(\theta))}.$$

Thus

$$\|\theta(t)\| \leq \|\theta_0\| + \sqrt{t(L(\theta_0) - \inf_{\theta \in \mathbb{R}^D} L(\theta))},$$

and hence no finite time blow-up is possible.

We conclude that for every finite nonempty dataset and every $\theta_0 \in \mathbb{R}^D$, the gradient-flow equation admits a unique solution defined for all $t \in [0, \infty)$. □

D.2 Proof for Proposition 4

Proposition 4. *Let $\Theta \subset \mathbb{R}^D$ be open, and let $H = (h_1, \ldots, h_k) : \Theta \to \mathbb{R}^k$, $h : \Theta \to \mathbb{R}$ be C^r functions for $r \geq 1$. Assume $\nabla_\theta h_1, \ldots, \nabla_\theta h_k$ are linearly independent at every $\theta \in \Theta$. Then the following two conditions are equivalent:*

(i) For every $\theta \in \Theta$, $\nabla_\theta h(\theta) \in \operatorname{span}\{\nabla_\theta h_1(\theta), \ldots, \nabla_\theta h_k(\theta)\}$.

(ii) For every $\theta_0 \in \Theta$, there exist neighborhoods $V \ni \theta_0, W \ni H(\theta_0)$ and $f \in C^r(W)$ such that

$$h(\theta) = f(H(\theta)) = f(h_1(\theta), \ldots, h_k(\theta)), \quad \forall \theta \in V. \tag{3}$$

Proof. The backward direction is obtained by taking the derivative of Equation 3. Indeed, for any $\theta \in \Theta$, the associated f induces the required linear coefficients since

$$\nabla_\theta h(\theta) = \sum_{i=1}^k \frac{\partial f}{\partial x_i}(H(\theta)) \nabla_\theta h_i(\theta) \in \operatorname{span}\{\nabla_\theta h_1(\theta), \ldots, \nabla_\theta h_k(\theta)\}.$$

Now, assume (i) holds. Since the gradients of the h_i are linearly independent in Θ, H is a submersion from $\Theta \subset \mathbb{R}^D$ to $\mathbb{R}^k$. Therefore, by the submersion theorem (Tu, 2010, Theorem 11.5), there exists an open neighborhood $\theta_0 \in V \subset \Theta$ and diffeomorphisms $\phi : V \to \mathbb{R}^D, \psi : H(V) \to \mathbb{R}^k$ such that

$$\psi \circ H \circ \phi^{-1} : \mathbb{R}^D \to \mathbb{R}^k$$

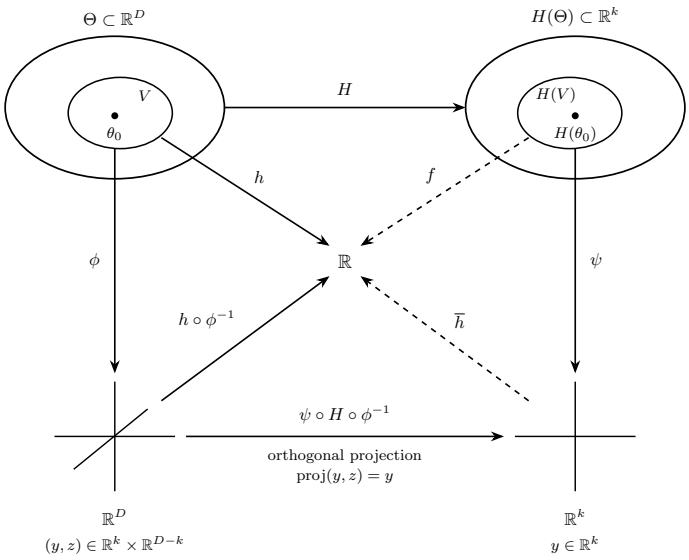


Figure 1: Commutative diagram showing the domains and ranges of different functions in the proof. The key step is to construct $\overline{h}$. The final construction is $h = f \circ H = \overline{h} \circ \psi \circ H$.

is the orthogonal projection $\mathrm{proj} : \mathbb{R}^D \to \mathbb{R}^k$ of the first k coordinates, i.e.,

$$\psi \circ H \circ \phi^{-1}(y, z) = \mathrm{proj}(y, z) = y, \quad \forall y \in \mathbb{R}^k, z \in \mathbb{R}^{D-k}.$$

Intuitively, ϕ and ψ flatten both the domain and the range of H. This makes explicit the idea that up to diffeomorphisms, H can only see variations in k directions. Next, by assumption (i), we will show that these directions entirely encompass the direction of variation of h up to diffeomorphism.

We claim that $h \circ \phi^{-1} : \mathbb{R}^D \to \mathbb{R}$ is independent of the last $D-k$ coordinates, and hence $\overline{h} : \mathbb{R}^k \to \mathbb{R}$ defined by

$$\overline{h}(y) = h \circ \phi^{-1}(y, z), \quad \forall z \in \mathbb{R}^{D-k}$$

is a well-defined function. If this is the case, then the projection map $\mathrm{proj} : \mathbb{R}^D \to \mathbb{R}^k$ to the first k coordinates satisfies

$$h \circ \phi^{-1}(y, z) = \overline{h}(y) = \overline{h}(\mathrm{proj}(y, z)),$$

and we would have for any $\theta \in V$

$$h(\theta) = h \circ \phi^{-1}(\phi(\theta)) = \overline{h}(\mathrm{proj}(\phi(\theta)) = \overline{h} \circ \psi \circ H \circ \phi^{-1} \circ \phi(\theta) = \overline{h} \circ \psi(H(\theta)) = \overline{h} \circ \psi(h_1(\theta), \ldots, h_k(\theta)).$$

Thus we can choose $f(x) := \overline{h} \circ \psi(x) = h \circ \phi^{-1}(\psi(x), 0)$. The fact that f is C^r is inherited from h. An illustrative commutative diagram is included in Figure 1 for ease of following the proof.

The final step is to prove the claim that $h \circ \phi^{-1} : \mathbb{R}^D \to \mathbb{R}$ is independent of the last $D - k$ coordinates. Let $y \in \mathbb{R}^k, z \in \mathbb{R}^{D-k}$. By chain rule and the fact that $\psi \circ H \circ \phi^{-1}$ is the orthogonal projection, observe that

$$0 = D_z \left(\psi \circ H \circ \phi^{-1}\right)\big|_{(y,z)} = D\psi|_{H \circ \phi^{-1}(y,z)} \; DH|_{\phi^{-1}(y,z)} \; D_z\phi^{-1}\big|_{(y,z)}.$$

Since ψ is a diffeomorphism, this implies $DH|_{\phi^{-1}(y,z)} \; D_z\phi^{-1}\big|_{(y,z)} = 0$, and thus

$$Dh_i|_{\phi^{-1}(y,z)} \; D_z\phi^{-1}\big|_{(y,z)} = 0, \quad \forall i \in \overline{1, \ldots, k}.$$

By the chain rule and assumption (i), we obtain

$$D_z(h \circ \phi^{-1})\big|_{(y,z)} = Dh|_{\phi^{-1}(y,z)} \; D_z\phi^{-1}\big|_{(y,z)} = 0.$$

Thus, $h \circ \phi^{-1}$ is independent of z, completing the proof. □

E Proof for Section 3

E.1 Proof for Proposition 5

Proposition 5. *Let Γ be a connected Lie group that acts on Θ via a smooth group action ψ. A necessary and sufficient condition for Γ to be a differentiable symmetry of $L_{\mathcal{D}}$ is*

$$D_g\psi(e,\theta)^\top \nabla_\theta L_{\mathcal{D}}(\theta) = 0, \quad \forall \theta \in \Theta. \tag{5}$$

Proof. Suppose $L_{\mathcal{D}}$ has the differentiable symmetry Γ, then

$$\mathcal{L}(g,\theta) := L_{\mathcal{D}} \circ \psi(g,\theta) = L_{\mathcal{D}}(\psi(g,\theta))$$

is a constant function with respect to g. Thus, for all $g \in \Gamma, \theta \in \Theta$, we have

$$0 = (D_g\mathcal{L})|_{(g,\theta)} = DL_{\mathcal{D}}|_{\psi(g,\theta)}\, D_g\psi|_{(g,\theta)}\,.$$

In particular, for $g = e$, we have

$$(D_g\mathcal{L})|_{(e,\theta)} = DL_{\mathcal{D}}|_{\psi(e,\theta)}\, D_g\psi|_{(e,\theta)} = DL_{\mathcal{D}}|_{\theta}\, D_g\psi|_{(e,\theta)} = DL_{\mathcal{D}}(\theta)\, D_g\psi(e,\theta) = 0, \quad \forall \theta \in \Theta.$$

Taking the transpose yields equation 5.

Conversely, suppose equation 5 holds for all $\theta \in \Theta$. For any $g \in \Gamma$ and any $v \in T_g\Gamma$, there exists a smooth path $\gamma : [0,1] \to \Gamma$ such that $\gamma(0) = g, \dot\gamma(0) = v$. We can reparametrize

$$\gamma(t) = (\gamma(t)g^{-1})g =: \epsilon(t)g, \quad \epsilon(0) = e.$$

Then

$$\frac{d}{dt}\mathcal{L}(\gamma(t),\theta) = \frac{d}{dt}L_{\mathcal{D}}(\psi(\epsilon(t)g,\theta)) = \frac{d}{dt}L_{\mathcal{D}}(\psi(\epsilon(t),\psi(g,\theta))) = DL_{\mathcal{D}}|_{\psi(\epsilon(t),\psi(g,\theta))}\, D_g\psi|_{(\epsilon(t),\psi(g,\theta))}\,\dot\epsilon(t).$$

In particular, for $t = 0$, we have

$$\frac{d}{dt}\mathcal{L}(\gamma(t),\theta)\bigg|_{t=0} = DL_{\mathcal{D}}|_{\psi(e,\psi(g,\theta))}\, D_g\psi|_{(e,\psi(g,\theta))}\,\dot\epsilon(0) = DL_{\mathcal{D}}|_{\psi(g,\theta)}\, D_g\psi|_{(e,\psi(g,\theta))}\,\dot\epsilon(0) = 0,$$

where the last equality follows from the assumption equation 5. This proves that

$$D_g\mathcal{L}(g,\theta)[v] = 0, \forall v \in T_g\Gamma.$$

Since $v \in T_g\Gamma$ was arbitrary, we have shown that $D_g\mathcal{L}(g,\theta) = 0$ for all $g \in \Gamma$.

Finally, by the connectedness of Γ, there exists a smooth path connecting any $g \in \Gamma$ to the identity e, and since the derivative along any such path is zero, we conclude that

$$\mathcal{L}(g,\theta) = \mathcal{L}(e,\theta) = L_{\mathcal{D}}(\psi(e,\theta)) = L_{\mathcal{D}}(\theta)$$

for all $g \in \Gamma$. □

E.2 Proof for Corollary 6

Corollary 6. *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a smooth group action $\psi(t,\theta)$. Let $\mathcal{D}$ be a dataset in $(\mathcal{X} \times \mathcal{Y})^*$. A necessary and sufficient condition for Γ to be a differentiable symmetry of $L_{\mathcal{D}}$ is*

$$\langle D_t\psi(0,\theta), \nabla_\theta L_{\mathcal{D}}(\theta)\rangle = 0, \quad \forall \theta \in \Theta. \tag{6}$$

Proof. Replace Γ with the additive group $\mathbb{R}$ and the identity element e with 0 in Proposition 5. Since $D_t\psi(0,\theta)$ is a vector field, we can rewrite the equation 6 using the inner product notation as desired. □

E.3 Proof for Corollary 7

Corollary 7. *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a smooth group action $\psi(t, \theta)$. A necessary and sufficient condition for Γ to be a differentiable symmetry of L is*

$$D_t\psi(0, \theta) \perp W_\theta, \quad \forall \theta \in \Theta. \tag{7}$$

Proof. The forward implication follows directly from Corollary 6 by noting that W_θ is the span of the gradient flow directions for all datasets of size 1.

For the reverse implication, for any dataset $\mathcal{D}$, its gradient flow vector field

$$-\nabla_\theta L_\mathcal{D}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell_{(x^{(i)}, y^{(i)})}(\theta)$$

belongs to W_θ. Therefore, if equation 7 holds then

$$\langle D_t\psi(0, \theta), -\nabla_\theta L_\mathcal{D}(\theta)\rangle = 0, \quad \forall \theta \in \Theta.$$

Corollary 6 then implies that ψ is a symmetry of $L_\mathcal{D}$. Since the dataset $\mathcal{D}$ was arbitrary, this completes the proof. □

E.4 Proof for Proposition 8

Proposition 8. *If h is a conservation law, then $\chi := \nabla h$ is a closed vector field in $W^\perp$. On the other hand, if χ is a closed vector field in $W^\perp$, then locally around every $\theta \in \Theta$, there exists a unique scalar potential function h such that $\chi = \nabla h$, up to an additive constant. Furthermore, if Θ is a star-shaped domain, then the scalar potential function h exists globally on Θ.*

Proof. The first part of the proposition follows directly from the definition of a conservation law and the fact that the gradient of any scalar function is a closed vector field. The second part follows from the Poincaré lemma (Lee, 2003, Theorem 11.50). The final part is the Poincaré lemma itself (Lee, 2003, Theorem 11.49). □

E.5 Proof for Proposition 9

Proposition 9. *If $(\mathbb{R}, \psi)$ is a symmetry of the loss function L, then $\chi := D_t\psi \mid_{t=0}$ is a smooth vector field in $W^\perp$. On the other hand, if $\chi \in W^\perp$ is a smooth vector field, then there exists a unique maximal local flow $\psi : \mathcal{O} \to \Theta$, where $\mathcal{O} \subset \mathbb{R} \times \Theta$ is the maximal open set containing $\{0\} \times \Theta$, for which the solution of*

$$\frac{d}{dt}\psi(t, \theta) = \chi(\psi(t, \theta)), \quad \psi(0, \theta) = \theta$$

is well-defined. This unique maximal local flow satisfies the following four equations for all $\theta \in \Theta$ and t, s whenever the terms are well-defined:

$$\psi(0, \theta) = \theta, \qquad \psi(t + s, \theta) = \psi(t, \psi(s, \theta)), \qquad \textit{(group action axioms)}$$
$$D_t\psi(0, \theta) = \chi(\theta), \qquad \textit{(infinitesimal generator)}$$
$$L(\psi(t, \theta)) = L(\theta). \qquad \textit{(loss invariance condition)}$$

Furthermore, if χ is complete, then ψ is a symmetry of the loss function L.

Proof. The first part of the proposition follows directly from Corollary 7.

The second part of the proposition follows from the standard theory of ordinary differential equations. Since χ is smooth, the existence and uniqueness of the maximal local flow ψ is guaranteed by the Picard-Lindelöff theorem. The first three properties of the maximal local flow listed in the proposition are standard results for the flow of a smooth vector field. The last property, $L(\psi(t, \theta)) = L(\theta)$, follows from the fact that the flow direction $\chi \in W^\perp$ is always perpendicular to the gradient flow distribution W.

Moreover, if the maximal local flow exists for all time, then by Corollary 7, ψ is a symmetry of the loss function L. □

E.6 Extension for Corollary 6, Corollary 7 and Propostion 9 to partial symmetries

It would have been more convenient to state all results in the context of partial symmetries, but we chose to first present the results in the context of global symmetries for clarity. Every global symmetry is of course also a partial symmetry. The nuance is just that you might not always be able to integrate a vector field to a global symmetry, but you can always integrate it to a local symmetry.

In a small neighborhood of a point, the local symmetry is indistinguishable from a global symmetry. Alternatively, if we assume that all vector fields that concern us are complete, then we can temporarily forget about the distinction between partial and global symmetries.

Proposition 23 (Extension of Corollary 6)**.** *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a partial group action $\psi(t, \theta)$. Let $\mathcal{D}$ be a dataset in $(\mathcal{X} \times \mathcal{Y})^*$. A necessary and sufficient condition for Γ to be a differentiable symmetry of $L_\mathcal{D}$ is*

$$\langle D_t\psi(0, \theta), \nabla_\theta L_\mathcal{D}(\theta)\rangle = 0, \quad \forall\theta \in \Theta.$$

Proof. (Γ, ψ) is a partial symmetry of $L_\mathcal{D}$ if and only if $L_\mathcal{D}(\psi(t, \theta))$ is constant with respect to t whenever defined, if and only if $\frac{d}{dt}L_\mathcal{D}(\psi(t, \theta)) = 0$ whenever defined. By the chain rule, this is equivalent to

$$\langle D_t\psi(t, \theta), \nabla_\theta L_\mathcal{D}(\psi(t, \theta))\rangle = 0$$

whenever defined. Setting $t = 0$ yields the forward implication.

On the other hand, suppose 6 holds for all $\theta \in \Theta$. Then it must hold that

$$\langle D_t\psi(0, \psi(s, \theta)), \nabla_\theta L_D(\psi(s, \theta))\rangle = 0$$

for all s whenever defined. But since ψ satisfies the group action property whenever defined, we get

$$D_t\psi(0, \psi(s, \theta)) = D_t\psi(s, \theta).$$

Thus, $\langle D_t\psi(s, \theta), \nabla_\theta L_\mathcal{D}(\psi(s, \theta))\rangle = 0$ for all s whenever defined, which is the equivalent condition above. □

Proposition 24 (Extension of Corollary 7)**.** *Let $\Gamma = \mathbb{R}$ be the additive group of real numbers that acts on Θ via a partial group action $\psi(t, \theta)$. A necessary and sufficient condition for Γ to be a partial symmetry of L is*

$$D_t\psi(0, \theta) \perp W_\theta, \quad \forall\theta \in \Theta.$$

Proof. The proof is identical to that of Corollary 7, except that we rely on Proposition 23, i.e., the extension of Corollary 6 to partial symmetries, instead of Corollary 6 itself. □

Proposition 25 (Extension of Proposition 9)**.** *If $(\mathbb{R}, \psi)$ is a partial symmetry of the loss function L, then $\chi := D_t\psi \mid_{t=0}$ is a smooth vector field in $W^\perp$. On the other hand, if $\chi \in W^\perp$ is a smooth vector field, then there exists a unique maximal local flow $\psi : \mathcal{O} \to \Theta$, where $\mathcal{O} \subset \mathbb{R} \times \Theta$ is the maximal open set containing $\{0\} \times \Theta$, for which the solution of*

$$\frac{d}{dt}\psi(t, \theta) = \chi(\psi(t, \theta)), \quad \psi(0, \theta) = \theta$$

is well-defined. This unique maximal local flow satisfies the following four equations for all $\theta \in \Theta$ and t, s whenever the terms are well-defined:

$$\begin{aligned}
&\psi(0, \theta) = \theta, \qquad \psi(t + s, \theta) = \psi(t, \psi(s, \theta)), && \textit{(group action axioms)}\\
&D_t\psi(0, \theta) = \chi(\theta), && \textit{(infinitesimal generator)}\\
&L(\psi(t, \theta)) = L(\theta). && \textit{(loss invariance condition)}
\end{aligned}$$

Furthermore, if χ is complete, then ψ is a symmetry of the loss function L.

Proof. The first part of the proposition follows directly from Proposition 24. The second part contains no difference compared to the corresponding part in Proposition 9, and hence does not require a separate proof. □

E.7 Proof for Corollary 10

Corollary 10. *Let $h : \Theta \to \mathbb{R}$ be a conservation law of the gradient flow dynamics. There exists a unique partial symmetry $(\mathbb{R}, \psi)$ of the loss function L such that*

$$D_t\psi(0,\theta) = \nabla h(\theta), \quad \forall\theta.$$

Conversely, let $(\mathbb{R}, \psi)$ be a partial symmetry of the loss function L. If $D_t\psi(0,\theta)$ is closed, then locally, there exists a unique conservation law h, up to an additive constant, such that

$$\nabla h(\theta) = D_t\psi(0,\theta), \quad \forall\theta.$$

Proof. Let h be a conservation law, then Proposition 8 implies that $\chi := \nabla h$ is a vector field in $W^\perp$. Applying Proposition 9 to the vector field χ completes the proof.

On the other hand, let ψ be a partial symmetry of L, then by Proposition 9, its infinitesimal generator $\chi := D_t\psi(0,\theta)$ is a vector field in $W^\perp$. If χ is also closed, then by Proposition 8, there exists a local conservation law h such that $\nabla h = \chi$, up to an additive constant. □

E.8 Proof for Theorem 12

Theorem 12. *Assume that the gradient flow distribution W is a smooth distribution of constant rank r, and the Lie completion $\mathrm{Lie}(W)$ is a smooth distribution of constant rank $\overline{r}$ in Θ. Then, for every $\theta \in \Theta$, there exists a neighborhood $U \ni \theta$ such that the following statements hold:*

- *(i) There exists $s := D - r$ partial symmetries of the loss function L that are independent in U, and they are complete.*
- *(ii) There exist $c := D - \overline{r}$ independent conservation laws of the gradient flow dynamics that are independent in U, and they are complete.*
- *(iii) The symmetry-conservation law gap is $s - c := \overline{r} - r \geq 0$. Equality holds if and only if W is involutive, i.e., $\mathrm{Lie}(W) = W$.*

Proof. Since W is a smooth distribution of constant rank r, its orthogonal complement $W^\perp$ is a smooth distribution of constant rank $D - r$. Thus, there exists a local smooth frame $\chi_1, \ldots, \chi_{D-r}$ that completely spans $W^\perp$ in a neighborhood $U \ni \theta$. By Proposition 9, each χ_i generates a unique partial symmetry $(\mathbb{R}, \psi_i)$ of the loss function L in U. Any other partial symmetry $(\mathbb{R}, \psi)$ of L must have its infinitesimal generator $D_t\psi(0,\theta)$ in $W^\perp$, and hence it is infinitesimally spanned by $\chi_1, \ldots, \chi_{D-r}$ in U. This proves (a).

Since $\mathrm{Lie}(W)$ is a smooth, involutive distribution of constant rank $\overline{r}$, its orthogonal complement $\mathrm{Lie}(W)^\perp$ is a smooth distribution of constant rank $D - \overline{r}$. By the Frobenius theorem, there exists a local coordinate system

$$u_1, \ldots, u_{\overline{r}}, v_1, \ldots, v_{D-\overline{r}}$$

around any $\theta \in \Theta$ such that the vector fields $\nabla v_1, \ldots, \nabla v_{D-\overline{r}}$ completely span $\mathrm{Lie}(W)^\perp$ in a neighborhood $U \ni \theta$. Note that ∇v_i is an exact vector field in $W^\perp$. By Proposition 8, the functions $h_i := v_i$ for $i = 1, \ldots, D - \overline{r}$ are conservation laws of the loss function L in U. Any other conservation law h must have its gradient ∇h lie in $\mathrm{Lie}(W)^\perp$, and hence it is infinitesimally spanned by $\nabla h_1, \ldots, \nabla h_{D-\overline{r}}$ in U. This proves (b).

Finally, since $W \subseteq \mathrm{Lie}(W)$, we have $r \leq \overline{r}$, and hence $s - c = \overline{r} - r \geq 0$. Equality holds if and only if $r = \overline{r}$, which is equivalent to W being involutive, i.e., $\mathrm{Lie}(W) = W$. □

F Proof for Section 4

F.1 Loss functions that satisfy Assumption 13

Assumption 13 (Separability of predictions)**.** *The loss function $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ satisfies*

$$\ell(z,y) = \ell(z',y), \forall y \in \mathcal{Y} \implies z = z'.$$

Here, we review some common loss functions that satisfy Assumption 13. This is not an exhaustive list since the assumption is very general. The idea is that the loss function should be able to distinguish between different outputs, which is a very mild requirement for most loss functions used in practice.

Lemma 26. *The following loss functions satisfy Assumption 13:*

1. *Assuming $\mathcal{Z} = \mathcal{Y}$, any loss function ℓ that satisfies $\ell(z, z) = 0$ and $\ell(z, y) > 0$ for all $z \neq y$.*
2. *Assuming $\mathcal{Z} = \mathcal{Y}$, any loss function ℓ that is also a metric on $\mathcal{Z}$, such as the Euclidean distance, or any L^p distance.*
3. *The quadratic loss*
$$\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad \ell(z, y) = \|z - y\|_2^2.$$
4. *The absolute loss*
$$\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad \ell(z, y) = \|z - y\|_1.$$
5. *Any L^p loss for $p \in (0, \infty)$*
$$\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad \ell(z, y) = \|z - y\|_p^p.$$
6. *The multiclass cross-entropy loss on the probability simplices*
$$\ell : \Delta_{++}^{V-1} \times \Delta_{+}^{V-1} \to \mathbb{R}, \quad \ell(z, y) = -\sum_{i=1}^{V} y_i \log z_i.$$
7. *The KL divergence loss on the probability simplices*
$$\ell : \Delta_{++}^{V-1} \times \Delta_{+}^{V-1} \to \mathbb{R}, \quad \ell(z, y) = \sum_{i=1}^{V} y_i \log \frac{y_i}{z_i}.$$
8. *The cosine loss on unit vectors*
$$\ell : S^{n-1} \times S^{n-1} \to \mathbb{R}, \quad \ell(z, y) = 1 - z^\top y.$$

Proof.

1. Choose $y = z$ to obtain $\ell(z, z) = 0$, then $\ell(z', z) = l(z, z) = 0$, implying $z' = z$ by the assumption that $\ell(z, y) > 0$ for all $z \neq y$.
2. Since ℓ is a metric, it satisfies the condition 1 above.
3. The quadratic loss is a metric.
4. The absolute loss is a metric.
5. The L^p loss is a metric for $p \geq 1$. For $p \in (0, 1)$, the L^p loss is not a metric, but it still satisfies the condition 1 above.
6. Multiclass cross-entropy loss: assume that
$$-\sum_{i=1}^{V} y_i \log z_i = -\sum_{i=1}^{V} y_i \log z_i' \qquad \text{for all } y \in \Delta_{+}^{V-1}.$$
For each $k \in \{1, \ldots, V\}$, choose $y = e_k$, where e_k is the k-th standard basis vector. Since $e_k \in \Delta_{+}^{V-1}$, we obtain
$$-\log z_k = -\log z_k'.$$
Because the logarithm is injective, we must have $z_k = z_k'$ for every index k. Since this holds for every k, we get $z = z'$.

7. The KL divergence loss satisfies the condition 1 above. Indeed, applying the inequality $\log t \leq t - 1, \forall t > 0$, we get

$$D_{\mathrm{KL}}(y\|z) = \sum_i y_i \log \frac{y_i}{z_i} = -\sum_i y_i \log \frac{z_i}{y_i} \geq -\sum_i y_i \left(\frac{z_i}{y_i} - 1\right) = -\sum_i (z_i - y_i) = 0,$$

with equality if and only if $z_i = y_i$ for all i.

8. The cosine loss satisfies the condition 1 above. Indeed, since z and y are unit vectors, the only way to have $\ell(z, y) = 0$ is if $z = y$.

□

F.2 Proof for Proposition 14

Proposition 14. *If the loss function ℓ satisfies Assumption 13, then any partial symmetry of the loss function L is also a partial functional symmetry of the model G.*

Proof. Let ψ be a partial symmetry of the loss function L. This means that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\ell(G(\psi(t, \theta), x), y) = \ell(G(\theta, x), y).$$

By Assumption 13, this implies that for all $x \in \mathcal{X}$,

$$G(\psi(t, \theta), x) = G(\theta, x).$$

This shows that ψ is a functional symmetry of the model G. □

F.3 Proof for Proposition 15

Before proceeding to the proof, we first provide Table 1 of examples of function combinations that satisfy the identifiability condition (9) in Proposition 15.

Table 1: Examples of function combinations compatible with Proposition 15. The listed operations occur within the indicated architectures. The categories are not mutually exclusive. The list is not exhaustive. We write $g_\theta(u) = g(\theta, u)$, and $\mathcal{N}(i)$ is the neighborhood of node i.

Combination Type	**Operation**	**Typical Examples**
Additive	$z = g_\theta(x) + f_\omega(x)$	ResNet (He et al., 2015), LoRA (Hu et al., 2022)
Concatenation	$z = [g_\theta(x);\ f_\omega(x, g_\theta(x))]$	Inception (Szegedy et al., 2015), U-Net (Ronneberger et al., 2015)
Composition	$z = f_\omega(g_\theta(x)),$ or $z = g_\theta(f_\omega(x))$	MLPs (Rumelhart et al., 1986), CNNs (Lecun et al., 1998)
Hadamard Product / Gating	$z = g_\theta(x) \odot f_\omega(x)$	LSTM (Hochreiter & Schmidhuber, 1997) and GRU gates (Cho et al., 2014)
Input-dependent Weighted Sum	$z = a_\omega(x)\, g_\theta(x) + f_\omega(x)$	Mixture-of-Experts (Shazeer et al., 2017)
Outer Product	$Z = g_\theta(x)\, f_\omega(x)^\top$	Bilinear CNNs (Lin et al., 2015)
Graph Aggregation	$z_i = \mathrm{AGG}_\omega\big(x_i, \{g_\theta(x_j, x_i, e_{ij}) : j \in \mathcal{N}(i)\}\big)$	MPNNs (Gilmer et al., 2017)
Shared Weights Across Branches	$z_i = f_{i,\omega}(x, g_\theta(x)), \quad i = 1, \ldots, m$	Grouped-query attention (Ainslie et al., 2023)
Repeated Composition	$z_t = g_\theta(z_{t-1}, x_t),$	RNNs (Cho et al., 2014)

Proposition 15 (Inheritance from subnetworks I). *Let $g : \Theta \times \mathcal{X}^g \to \mathcal{Z}^g$ and $G : (\Theta \times \Omega) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$g(\theta, \cdot) = g(\theta', \cdot) \quad \Longrightarrow \quad G((\theta, \omega), \cdot) = G((\theta', \omega), \cdot) \qquad \textit{for every } \omega \in \Omega. \tag{9}$$

Every partial symmetry $(\mathbb{R}, \psi)$ or conservation law h of g for a loss satisfying Assumption 13 extends to a partial symmetry $(\mathbb{R}, \tilde{\psi})$ or conservation law $\tilde{h}$ of G for any loss function via

$$\tilde{\psi}\,(t, (\theta, \omega)) = (\psi(t, \theta), \omega), \qquad \textit{or} \qquad \tilde{h}(\theta, \omega) = h(\theta). \tag{10}$$

Proof.

- Let $(\mathbb{R}, \psi)$ be a partial loss symmetry of the subnetwork g with respect to a loss satisfying Assumption 13. By Proposition 14, $(\mathbb{R}, \psi)$ is also a partial functional symmetry of g, i.e.,

  $$g(\psi(t, \theta), x) = g(\theta, x), \quad \forall \theta \in \Theta, x \in \mathcal{X}^g, t \text{ whenever well-defined.}$$

  Condition 9 implies that for all $(\theta, \omega) \in \Theta \times \Omega$, $x \in \mathcal{X}$, and t whenever well-defined,

  $$G(\tilde{\psi}(t, (\theta, \omega)), x) = G((\psi(t, \theta), \omega), x) = G((\theta, \omega), x).$$

  Thus, $(\mathbb{R}, \tilde{\psi})$ is a partial functional symmetry of the larger network G, and hence also a partial loss symmetry of G with respect to any loss function.

- Let h be a conservation law of the subnetwork g with respect to a loss satisfying Assumption 13. By Corollary 10, there exists a unique partial loss symmetry $(\mathbb{R}, \psi)$ of g associated with the underlying vector field $\chi := \nabla h$. The previous argument shows that the extension $(\mathbb{R}, \tilde{\psi})$ is a partial loss symmetry of the larger network G for any loss function.

  Importantly, the underlying vector field of $(\mathbb{R}, \tilde{\psi})$ is exactly the extension of χ to the larger parameter space $\Theta \times \Omega$, i.e.,

  $$\tilde{\chi} := D_t \tilde{\psi}(0, (\theta, \omega)) = (\chi, 0).$$

  By Proposition 9, the underlying vector field $\tilde{\chi}$ of $(\mathbb{R}, \tilde{\psi})$ is in the annihilator of the gradient flow distribution of the larger network G, i.e., $\tilde{\chi} \in W^{\perp}$ for any loss function. Note that this vector field is exactly the gradient of the extended function $\tilde{h}$, i.e., $\tilde{\chi} = \nabla \tilde{h}$. Since $\tilde{\chi} \in W^{\perp}$ and $\tilde{h}$ is a scalar potential function of $\tilde{\chi}$, Proposition 8 implies that $\tilde{h}$ is a conservation law of the larger network G with respect to any loss function.

□

F.4 Proof for Proposition 16

Proposition 16 (Inheritance from subnetworks II). *Let $g^j : \Theta^j \times \mathcal{X}^{g^j} \to \mathcal{Z}^{g^j}, j = 1, 2$ and $G : (\Theta^1 \times \Theta^2) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$[G((\theta_1, \theta_2), \cdot) = G((\theta_1', \theta_2'), \cdot)] \implies [g^j(\theta_j, \cdot) = g^j(\theta_j', \cdot), \forall j = 1, 2]. \tag{11}$$

For $j = 1, 2$, let $(\mathbb{R}, \psi_i^j)_{i=1}^{s_j}$ be a complete set of partial functional symmetries of g^j, and let $h_1^j, \ldots, h_{c_j}^j$ be a complete set of conservation laws of g^j w.r.t. a loss satisfying Assumption 13.

Every partial symmetry $(\mathbb{R}, \psi)$ or conservation law h of G with respect to a loss satisfying Assumption 13 is infinitesimally generated by the extensions

$$\tilde{\psi}_i^1(t, (\theta_1, \theta_2)) = (\psi_i^1(t, \theta_1), \theta_2), \quad \tilde{\psi}_i^2(t, (\theta_1, \theta_2)) = (\theta_1, \psi_i^2(t, \theta_2)), \quad \textit{or} \quad \tilde{h}_i^j(\theta_1, \theta_2) = h_i^j(\theta_j),$$

respectively. The extensions themselves need not be partial symmetries nor conservation laws of G.

Proof.

- Let $(\mathbb{R}, \psi)$ be a partial loss symmetry of the larger network G with respect to a loss satisfying Assumption 13, where for any t and $(\theta_1, \theta_2) \in \Theta^1 \times \Theta^2$, we have

$$\psi(t, (\theta_1, \theta_2)) = \begin{pmatrix} \psi^1(t, (\theta_1, \theta_2)) \\ \psi^2(t, (\theta_1, \theta_2)) \end{pmatrix} \in \Theta^1 \times \Theta^2.$$

  By Proposition 14, $(\mathbb{R}, \psi)$ is also a partial functional symmetry of G, i.e.,

$$G(\psi(t, (\theta_1, \theta_2)), \cdot) = G((\theta_1, \theta_2), \cdot), \quad \forall (\theta_1, \theta_2) \in \Theta^1 \times \Theta^2, t \text{ whenever well-defined.}$$

  Condition 11 now implies that for all $j = 1, 2$, we have

$$g^j(\psi^j(t, (\theta_1, \theta_2)), \cdot) = g^j(\theta_j, \cdot), \quad \forall (\theta_1, \theta_2) \in \Theta^1 \times \Theta^2, t \text{ whenever well-defined.}$$

  Observe that the right hand side of the previous equation does not depend on t, and hence so does the left hand side. By taking its derivative with respect to t at $t = 0$, we obtain

$$D_{\theta_j} g^j(\psi^j(0, (\theta_1, \theta_2)), \cdot)\, D_t \psi^j(0, (\theta_1, \theta_2)) = 0, \quad \forall (\theta_1, \theta_2) \in \Theta^1 \times \Theta^2.$$

  By definition, the parameter term inside g^j is $\psi^j(0, (\theta_1, \theta_2)) = \theta_j$, and so we have

$$D_{\theta_j} g^j(\theta_j, \cdot)\, D_t \psi^j(0, (\theta_1, \theta_2)) = 0, \quad \forall (\theta_1, \theta_2) \in \Theta^1 \times \Theta^2. \tag{14}$$

  For any fixed θ_2, equation 14 shows that the vector field $D_t \psi^1(0, (\theta_1, \theta_2))$ is an infinitesimal functional symmetry direction of g^1 at any θ_1. Thus, it is infinitesimally generated by the complete set of partial symmetries $(\mathbb{R}, \psi_i^1)_{i=1}^{s_1}$ of g^1, i.e.,

$$D_t \psi^1(0, (\theta_1, \theta_2)) \in \operatorname{span}\{D_t \psi_i^1(0, \theta_1), i = 1, \ldots, s_1\}.$$

  Note that

$$D_t \tilde{\psi}_i^1(0, (\theta_1, \theta_2)) = \begin{pmatrix} D_t \psi_i^1(0, \theta_1) \\ 0 \end{pmatrix}, \forall i = 1, \ldots, s_1$$

  and so

$$\begin{aligned} \begin{pmatrix} D_t \psi^1(0, (\theta_1, \theta_2)) \\ 0 \end{pmatrix} &\in \operatorname{span}\left\{ \begin{pmatrix} D_t \psi_i^1(0, \theta_1) \\ 0 \end{pmatrix}, i = 1, \ldots, s_1 \right\} \\ &= \operatorname{span}\{D_t \tilde{\psi}_i^1(0, (\theta_1, \theta_2)), i = 1, \ldots, s_1\}. \end{aligned}$$

  Similarly, for any fixed θ_1, the same argument shows that

$$\begin{pmatrix} 0 \\ D_t \psi^2(0, (\theta_1, \theta_2)) \end{pmatrix} \in \operatorname{span}\{D_t \tilde{\psi}_i^2(0, (\theta_1, \theta_2)), i = 1, \ldots, s_2\}.$$

  Combining the two results shows

$$D_t \psi(0, (\theta_1, \theta_2)) = \begin{pmatrix} D_t \psi^1(0, (\theta_1, \theta_2)) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ D_t \psi^2(0, (\theta_1, \theta_2)) \end{pmatrix}$$

  is in the span of $D_t \tilde{\psi}_i^j(0, (\theta_1, \theta_2)), j = 1, 2, i = 1, \ldots, s_j$. Thus, any partial loss symmetry of the larger network G is infinitesimally generated by the extensions of the complete set of partial functional symmetries of the subnetworks g^1 and g^2.

- Let h be a conservation law of the larger network G with respect to a loss satisfying Assumption 13. By Corollary 10, there exists a unique partial loss symmetry $(\mathbb{R}, \psi)$ of G given by

$$\psi(t, (\theta_1, \theta_2)) = \begin{pmatrix} \psi^1(t, (\theta_1, \theta_2)) \\ \psi^2(t, (\theta_1, \theta_2)) \end{pmatrix}$$

  associated with the underlying vector field $\chi := \nabla h = (\nabla_{\theta_1} h, \nabla_{\theta_2} h)$.

  The previous argument shows that for any fixed θ_2,

$$\nabla_{\theta_1} h(\theta_1, \theta_2) = D_t \psi^1(0, (\theta_1, \theta_2))$$

is an infinitesimal functional symmetry direction of g^1 at θ_1. Thus, it is also an infinitesimal loss symmetry direction of g^1, i.e., $\nabla_{\theta_1} h \in (W^1)^\perp$. Note that $\nabla_{\theta_1} h$ is exactly the gradient of the function $h(\theta_1, \theta_2)$ with fixed θ_2. Proposition 8 implies that $h(\cdot, \theta_2)$ is a conservation law of the subnetwork g^1. Since $h_1^1, \ldots, h_{c_1}^1$ is a complete set of conservation laws of g^1, we have

$$\nabla_{\theta_1} h(\theta_1, \theta_2) \in \operatorname{span}\{\nabla_{\theta_1} h_i^1(\theta_1) : i = 1, \ldots, c_1\}, \quad \forall \theta_1 \in \Theta^1.$$

Similarly, for any fixed θ_1, the same argument shows that

$$\nabla_{\theta_2} h(\theta_1, \theta_2) \in \operatorname{span}\{\nabla_{\theta_2} h_i^2(\theta_2) : i = 1, \ldots, c_2\}, \quad \forall \theta_2 \in \Theta^2.$$

Combining the two results shows that

$$\nabla h = (\nabla_{\theta_1} h, \nabla_{\theta_2} h) = \begin{pmatrix} \nabla_{\theta_1} h \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \nabla_{\theta_2} h \end{pmatrix}$$

is in the span of $\nabla \tilde{h}_i^j, j = 1, 2, i = 1, \ldots, c_j$.

We conclude that any conservation law of the larger network G is infinitesimally generated by the extensions of the complete set of conservation laws of the subnetworks g^1 and g^2.

□

F.5 Proof for Theorem 17

Theorem 17 (Inheritance from subnetworks III)**.** *Let $g^j : \Theta^j \times \mathcal{X}^{g^j} \to \mathcal{Z}^{g^j}, j = 1, \ldots, k$ and $G : (\Theta^1 \times \ldots \times \Theta^k) \times \mathcal{X} \to \mathcal{Z}$ satisfy*

$$[G((\theta_1, \ldots, \theta_k), \cdot) = G((\theta_1', \ldots, \theta_k'), \cdot)] \iff [g^j(\theta_j, \cdot) = g^j(\theta_j', \cdot), \forall j = 1, \ldots, k]. \tag{12}$$

For $j = \overline{1, k}$, let $(\mathbb{R}, \psi_i^j)_{i=1}^{s_j}$ be a complete set of partial functional symmetries of g^j, and let $h_1^j, \ldots, h_{c_j}^j$ be a complete set of conservation laws of g^j w.r.t. a loss satisfying Assumption 13.

Assume that the loss used for the larger network G also satisfies Assumption 13. A complete set of partial functional symmetries of G is given by the extensions

$$\tilde{\psi}_i^j(t, (\theta_1, \ldots, \theta_k)) = (\theta_1, \ldots, \psi_i^j(t, \theta_j), \ldots, \theta_k), \qquad j = \overline{1, k},\ i = \overline{1, s_j},$$

and a complete set of conservation laws of G is given by

$$\tilde{h}_i^j(\theta_1, \ldots, \theta_k) = h_i^j(\theta_j), \qquad j = \overline{1, k},\ i = \overline{1, c_j}.$$

Proof. Applying Proposition 15 for each subnetwork g^j shows that the extensions of the partial functional symmetries and conservation laws of g^j are also partial functional symmetries and conservation laws of G. Applying Proposition 16 recursively shows that any partial functional symmetry or conservation law of G is infinitesimally generated by the extensions of the partial functional symmetries and conservation laws of the subnetworks g^j.

The final step is to show that the extensions of the partial functional symmetries and conservation laws of the subnetworks g^j are independent. But this follows from the fact that the extensions of the partial functional symmetries and conservation laws of each subnetwork g^j are independent, and the extensions of the partial functional symmetries and conservation laws of different subnetworks g^j act on disjoint parameter spaces.

Therefore, the extensions of the partial functional symmetries and conservation laws of the subnetworks g^j form a complete set of partial functional symmetries and conservation laws of G. □

G Proof for Section 5

G.1 Proof for Proposition 18

To prove this result, we first recall (Tran et al., 2025b, Theorem 3.1), which gives an important result on the functional symmetries of self-attention-like mechanisms. This allows us to prove the compositional identifiability condition 12.

Theorem 27 (Tran et al. (2025b)). *Let D be a positive integer. Assume that for a positive integer k, matrices $A_1, A_2, \ldots, A_k \in \mathbb{R}^{D\times D}$ and $B_1, B_2, \ldots, B_k \in \mathbb{R}^{D\times D}$, we have*

$$F(X; \{A_i, B_i\}_{i=1}^k) := \sum_{i=1}^k \mathrm{softmax}(XA_iX^\top)XB_i = 0,$$

for all positive integers L and $X \in \mathbb{R}^{L\times D}$. Then, if $A_1, A_2, \ldots, A_k$ are pairwise distince, then

$$B_1 = \cdots = B_k = 0.$$

We also summarize (Marcotte et al., 2023, Section 4.1), which gives a complete characterization of indepdendent conservation laws of the matrix factorization $UV^\top$.

Lemma 28. *Let $U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}$ be full column rank matrices. Then, the following upper triangular components of the following matrix function form a complete set of independent conservation laws of the matrix factorization $\phi(U, V) = UV^\top$:*

$$H(U, V) = U^\top U - V^\top V.$$

Proof. The definition of the matrix factorization $UV^\top$ is given at the beginning of Section 4.1 in Marcotte et al. (2023). The final corollary of (Marcotte et al., 2023, Section 4.1) shows that the functions $H(U, V)$ form a complete set of independent conservation laws of the matrix factorization $UV^\top$.

Intuitively, the results here can be interpretted as follows. The matrix factorization $UV^\top$ is invariant under the group action of $\mathrm{GL}_r(\mathbb{R})$ given by

$$S \circ (U, V) = (US, V(S^\top)^{-1}).$$

The infinitesimal generators of this Lie group generate a codistribution $W^\perp = \ker D\phi$. Its orthogonal complement $W = \mathrm{range}(D\phi^\top)$ is the 'gradient flow distribution' of the matrix factorization $UV^\top$. By calculating the rank of $\mathrm{Lie}(W)$ under the assumptions of the lemma, the authors showed that the rank of $\mathrm{Lie}(\mathrm{W})^\perp$ matches the number of independent functions defined by $H(U, V)$, and hence they form a complete set of independent conservation laws. This result does not rely on the specific form of the loss function, as it is a property of the matrix factorization $UV^\top$ itself. □

We can now proceed with the proof of Proposition 18.

Proposition 18 (Grouped query attention). *Let $n_H = kn_G$ and $D_h \le D$ be positive integers. Consider the grouped query attention mechanism G with n_H heads and n_G groups with parameters*

$$\theta = (Q_1^j, \ldots, Q_k^j, K^j, V^j, O_1^j, \ldots, O_k^j)_{j=1}^{n_G} \in \left(\left(\mathbb{R}^{D\times D_h}\right)^k \times \mathbb{R}^{D\times D_h} \times \mathbb{R}^{D\times D_h} \times \left(\mathbb{R}^{D\times D_h}\right)^k\right)^{n_G},$$

where for any $X \in \cup_{L=1}^\infty \mathbb{R}^{L\times D}$,

$$G(\theta, X) = \sum_{j=1}^{n_G}\sum_{i=1}^{k} \mathrm{softmax}\left(\frac{XQ_i^j(K^j)^\top X^\top}{\sqrt{D_h}}\right) XV^j(O_i^j)^\top.$$

Let θ_0 be a point in the parameter space so that Q_i^j, K^j, V^j, O_i^j are full rank and the matrices $Q_i^j(K^j)^\top$ are pairwise distinct for $j = \overline{1, n_G}, i = \overline{1, k}$. Locally around θ_0, the following hold:

- *All functionally equivalent parameters of the grouped query attention mechanism G are given by the following partial group action of $(\mathrm{GL}_{D_h}(\mathbb{R}) \times \mathrm{GL}_{D_h}(\mathbb{R}))^{n_G}$:*

  $$(S^j, T^j)_{j=1}^{n_G} \cdot (Q_1^j, \ldots, Q_k^j, K^j, V^j, O_1^j, \ldots, O_k^j)_{j=1}^{n_G} = (\tilde{Q}_1^j, \ldots, \tilde{Q}_k^j, \tilde{K}^j, \tilde{V}^j, \tilde{O}_1^j, \ldots, \tilde{O}_k^j)_{j=1}^{n_G},$$

  where

  $$\tilde{Q}_i^j = Q_i^j S^j, \quad \tilde{K}^j = K^j(S^j)^{-\top}, \quad \tilde{V}^j = V^j T^j, \quad \tilde{O}_i^j = O_i^j(T^j)^{-\top}.$$

- *The upper triangular elements of the following matrix functions form a complete set of conservation laws with respect to any loss satisfying Assumption 13:*

  $$\sum_{i=1}^k (Q_i^j)^\top Q_i^j - (K^j)^\top K^j, \quad (V^j)^\top V^j - \sum_{i=1}^k (O_i^j)^\top O_i^j, \quad j \in \overline{1, n_G}.$$

The case $n_H = n_G$ corresponds to standard multi-head self-attention, and all results still hold.

Proof. Let Θ be an open neighborhood of θ_0 in the parameter space of G such that the assumptions hold for all $\theta \in \Theta$. Choose small open balls around the distinct attention matrices $Q_i^j (K^j)^\top$ such that they are pairwise disjoint. Shrink Θ if necessary so that for all $\theta \in \Theta$, the attention matrices $Q_i^j (K^j)^\top$ are contained in the corresponding disjoint balls. This ensures that functional symmetries induced by permutations of the heads are not allowed, and hence the attention matrices $Q_i^j (K^j)^\top$ are pairwise distinct for all $\theta \in \Theta$. This will be the local neighborhood in which we prove our results.

Step 1. Identifiability result.

The main tool is (Tran et al., 2025b, Theorem 3.1). For other networks, one would need to prove a similar result for the specific architecture.

Let

$$\overline{\theta} = (\overline{Q}_1^j, \dots, \overline{Q}_k^j, \overline{K}^j, \overline{V}^j, \overline{O}_1^j, \dots, \overline{O}_k^j)_{j=1}^{n_G} \in \Theta$$

be another set of parameters such that $G(\theta, \cdot) = G(\overline{\theta}, \cdot)$. Then, for all X with arbitrary number of tokens, each of size D, we have

$$\sum_{j=1}^{n_G} \sum_{i=1}^{k} \mathrm{softmax}\left(X \frac{Q_i^j (K^j)^\top}{\sqrt{D_h}} X^\top \right) X V^j (O_i^j)^\top - \sum_{j=1}^{n_G} \sum_{i=1}^{k} \mathrm{softmax}\left(X \frac{\overline{Q}_i^j (\overline{K}^j)^\top}{\sqrt{D_h}} X^\top \right) X \overline{V}^j (\overline{O}_i^j)^\top = 0.$$

Since the term on the left contains $n_H = k n_G$ terms with pairwise distinct attention matrices $Q_i^j (K^j)^\top$ and nonzero weight matrices $V^j (O_i^j)^\top$, Theorem 27 implies that the attention matrices and weight matrices must be equal, i.e.,

$$Q_i^j (K^j)^\top = \overline{Q}_i^j (\overline{K}^j)^\top, \quad V^j (O_i^j)^\top = \overline{V}^j (\overline{O}_i^j)^\top.$$

By the matrix factorization result (Piziak & Odell, 1999, Theorem 2) and the full rank assumption, we have that for each $j = 1, \dots, n_G$, $i = 1, \dots, k$ there exists invertible matrices $S_i^j \in \mathbb{R}^{D_h \times D_h}, T_i^j \in \mathbb{R}^{D_h \times D_h}$ such that

$$\overline{Q}_i^j = Q_i^j S_i^j, \quad \overline{K}^j = K^j ((S_i^j)^\top)^{-1}, \quad \overline{V}^j = V^j T_i^j, \quad \overline{O}_i^j = O_i^j ((T_i^j)^\top)^{-1}.$$

But since K^j, V^j are full rank, we must have

$$S_1^j = \cdots = S_k^j =: S^j, \quad T_1^j = \cdots = T_k^j =: T^j$$

for each $j = 1, \dots, n_G$.

It is easy to verify that for any invertible matrices $S^j, T^j \in \mathbb{R}^{D_h \times D_h}$, the following transformations of the parameters leave the grouped query attention mechanism G functionally invariant:

$$\overline{Q}_i^j = Q_i^j S^j, \quad \overline{K}^j = K^j ((S^j)^\top)^{-1}, \quad \overline{V}^j = V^j T^j, \quad \overline{O}_i^j = O_i^j ((T^j)^\top)^{-1}.$$

Thus, all functionally equivalent parameters of the grouped query attention mechanism G are given by the following group action of $(\mathrm{GL}_{D_h}(\mathbb{R}) \times \mathrm{GL}_{D_h}(\mathbb{R}))^{n_G}$:

$$(S^j, T^j)_{j=1}^{n_G} \cdot (Q_1^j, \dots, Q_k^j, K^j, V^j, O_1^j, \dots, O_k^j)_{j=1}^{n_G} = (\tilde{Q}_1^j, \dots, \tilde{Q}_k^j, \tilde{K}^j, \tilde{V}^j, \tilde{O}_1^j, \dots, \tilde{O}_k^j)_{j=1}^{n_G},$$

where

$$\tilde{Q}_i^j = Q_i^j S^j, \quad \tilde{K}^j = K^j ((S^j)^\top)^{-1}, \quad \tilde{V}^j = V^j T^j, \quad \tilde{O}_i^j = O_i^j ((T^j)^\top)^{-1}.$$

Step 2. Reduction to subnetworks

The main tool is Theorem 17 from our framework. It is reusable for other architectures, as long as the compositional identifiability condition 12 holds.

Stack the query matrices Q_i^j and output matrices O_i^j for each group j into the matrices

$$Q^j = \begin{pmatrix} Q_1^j \\ \vdots \\ Q_k^j \end{pmatrix} \in \mathbb{R}^{kD \times D_h}, \quad O^j = \begin{pmatrix} O_1^j \\ \vdots \\ O_k^j \end{pmatrix} \in \mathbb{R}^{kD \times D_h}.$$

Define the following functions

$$g_1^j(Q_1^j,\dots,Q_k^j,K^j) = g_1^j(Q^j,K^j) := Q^j(K^j)^\top = \begin{pmatrix} Q_1^j(K^j)^\top \\ \vdots \\ Q_k^j(K^j)^\top \end{pmatrix} \in \mathbb{R}^{kD\times D},$$

$$g_2^j(V^j,O_1^j,\dots,O_k^j) = g_2^j(V^j,O^j) = V^j(O^j)^\top = (V^j(O_1^j)^\top,\dots,V^j(O_k^j)^\top) \in \mathbb{R}^{D\times kD}.$$

By the identifiability result in Step 1, the compositional identifiability conditon 12 holds for the grouped query attention mechanism G with respect to the functions g_1^j and g_2^j for all $j = 1,\dots,n_G$. Thus, by Theorem 17, the complete set of symmetries and conservation laws of G can be obtained by combining the extensions of complete sets of symmetries and conservation laws of g_1^j and g_2^j for all $j = 1,\dots,n_G$.

Step 3. Complete characterization of conservation laws of subnetworks.

The main tool is Lemma 28. For other networks, one would need to prove a similar result for the specific architecture.

By Lemma 28, the complete set of independent conservation laws of g_1^j and g_2^j are given by

$$H_1^j(Q^j,K^j) = (Q^j)^\top Q^j - (K^j)^\top K^j = \sum_{i=1}^{k}(Q_i^j)^\top Q_i^j - (K^j)^\top K^j$$

$$H_2^j(V^j,O^j) = (V^j)^\top V^j - (O^j)^\top O^j = (V^j)^\top V^j - \sum_{i=1}^{k}(O_i^j)^\top O_i^j.$$

Applying Theorem 17, we obtain the complete set of independent conservation laws of G, as stated in the proposition. □

G.2 Proof for Proposition 19

Proposition 19 (Polynomial neural network)**.** *Consider the polynomial neural network (PNN) mechanism G with parameters $\theta = (W_j,b_j)_{j=1}^L \in \prod_{j=1}^L \left(\mathbb{R}^{D_j\times D_{j-1}}\times\mathbb{R}^{D_j}\right)$ defined by*

$$G(\theta,x) = f_L\circ\rho_{r_{L-1}}\circ f_{L-1}\circ\cdots\circ\rho_{r_1}\circ f_1(x),$$

where $\rho_{r_j}(z) = (z_1^{r_j},\dots,z_{D_j}^{r_j})$ is a monomial activation function of degree r_j and $f_j(z) = W_jz+b_j$ for $j = 1,\dots,L$. If the architecture is finite-to-one at a generic θ_0, then locally around θ_0, a complete set of conservation laws of G with respect to any loss satisfying Assumption 13 is given by

$$h_i^j(\theta) = \|(W_j)_{i,:}\|_2^2 + (b_j)_i^2 - r_j\|(W_{j+1})_{:,i}\|_2^2, \quad i = \overline{1,D_j}, j = \overline{1,L-1}.$$

Proof. We recall from Usevich et al. (2025) that a polynomial neural network with a specific parameter θ_0 is called finite-to-one if there are only finitely many functionally equivalent classes of parameters under the following transformations

$$W_l' = P_lD_lW_lD_{l-1}^{-r_{l-1}}P_{l-1}^\top, \quad b_l' = P_lD_lb_l,$$

where $P_l\in\mathbb{Z}^{D_l\times D_l}$ is a permutation matrix, $D_l\in\mathbb{R}^{D_l\times D_l}$ is a diagonal matrix, and $P_0 = D_0 = I, P_L = D_L = I$.

We first restrict our attention to a small enough neighborhood of θ_0 to eliminate all discrete symmetries of the PNN mechanism G. Let Θ be an open neighborhood of θ_0 in the parameter space of G such that it avoids all other functionally equivalent classes of parameters. Further shrink Θ if necessary so permutations by P_l, P_{l-1} are not allowed. Then, by the finite-to-one assumption, any functionally equivalent parameters $\theta'\in\Theta$ must be related to θ by a transformation of the form

$$W_l' = D_lW_lD_{l-1}^{-r_{l-1}}, \quad b_l' = D_lb_l,$$

where $D_l\in\mathbb{R}^{D_l\times D_l}$ is an invertible diagonal matrix, and $D_0 = D_L = I$. This will be the local neighborhood in which we prove our results.

In other words, the functionally equivalent parameters are given by the partial group action of $(\mathbb{R}^*)^{D_1} \times \cdots \times (\mathbb{R}^*)^{D_{L-1}}$ on the parameter space of G given by

$$(t_{l,i})_{i=\overline{1,D_l}}^{l=\overline{1,L-1}} \circ (W_l, b_l)_{l=1}^{L}$$
$$= \left((t_{l,1}E_1 + \cdots + t_{l,D_l}E_{D_l}) W_l (t_{l-1,1}^{-r_{l-1}} E_1 + \cdots + t_{l-1,D_{l-1}}^{-r_{l-1}} E_{D_{l-1}}), (t_{l,1}E_1 + \cdots + t_{l,D_l}E_{D_l}) b_l \right),$$

where $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ is the multiplicative group of nonzero real numbers and E_i is the square matrix of the appropriate size with 1 in the (i,i)-th entry and 0 elsewhere. This Lie group has dimension $s := \sum_{j=1}^{L-1} D_j$, so the number of partial independent functional symmetries is at most s. By Assumption 13 and Proposition 14, the number of partial independent loss symetries is also at most s. Finally, by Theorem 12, the number of independent conservation laws is at most s.

We will show that there exists exactly this many independent conservation laws, which then must form a complete set of conservation laws of the PNN mechanism G. Observe that the following functions are functional symmetries of the PNN mechanism G:

$$\psi_i^j(t, \theta) = (W_1, b_1, \ldots, e^{tE_i} W_j, e^{tE_i} b_j, W_{j+1} e^{-t r_l E_i}, \ldots, W_L, b_L),\ j = \overline{1, L-1},\ i = \overline{1, D_j}.$$

Their infinitesimal generators are given by

$$\chi_i^j(\theta) = D_t \psi_i^j(0, \theta) = (0, 0, \ldots, E_i W_j, E_i b_j, -r_l W_{j+1} E_i, \ldots, 0, 0),\ j = \overline{1, L-1},\ i = \overline{1, D_j}.$$

By the assumption that θ_0 is generic and the fact that each χ_i^j involves only the i-th row of W_j and the i-th column of W_{j+1}, these vector fields must be linearly independent. It can be verified that the functions

$$h_i^j(\theta) = \|(W_j)_{i,:}\|_2^2 + (b_j)_i^2 - r_j \|(W_{j+1})_{:,i}\|_2^2, \quad i = \overline{1, D_j}, j = \overline{1, L-1}$$

have gradients $\nabla h_i^j = 2\chi_i^j$. By Proposition 9 and Proposition 8, the functions h_i^j are conservation laws of the PNN mechanism G. Since their gradients ∇h_i^j are linearly independent, the functions h_i^j are also independent. Finally, since there are exactly $\sum_{j=1}^{L-1} D_j$ independent functions h_i^j, they form a complete set of conservation laws of the PNN mechanism G. □

G.3 Proof for Proposition 20

Proposition 20 (Deep square linear network)**.** *Consider the deep square linear network mechanism G with parameters $\theta = (W_j)_{j=1}^L \in \prod_{j=1}^L \mathbb{R}^{D \times D}$ defined by*

$$G(\theta, x) = W_L W_{L-1} \ldots W_1 x.$$

If $W_L \ldots W_1$ is full rank, then locally around $(W_1, \ldots, W_L)$, there are at most $(L-1)D^2$ independent conservation laws of G with respect to any loss satisfying Assumption 13.

Proof. Lemma 1 from Lindsey & Menon (2026) shows that all symmetries of the deep square linear network is given by the group action of $(\mathrm{GL}_D(\mathbb{R}))^{L-1}$ on the parameters $(W_1, \ldots, W_L)$ via

$$(S_1, \ldots, S_{L-1}) \cdot (W_1, \ldots, W_L) = (S_1 W_1, S_2 W_2 S_1^{-1}, \ldots, S_{L-1} W_{L-1} S_{L-2}^{-1}, W_L S_{L-1}^{-1}).$$

Since the dimension of $(\mathrm{GL}_D(\mathbb{R}))^{L-1}$ is $(L-1)D^2$, Theorem 12 implies that there are at most $(L-1)D^2$ independent conservation laws of G with respect to any loss satisfying Assumption 13. □

H Learning to Scale

Consider the problem of using a two layer linear network G to learn a scaling function f given by

$$G((U, V), x) = UV^\top x, \quad f(x) = \alpha x = \alpha I x.$$

We specifically focus on the cases where $U, V \in \mathbb{R}$ or $U, V \in \mathbb{R}^{2\times 2}$, and the loss function is given by the mean squared error

$$\ell((U, V), x) = \|G((U, V), x) - f(x)\|_2^2 = \|UV^\top x - \alpha x\|_2^2.$$

H.1 THE SCALAR CASE

In the scalar case, a complete set of independent conservation laws of the two layer linear network G is given by a single scalar function

$$h(U, V) = U^2 - V^2.$$

The symmetry-law gap is 0. This informs us that this conservation law should be sufficient to fully determines the local minimizer that gradient flow will converge to given any initial condition.

Indeed, assume that $\alpha > 0$, and that the weights are initialized at $U_0 > 0, V_0 > 0$. Then the conservation law value is given by

$$h(U_0, V_0) = U_0^2 - V_0^2 = h_0 \in \mathbb{R}.$$

If gradient flow converges, it will converge to a minimizer (U^*, V^*) with the same conservation law value. This gives us the following system of equations:

$$\begin{cases} (U^*)^2 - (V^*)^2 = h_0 \\ U^* V^* = \alpha, \end{cases}$$

which yields a degree 4 polynomial equation in U^* via the substitution $V^* = \alpha / U^*$. Solving this gives a unique solution in the positive quadrant (the same quadrant as the initial condition) given by

$$(U^*, V^*) = \left(\sqrt{\frac{h_0 + \sqrt{h_0^2 + 4\alpha^2}}{2}}, \frac{\alpha}{\sqrt{\frac{h_0 + \sqrt{h_0^2 + 4\alpha^2}}{2}}} \right). \tag{15}$$

Thus, we can predict that if gradient flow converges, it will converge to the unique minimizer (U^*, V^*) given by equation 15, which is fully determined by the conservation law value h_0 at initialization.

H.2 THE 2X2 MATRIX CASE

In the 2×2 matrix case, a complete set of independent conservation laws of the two layer linear network G is given by the following matrix function:

$$H(U, V) = U^\top U - V^\top V = \begin{pmatrix} h_1(U, V) & h_2(U, V) \\ h_2(U, V) & h_3(U, V) \end{pmatrix}.$$

The symmetry-law gap is 1. This informs us that the conservation laws are not sufficient to fully determine the local minimizer that gradient flow will converge to given any initial condition. Rather, gradient flow will converge to a point on a one dimensional manifold of local minimizers that share the same conservation law values.

Indeed, assume that the weights are initialized at $U_0, V_0 \in \mathbb{R}^{2\times 2}$, and that gradient flow converges to a minimizer (U^*, V^*) with the same conservation law values. This gives us the following system of equations:

$$H(U, V) = H(U_0, V_0) := H_0 \tag{16}$$

$$UV^\top = \alpha I. \tag{17}$$

Solving this system of equations yield a one dimensional manifold of solutions induced by the orthogonal group $\mathrm{O}(2)$. We present this in Lemma 29, whose proof we defer to the end of this section.

Lemma 29. *Let $H_0 \in \mathbb{R}^{2\times 2}$ be a symmetric matrix and $0 \neq \alpha \in \mathbb{R}$ be given. The entire solution set of $(U, V) \in (\mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2})$ that satisfies equation 16 and equation 17 is given by the group action of the orthogonal group $\mathrm{O}(2)$ on the solution $(P^{1/2}, \alpha(P^{1/2})^{-\top})$:*

$$Q \cdot (P^{1/2}, \alpha P^{-1/2}) = (QP^{1/2}, \alpha Q P^{-1/2}), \quad Q \in \mathrm{O}(2),$$

where $P = \frac{H_0 + \sqrt{H_0^2 + 4\alpha^2 I}}{2}$ is a positive definite matrix.

Thus, the conservation laws are not sufficient to fully determine the local minimizer that gradient flow will converge to given any initial condition. Nevertheless, they are still sufficient to determine important numerical properties of the minimizer, such as the singular values or condition numbers of U^* and V^*. In particular, since these properties are invariant under the action of the orthogonal group $\mathrm{O}(2)$, the singular values and condition numbers of U^* (resp. V^*) are the same as those of $P^{1/2}$ (resp. $\alpha P^{-1/2}$), where

$$P = \frac{H_0 + \sqrt{H_0^2 + 4\alpha^2 I}}{2} \tag{18}$$

is determined at initialization by the conservation law value H_0. Therefore, if one desires to have a well-conditioned minimizer (U^*, V^*), one should initialize the weights such that P is well-conditioned, e.g., $P \approx \lambda I$ for some $\lambda > 0$.

Proof of Lemma 29. To solve this system, note that the condition 17 implies $V = \alpha U^{-\top}$, and so equation 16 is equivalent to

$$H_0 = U^\top U - V^\top V = U^\top U - \alpha^2 U^{-1} U^{-\top}.$$

Setting the positive definite Gram matrix $P := U^\top U$, we obtain the equivalent equation

$$H_0 = P - \alpha^2 P^{-1}. \tag{19}$$

Let R be a orthogonal matrix that diagonalizes P, i.e., $R^\top P R = D$ where D is a diagonal matrix. We can show that R also diagonalizes H_0 as follows:

$$R^\top H_0 R = R^\top (P - \alpha^2 P^{-1}) R = D - \alpha^2 D^{-1}.$$

Thus, to solve for P in equation 19 is equivalent to solving for D in

$$D - \alpha^2 D^{-1} = R^\top H_0 R,$$

but this reduces to solving a quadratic equation for each diagonal entry of D, which has a unique positive diagonal matrix solution

$$D = \frac{R^\top H_0 R + \sqrt{(R^\top H_0 R)^2 + 4\alpha^2 I}}{2}.$$

By reversing the diagonalization, we obtain that

$$P = R D R^\top = \frac{H_0 + \sqrt{H_0^2 + 4\alpha^2 I}}{2}.$$

Finally, since P is the Gram matrix of U, it is known that U must be of the form $U = QP^{1/2}$ for some orthogonal matrix Q. In other words, the solution set of (U, V) that satisfies equation 16 and equation 17 must be given by the group action of the orthogonal group $\mathrm{O}(2)$ on the solution $(P^{1/2}, \alpha (P^{1/2})^{-\top})$:

$$Q \cdot (P^{1/2}, \alpha P^{-1/2}) = (QP^{1/2}, \alpha Q P^{-1/2}).$$

It is easy to verify that these are indeed solutions, and so they are all the solutions. □

I Empirical Examination of Conservation Laws in a Language Model

In this section, we empirically examine the conservation laws of the multi-head self-attention and grouped query attention mechanisms during the training of a language model with SGD. The theoretical basis for this experiment is given by Proposition 5.1 of Marcotte et al. (2025), which states that for constant step size, the conservation law error grows linearly with the number of step size.

We use a modified version of the character level nanoGPT model (Karpathy, 2022) with 6 layers, each with 6 attention heads. Layers 0, 2, 3, 5 are standard multi-head self-attention layers, while layers 1 and 4 are grouped query attention layers with 3 groups of 2 query heads. We train from scratch on the Tiny Shakespeare dataset (Karpathy, 2015) for 20,000 updates using plain SGD with a constant learning rate of 10^{-2} and no weight decay. We set dropout to 0 and use training without

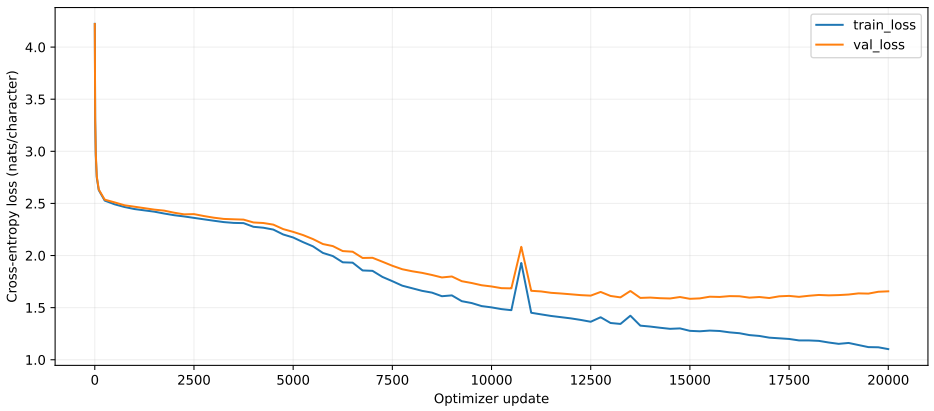


Figure 2: Training and validation loss of the language model with SGD. Both losses decrease substantially over the course of training.

KING: step 0

```
KING:iW,xV!fF,vOTTtcukUaGBS&HZ?
JGP!OOT WuurJ'uuWX-o--pGM3???J?Hh-a;aHwu&z:
     upupIIyy;-,
     NllkwmGFTOggmkHjg$yiPYqeMlbCLSYOeUM;
     JzlUIqhETTHH:c:g;Vre&dN-pugJo?
     aHeMhnfSfHH ks&jn''IIM''d$uS'S
d:qsgJB-Q:o
SSJB$,ssTI;mLs?VVoN:VN;zsm.HqgcPVWhzn
?Ghjs?u;Yje
```

KING: step 20,000

```
KING:
What from Tutualte the oratops of this
     business be.

GLOUCESTER:
The quarrel of thy smooth, many by thee,
So hated, and I have shed unrest, that have
     heard
A knotny before the brother than words,
When he has has want to woe Romans soul of
```

Figure 3: Opening excerpts from fixed-prompt samples before training and after 20,000 optimizer updates. The language model has learned to produce text that contains recognizable words and dialogue-like formatting, demonstrating that it has learned some structure of the English language.

gradient clipping to avoid any potential interference with the conservation laws. Training was performed on a single NVIDIA RTX 3080 GPU with 10GB of memory, and took less than 2 hours to complete.

The training loss and validation loss are shown in Figure 2, which shows that the model is learning effectively. Sample outputs from the language model with the prompt "KING:" before training and after 20,000 updates are shown in Figure 3.

For each layer and each group j, define the Gram matrices of the query, key, value, and output weights as

$$G_Q^j = \sum_{i \in j} Q_i^\top Q_i, \quad G_K^j = K_j^\top K_j, \qquad G_V^j = V_j^\top V_j, \quad G_O^j = \sum_{i \in j} O_i^\top O_i.$$

Following the setting in Appendix E of Tran et al. (2026), we track both the conservation law matrices and a non-conserved sum of Gram matrices as a control:

$$H_{QK}^j = G_Q^j - G_K^j, \quad H_{VO}^j = G_V^j - G_O^j, \qquad C_{QK}^j = G_Q^j + G_K^j, \quad C_{VO}^j = G_V^j + G_O^j.$$

While the conservation law matrices H_{QK}^j and H_{VO}^j are expected to be approximately preserved during training, the non-conserved sum matrices C_{QK}^j and C_{VO}^j are not expected to be preserved.

To compare the conservation law error and non-conserved sum error across different layers and groups, we normalize the errors by the initial scale of the Gram matrices

$$s_{QK}^j = \|G_Q^j(0)\|_F + \|G_K^j(0)\|_F, \qquad s_{VO}^j = \|G_V^j(0)\|_F + \|G_O^j(0)\|_F.$$

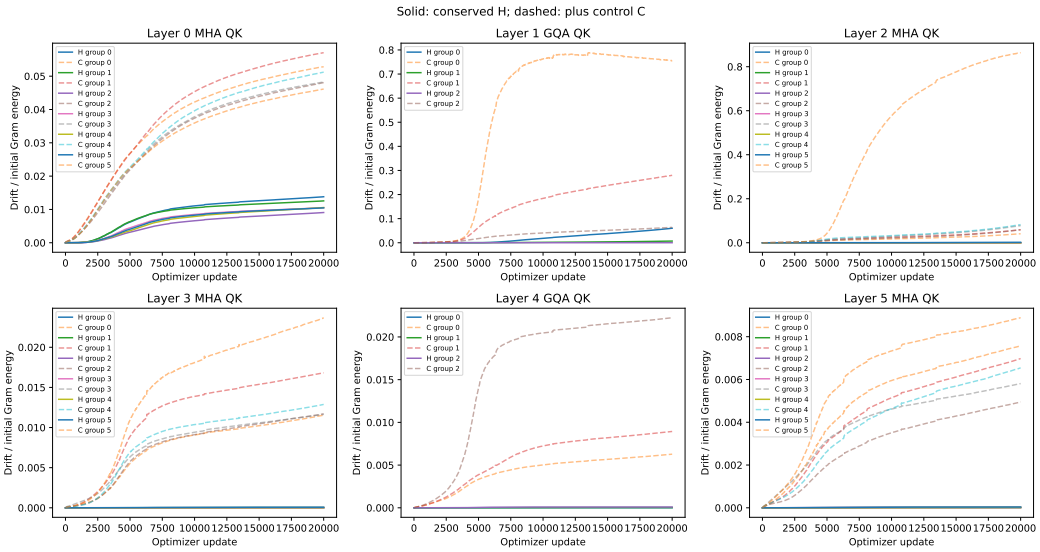


Figure 4: Conservation law error and non-conserved sum error for the Q, K blocks of the multi-head self-attention and grouped query attention mechanisms during training of a language model with SGD. Solid curves show the normalized drift of conservation laws, while dashed curves show the normalized drift of the non-conserved sum. The results show that the conservation law error drifts much less than the non-conserved sum error.

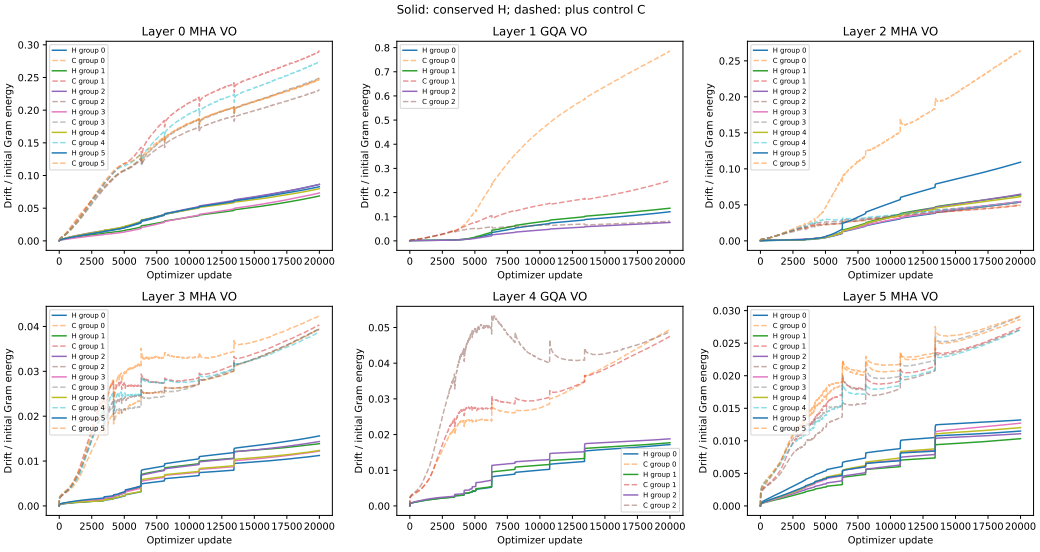


Figure 5: Conservation law error and non-conserved sum error for the V, O blocks of the multi-head self-attention and grouped query attention mechanisms during training of a language model with SGD. Solid curves show the normalized drift of conservation laws, while dashed curves show the normalized drift of the non-conserved sum. For most layers, the result shows that the conservation law error drift much less than the non-conserved sum error. However, for the V, O blocks in layers 1 and 2, this trend only holds for the first 5,000 updates.

and plot the normalized conservation law error

$$\frac{\|H^j_{QK}(t) - H^j_{QK}(0)\|_F}{s^j_{QK}}, \quad \frac{\|H^j_{VO}(t) - H^j_{VO}(0)\|_F}{s^j_{VO}}$$

as well as the normalized non-conserved sum error

$$\frac{\|C^j_{QK}(t) - C^j_{QK}(0)\|_F}{s^j_{QK}}, \quad \frac{\|C^j_{VO}(t) - C^j_{VO}(0)\|_F}{s^j_{VO}}.$$

The implementation adds 10^{-12} to each denominator to avoid division by zero.

We report the results in Figure 4 for the Q, K blocks and Figure 5 for the V, O blocks. We observe that many conservation law matrices exhibit a smaller drift than the non-conserved sum matrices, particularly for the Q, K blocks. However, for the V, O blocks in layers 1 and 2, this trend only holds for the first 5,000 updates.

This experiment illustrates that the conservation laws can be approximately preserved during training with SGD, but the degree of preservation can vary across different layers and blocks of the model, and may deteriorate over time.