Keep the Momentum: Conservation Laws beyond Euclidean Gradient Flows

Sibylle Marcotte [1] **Rémi Gribonval** [2] **Gabriel Peyré** [1 3]

Abstract

Conservation laws are well-established in the context of Euclidean gradient flow dynamics, notably for linear or ReLU neural network training. Yet, their existence and principles for non-Euclidean geometries and momentum-based dynamics remain largely unknown. In this paper, we characterize "all" conservation laws in this general setting. In stark contrast to the case of gradient flows, we prove that the conservation laws for momentum-based dynamics exhibit temporal dependence. Additionally, we often observe a "conservation loss" when transitioning from gradient flow to momentum dynamics. Specifically, for linear networks, our framework allows us to identify all momentum conservation laws, which are less numerous than in the gradient flow case except in sufficiently over-parameterized regimes. With ReLU networks, no conservation law remains. This phenomenon also manifests in non-Euclidean metrics, used e.g. for Nonnegative Matrix Factorization (NMF): all conservation laws can be determined in the gradient flow context, yet none persists in the momentum case.

1. Introduction

Discovering functions that remain unchanged during the optimization of neural networks is important to gain insight into the properties of trained models. While these laws are understood in Euclidean gradient flows, they are much less studied in non-Euclidean metrics or momentum dynamics.

Conservation laws of Euclidean gradient flows are known and extensively used to study the training of linear and ReLU networks without momentum. These laws correspond to "balancedness properties" between neurons across layers (Saxe et al., 2013; Du et al., 2018; Arora et al., 2019). They can be leveraged to understand which specific attributes (e.g. sparsity, low-rank, etc.) the optimization process tends to select from the potentially infinite pool of solutions (Saxe et al., 2013; Bah et al., 2022; Arora et al., 2018; Tarmoun et al., 2021; Min et al., 2021). They can also be used to prove under restrictive conditions the global convergence of gradient descent (Du et al., 2018; Arora et al., 2019; Bah et al., 2022; Chizat & Bach, 2020; Ji & Telgarsky, 2019; Min et al., 2021). For linear and ReLU neural networks, the number of conservation laws for a gradient flow dynamic is determined by the dimension of a Lie algebra and there are no more conservation laws than these "balancedness" laws (Marcotte et al., 2024).

Momentum and non-Euclidean metrics. Momentum and non-Euclidean metrics are two key ideas to accelerate the convergence of optimization schemes, enable the use of larger step sizes, and take into account constraints on the weights. The initial idea dates back to Polyak's heavy ball (Polyak, 1964), which introduces a momentum in the gradient descent algorithm to perform an extrapolation step. In the continuous-time limit of using small step sizes, this corresponds to using a second-order differential equation. Nesterov's acceleration (Nesterov, 1983) goes one step further by progressively increasing the momentum strength during the dynamics, reaching a faster convergence rate on the class of smooth functions. A complementary idea to better capture the curvature of the loss is to use spatially varying metrics. The simplest cases are data-independent metrics, which are Hessian of some potential function (Raskutti & Mukherjee, 2015). This corresponds to the continuous-time counterpart to the mirror descent algorithm (Nemirovskij & Yudin, 1983), which is closely related to optimization schemes using Bregman's divergences (Bregman, 1967). Another advantage of these non-Euclidean metrics is that they can naturally enforce constraints, such as positivity when using the mirror descent metric associated with the Shannon entropy potential (Bubeck et al., 2015). Data-dependent metrics estimate the local curvature using variations around the idea of natural gradient (Amari, 1998). They are popular for training neural networks, using efficient low-rank approximations of the metric (Martens, 2010; Martens & Sutskever, 2012; Martens & Grosse, 2015).

[1]ENS - PSL Univ. [2]Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP. [3]CNRS. Correspondence to: Sibylle Marcotte <sibylle.marcotte@ens.fr>.

Proceedings of the 41st International Conference on Machine Learning,Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Conservation laws and momentum. In sharp contrast with first-order flows, conservation laws of momentum flows and non-Euclidean metrics remain mostly unexplored. The simplest approach to derive conservation laws is to apply Noether's theorem (Noether, 1918) to transformations leaving the loss invariant. Leveraging Lagrangian's formulations of the flows (Wibisono et al., 2016), this leads to preserving some form of inertial quantities (Tanaka & Kunin, 2021). While gradient flows can be seen as a small-momentum limit of second-order flow, this limit is singular. In particular, invariances of the loss do not immediately lead to conservation laws for gradient flows (Zhao et al., 2023; Tanaka & Kunin, 2021; Kunin et al., 2021). One of the goals of this paper is to expose the fundamental differences in the structure and number of the conservation laws for gradient and momentum flows.

Contributions. We define the **concept of conservation laws for Momentum flows** (Section 2.3.2) and show how to extend the framework from paper (Marcotte et al., 2024) for non-Euclidean gradient flows and momentum flow settings (Proposition 2.9). We prove the **time independence of conservation laws in the gradient flow case**, and **a non-trivial time dependence in the momentum case** (Theorem 2.1). We uncover **new conservation laws for linear networks in the Euclidean momentum case** (Theorem 4.1). These new laws **are complete**, as proved theoretically for depth-two cases (Theorem 4.3, Proposition 4.2), and algorithmically through formal computations for deeper cases (Section 4.1); We show that, in contrast, there is **no conservation law for ReLU networks in the Euclidean momentum case**, as proved theoretically for two-layer cases, and by formal computations for deeper cases (Section 4.2). We shed light on a quasi systematic **loss of conservation** when transitioning from Euclidean gradient flows to the Euclidean momentum setting (Section 4); this loss also occurs in a non-Euclidean context, such as in Non-negative Matrix Factorization **(NMF)** or for Input Convex Neural Networks (**ICNN**) implemented with two-layer ReLU networks where we find out **new conservation laws for gradient flows** (Theorem 4.5 and Theorem 4.6, and find **none in the momentum case** (See Section 4.3, Section 4.4) ; **We obtain new conservation laws in the Natural Gradient Flow case** (Section 4.5).

2. Conservation Laws for Momentum Flows

We formalize the concept of conservation laws for momentum dynamics and establish their generic time-dependence properties. We characterize their most important properties in relation to certain linear spaces of vector fields.

2.1. Momentum dynamics

We explore learning problems with features $x_i \in \mathbb{R}^m$ and targets $y_i \in \mathcal{Y}$ (typically for regression, with $\mathcal{Y} = \mathbb{R}^n$) or labels (for classification) within the scope of supervised learning. In the context of unsupervised or self-supervised learning, the y_i can be treated as constant. We denote $z_i := (x_i, y_i)$ and $Z = (x_i, y_i)_i$. Prediction is accomplished through a parametric function $g(\theta, \cdot) : \mathbb{R}^m \to \mathbb{R}^n$ (such as a neural network), which is trained by empirically minimizing a **cost** over the parameters $\theta \in \Theta \subseteq \mathbb{R}^D$

$$\mathcal{E}_Z(\theta) := \sum_i \ell(g(\theta, x_i), y_i), \tag{1}$$

with ℓ a **loss** function. The parameter set $\Theta \subseteq \mathbb{R}^D$ is typically either the whole parameter space $\mathbb{R}^D$ or an open set of "non-degenerate" parameters for linear or ReLU networks.

This paper studies quantities that are preserved during the minimization of $\mathcal{E}_Z$ defined in (1) using dynamical flows. First-order dynamics corresponds to gradient flows

$$\dot{\theta}(t) = -M_Z(\theta(t))\nabla\mathcal{E}_Z(\theta(t)) \tag{2}$$

where $M_Z(\theta) \in \mathbb{R}^{D\times D}$ is typically a positive semi-definite matrix. Conservation laws $h(\theta)$ for these flows have been studied in-depth in the Euclidean case (M is the identity matrix Id_D) (Marcotte et al., 2024), and we study what happens in a non-Euclidean setting. Another goal is to go beyond first-order dynamics and to analyze which functions $h(t, \theta, \dot{\theta})$ are preserved during the momentum flow of $\mathcal{E}_Z$:

$$\ddot{\theta}(t) + \tau(t)\dot{\theta}(t) = -M_Z(t, \theta(t), \dot{\theta}(t))\nabla\mathcal{E}_Z(\theta(t)). \tag{3}$$

We will always assume that $t \mapsto \tau(t)$ is infinitely smooth. To anticipate our findings, let us immediately mention that introducing momentum leads to *conserved quantities that depend on time and velocity*, and results in many cases of interest in a *reduction of conservation properties*. The latter phenomenon consistently emerges across all our examples, whether in Euclidean ($M = \mathrm{Id}_D$) or non-Euclidean settings, such as in non-negative matrix factorization. Consequently, we will draw comparisons with the gradient flow scenario.

2.2. Main Examples

We consider several settings of practical interest.

Examples of models. Prime examples are ***two-layer linear or ReLU networks***, where $\theta = (U, V)$ with matrices U, V of appropriate sizes and $g(\theta, x) := U\sigma(V^\top x)$. Here $\sigma = \mathtt{id}$ for the linear case, while σ is the entrywise ReLU activation function for ReLU networks. In the latter case, deeper examples as well as biases can also be considered.

Example of flows. In terms of flows, we consider: **Gradient flows (GF)**, corresponding to (2). This can informally

be thought of as using $\tau = \infty$ in (3) (noticing that the matrix M_Z in (3) can depend on τ via t). **Heavy ball**, i.e. (3) with a fixed $\tau(t) = \tau < \infty$. This corresponds to a continuous-time limit of Poliak's heavy ball acceleration method (Polyak, 1964). **Nesterov acceleration**, i.e. (3) with $\tau(t) = 3/t$. This corresponds to the flow introduced by (Su et al., 2016) as a continuous-time limit of Nesterov accelerated gradient descent scheme (Nesterov, 1983).

Example of metrics. We illustrate our findings on: **Euclidean geometry**, i.e. $M_Z(t, \theta, \dot\theta) = \mathrm{Id}_D$. **Mirror geometry**, associated to the Shannon entropy potential, uses $M_Z(\theta) = \mathrm{diag}(\theta)$ for gradient flows, and in the Heavy ball case, it uses $M_Z(t, \theta, \dot\theta) = \mathrm{diag}(\dot\theta + \tau\theta)$ (Wibisono et al., 2016). The associated flow is a continuous time limit of mirror descent (Nemirovskij & Yudin, 1983). While such approaches were initially developed for first-order schemes, their extension to second-order flows that we use is derived in (Wibisono et al., 2016) as a flow for a Bregman-type Lagrangian. If this paper particularly focuses on the case of the Shannon entropy potential as an example, note that our theory applies to any mirror potential. **Natural gradient** (Amari, 1998) avoids the issues of Newton's method using a first-order estimation of the curvature (hence the naming "Hessian free"). Assuming again $\tau(t) = \tau$ is constant, it uses a data-dependent metric $M_Z(t, \theta, \dot\theta) = H_Z(\dot\theta + \tau\theta)^\dagger$, where $A^\dagger$ denotes the pseudo-inverse of A and where (for the mean square loss function) $H_Z(\xi) := \frac{1}{n}\sum_i \partial_1 g(\xi, x_i)^\top \partial_1 g(\xi, x_i)$ is a proxy for the Hessian that captures the curvature of the loss. Here $\partial_1 g$ is the differential of g with respect to its first variable.

Running examples. We consider several examples. **Principal component analysis (PCA)** corresponds to linear neural networks ($\sigma = \mathtt{id}$) with Euclidean geometry to perform dimensionality reduction via matrix factorization. **Multilayer Perceptrons (MLP)** use σ =ReLU and Euclidean geometry. **Non-negative matrix factorization (NMF)** uses a linear network (i.e. σ =Id) with mirror geometry (for the Shannon entropy) to impose positivity on the factors (Lee & Seung, 1999). **Input Convex Neural Networks (ICNN)** (Amos et al., 2017) use a hybrid Euclidean/Mirror (for the Shannon entropy) geometry with σ =ReLU to impose positivity on some weights, to represent convex functions. It finds applications in implicit deep learning (Amos et al., 2017) and to compute optimal transport in high dimension (Makkuva et al., 2020).

This table indicates the number of conservation laws that we characterize (for two layers and r hidden neurons) for gradient flow ($\tau(t) = \infty$) and for momentum flows ($\tau(t) < +\infty$).

metric \ σ	Linear		ReLu	
Euclid. $\mathbb{R}$	PCA $\frac{r(r+1)}{2}$	$\frac{r(r-1)}{2}$	MLP r	0
Mirror ≥ 0	NMF r	0	ICNN r	0

For the example of PCA, the information presented in the table corresponds to the case where r is small enough, and the general case is fully addressed in Proposition 4.4.

2.3. Conserved functions

In the GF (resp. MF – momentum flow) scenario, a function $h(t, \theta)$ (resp. $h(t, \theta, \dot\theta)$) is conserved if for each solution[1] $\theta(\cdot)$ to the ODE (2) (resp. (3)) with arbitrary initialization, the quantity $h(t, \theta(t))$ (resp. $h(t, \theta(t), \dot\theta(t))$) remains constant in time.

2.3.1. Time-dependence: GF *vs* MF

We postpone the formal definition of conserved functions and their characterization to first highlight an important fact (See Appendix A for a proof): in the momentum setting conservation laws can depend non-trivially on time. This is in sharp contrast with the gradient flow case.

Theorem 2.1 (Structure theorem)**.** *Let $h(t, \theta)$ (resp. $h(t, \theta, \dot\theta)$) be a conserved function for the ODE (2) (resp. (3) with $\tau(t) = \tau$) when its right-hand side is zero. For all t and θ, one has $h(t, \theta) = h(0, \theta)$ (resp. $h(t, \theta, \dot\theta) = H(\theta + \frac{\dot\theta}{\tau}, \dot\theta \exp(\tau t))$, where $H(a, b) := h(0, a - \frac{b}{\tau}, b)$). In particular, conserved functions in the MF scenario can be expressed with $2D$ variables (instead of $2D + 1$).*

Conservation laws (to be soon formally defined) are notably conserved functions of the ODE when the right-hand side is zero, hence the above theorem directly applies.

2.3.2. Formal definition via phase-space lifting

To formally define conserved functions in the momentum case, we generalize the related notions from (Marcotte et al., 2024) to encompass dependencies on time t and velocity $\dot\theta$.

Notations Given an open subset $\Omega \subseteq \Theta \subseteq \mathbb{R}^D$, we denote $\tilde\Omega := \mathbb{R} \times \Omega \times \mathbb{R}^D \subseteq \mathbb{R}^{2D+1}$. In particular $\tilde\Theta := \mathbb{R} \times \Theta \times \mathbb{R}^D$.

Definition 2.2 (Conservation through a flow)**.** Consider an open subset $\Omega \subseteq \mathbb{R}^D$ and a function $F \in \mathcal{C}^1(\tilde\Omega, \mathbb{R}^{2D})$. By the Cauchy-Lipschitz theorem, for each initial condition $\mathtt{init} := (t_{\text{init}}, \omega_{\text{init}}) \in \tilde\Omega$, there exists a unique maximal solution $t \in (t_{\text{init}} - \eta_{\text{init}}, t_{\text{init}} + \eta_{\text{init}}) \mapsto \omega(t, \mathtt{init})$ of the ODE $\dot\omega(t) = F(t, \omega(t))$ with $\omega(t_{\text{init}}) = \omega_{\text{init}}$. A function $h : \tilde\Omega \subseteq \mathbb{R}^{2D+1} \to \mathbb{R}$ is *conserved on Ω through the flow F* if $h(t, \omega(t, \mathtt{init})) = h(t_{\text{init}}, \omega_{\text{init}})$ for each choice of $\mathtt{init} \in \tilde\Omega$ and every $t \in (t_{\text{init}} - \eta_{\text{init}}, t_{\text{init}} + \eta_{\text{init}})$.

Rewrite (3) as $\dot\omega(t) = F(t, \omega(t))$ with $\omega := (\theta, \dot\theta)$ and

$$F_Z\left(t, \begin{pmatrix}\theta \\ \dot\theta\end{pmatrix}\right) := \begin{pmatrix}\dot\theta \\ -\tau(t)\dot\theta - M_Z(t, \theta, \dot\theta)\nabla\mathcal{E}_Z(\theta)\end{pmatrix}. \quad (4)$$

With this expression, Definition 2.2 can be specialized.

[1]The existence of such a solution is discussed in Definition 2.2.

Definition 2.3 (Conservation during the flow (3) with a given dataset)**.** Given an open subset $\Omega \subseteq \Theta \subseteq \mathbb{R}^D$ and a dataset Z such that $(t, \theta, \dot{\theta}) \mapsto M_Z(t, \theta, \dot{\theta}) \nabla \mathcal{E}_Z(\theta) \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^D)$, a function $h : \tilde{\Omega} \subseteq \mathbb{R}^{2D+1} \to \mathbb{R}$ is *conserved on* Ω *during the flow* (3) if it is conserved through F_Z.

The next definition allows us to study which functions are conserved during *"all" flows* defined by the ODE (3). The smoothness assumptions will enable simpler characterizations of such functions in due time.

Definition 2.4 (Conservation during the flow (3) with "any" dataset)**.** Consider an open subset $\Omega \subset \Theta$ and a loss $\ell(z, y)$ such that $\ell(\cdot, y)$ is $\mathcal{C}^2$-differentiable for all $y \in \mathcal{Y}$. A function $h : \tilde{\Omega} \subseteq \mathbb{R}^{2D+1} \to \mathbb{R}$ is *conserved on* Ω *during the flow* (3) *for any data set* if, for each data set Z *such that* $g(\cdot, x_i) \in \mathcal{C}^2(\Omega, \mathbb{R})$ for each i and $(t, \theta, \dot{\theta}) \mapsto M_Z(t, \theta, \dot{\theta}) \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^D)$, the function h is conserved on Ω during the flow (3).

While the above definitions are local, we are rather interested in functions defined *for the whole parameter space* Θ, hence the following definition which mimics its equivalent in the gradient flow case (Marcotte et al., 2024).

Definition 2.5. A function $h : \tilde{\Theta} \mapsto \mathbb{R}$ is *locally conserved during the flow* (3) *on* Θ *for any data set* if for each open subset $\Omega \subseteq \Theta$, h is conserved on Ω for any data set.

Example 2.6. As a first simple example, consider a two-layer *linear* neural network in dimension 1 (both for the input and output), with two hidden neurons. In that case, the parameter is $\theta = (u_1, u_2, v_1, v_2) \subseteq \mathbb{R}^4 = \Theta$ and the model writes $g(\theta, x) = (u_1 v_1 + u_2 v_2)x$. Computing the derivative of $t \mapsto h(t, \theta(t), \dot{\theta}(t))$, where $h(t, \theta, \dot{\theta}) := u_1 \dot{u}_2 - \dot{u}_1 u_2 + v_1 \dot{v}_2 - \dot{v}_1 v_2$, one can directly check that it vanishes, hence h is *locally conserved on* $\tilde{\Theta} = \mathbb{R}^9$ *for any data set during the flow* (3) *for* $M = \mathrm{Id}_4$ *and* $\tau = 0$.

A characterization of $\mathcal{C}^1$ conserved functions (see proof in Appendix B) is the "orthogonality" of their gradients to an associated vector field. This is the momentum analog to a similar property for gradient flows (Marcotte et al., 2024).

Proposition 2.7 (Smooth functions conserved through a given flow)**.** *Given* $F \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^{2D})$*, a function* $h \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R})$ *is conserved through the flow induced by* F *if and only if* $\langle \nabla h(\alpha), (1, F(\alpha)^\top)^\top \rangle = 0$ *for all* $\alpha \in \tilde{\Omega}$.

The following characterization is proved in Appendix C.

Proposition 2.8. *A function* $h : \tilde{\Theta} \mapsto \mathbb{R} \in \mathcal{C}^1(\tilde{\Theta}, \mathbb{R})$ *is locally conserved on* Θ *for any data set if and only if* $\nabla h(\alpha) \perp W_\alpha^{\mathtt{mom}}$ *for all* $\alpha \in \tilde{\Theta}$*, where for all* $\alpha := (t, \theta, \dot{\theta}) \in \tilde{\Theta}$*:*

$$W_\alpha^{\mathtt{mom}} := \operatorname*{span}_{Z \in \mathcal{Z}_\theta} \left\{ (1, F_Z(\alpha)^\top)^\top \right\},$$

with $\mathcal{Z}_\theta$ *the set of all data sets* $Z = (x_i, y_i)_i$ *such that there exists a neighborhood* Ω *of* θ *such that for all* i*,* $g(\cdot, x_i) \in \mathcal{C}^2(\Omega, \mathbb{R}^n)$ *and such that* $M_Z(\cdot) \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^{D \times D})$.

The subspace $W_\alpha^{\mathtt{mom}} \subseteq \mathbb{R}^{2D+1}$ (that characterizes conserved functions for a momentum flow) is linked with its counterpart in a Euclidean GF dynamic (Marcotte et al., 2024), a subspace denoted $W_\theta \subseteq \mathbb{R}^D$ where $\alpha = (t, \theta, \dot{\theta})$. The space W_θ characterizes conserved functions for a gradient flow in an Euclidean setting, and its definition is recalled in Appendix D where the following proposition is proved.

Proposition 2.9. *Assume that* $M_Z(t, \theta, \dot{\theta}) = M(t, \theta, \dot{\theta})$ *does not depend on the data set* Z *and that* $M(\cdot) \in \mathcal{C}^1(\tilde{\Theta}, \mathbb{R}^{D \times D})$*. Assume that for each* $y \in \mathcal{Y}$ *the loss* $\ell(z, y)$ *is* $\mathcal{C}^2$*-differentiable with respect to* $z \in \mathbb{R}^n$ *and that for each* $\theta \in \Theta$*, there exists a training dataset* $Z = (x_i, y_i)_i$ *such that* $\nabla \mathcal{E}_Z(\theta) = 0$ *and such that for all* i*,* $\theta \mapsto g(\theta, x_i)$ *is* $\mathcal{C}^2$ *in a neighborhood of* θ*. Then, for each* $\alpha := (t, \theta, \dot{\theta}) \in \tilde{\Theta}$*:*

$$W_\alpha^{\mathtt{mom}} = \mathbb{R} \begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ M(t, \theta, \dot{\theta}) W_\theta \end{pmatrix}. \tag{5}$$

When the subspace W_θ for Euclidean gradient flows is known (Marcotte et al., 2024), the above link allows us to leverage this knowledge in *non-Euclidean, momentum* flow scenarios. Extensions to matrices M_Z that *depend* on the dataset Z, e.g. with natural gradient metrics, are briefly discussed (in the gradient case) in Section 4.5.

Remark 2.10. The assumption on the loss ℓ in Proposition 2.9 holds for classical losses, see Appendix D.

Example 2.11. Revisiting Example 2.6, we know that $h(t, \theta, \dot{\theta}) := u_1 \dot{u}_2 - \dot{u}_1 u_2 + v_1 \dot{v}_2 - \dot{v}_1 v_2$ is conserved for any data set (with $M = \mathrm{Id}_4$) during (3), hence by Proposition 2.8 one has for each $\alpha = (t, \theta, \dot{\theta})$: $\nabla h(\alpha) \perp W_\alpha^{\mathtt{mom}}$. By (5), one has in particular $\nabla h(\alpha) = (0, \dot{u}_2, -\dot{u}_1, \dot{v}_2, -\dot{v}_1, -u_2, u_1, -v_2, v_1)^\top \perp (1, \dot{u}_1, \dot{u}_2, \dot{v}_1, \dot{v}_2, 0, \ldots, 0) = \begin{pmatrix} 1 \\ \dot{\theta} \\ 0 \end{pmatrix}$ (remember that $\tau = 0$). This will be further elaborated on in Example 2.18.

The analog of Proposition 2.9 for gradient flows was established in (Marcotte et al., 2024) in the Euclidean case, and is naturally extended to the non-Euclidean case.

Proposition 2.12 (Locally conserved function for any data set for (2))**.** *Assume* $M_Z(\cdot) \in \mathcal{C}^1(\Theta, \mathbb{R}^D)$*. A function* $h : \Theta \mapsto \mathbb{R} \in \mathcal{C}^1(\Theta, \mathbb{R})$ *is locally conserved on* Θ *for any data set during the flow* (2) *if and only if* $\nabla h(\theta) \perp W_\theta^{grad} := \mathrm{span}\{M_Z(\theta) \nabla \mathcal{E}_Z(\theta) : Z \in \mathcal{Z}'_\theta\}$*, with* $\mathcal{Z}'_\theta$ *the set of all data sets* $Z = (x_i, y_i)_i$ *such that there exists a neighborhood* Ω *of* θ *such that for all* i*,* $g(\cdot, x_i) \in \mathcal{C}^2(\Omega, \mathbb{R}^n)$*. When* M_Z *does not depend on* Z*, we simple have* $W_\theta^{grad} = M(\theta) W_\theta$.

2.4. From conserved functions to conservation laws

To provide an algorithmic procedure to determine these functions, (Marcotte et al., 2024) makes the fundamental hypothesis that the **model** $g(\theta, x)$ can be (locally) factored via a **reparametrization** ϕ as $f(\phi(\theta), x)$. They require that the model $g(\theta, x)$ satisfies the following central assumption.

Assumption 2.13 (Local reparameterization)**.** There exists a dimension d and a function $\phi \in \mathcal{C}^2(\Theta, \mathbb{R}^d)$ such that: for each parameter θ_0 in the open set $\Theta \subseteq \mathbb{R}^D$, for each $x \in \mathcal{X}$ such that $\theta \mapsto g(\theta, x)$ is $\mathcal{C}^2$ in a neighborhood of θ_0, there is a neighborhood Ω of θ_0 and a function $f(\cdot, x) \in \mathcal{C}^2(\phi(\Omega), \mathbb{R}^n)$ such that

$$\forall \theta \in \Omega, \quad g(\theta, x) = f(\phi(\theta), x). \tag{6}$$

Moreover, (Marcotte et al., 2024) shows that such reparametrizations exist for linear and layered ReLU neural networks of any depth. They are respectively denoted $\phi_{\texttt{Lin}}(\cdot)$ and $\phi_{\texttt{ReLU}}(\cdot)$, and detailed in (Marcotte et al., 2024, Examples 2.10, 2.11 and Appendix C). We recall the expression of such reparametrizations in the two-layer case.

Example 2.14. (Factorization for two-layer *linear* neural networks) In the two-layer case, with r neurons, denoting $\theta = (U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ (so that $D = (n+m)r$), we can factorize $g(\theta, x) := UV^\top x$ by the reparametrization $\phi_{\texttt{Lin}}(\theta) := UV^\top \in \mathbb{R}^{n\times m}$ (identified with $\mathbb{R}^d$, $d = mn$).

Example 2.15 (Factorization for two-layer ReLU networks). Consider $g(\theta, x) = \left(\sum_{j=1}^r u_{k,j}\sigma(\langle v_j, x\rangle + b_j) + c_k\right)_{k=1}^n$, with $\sigma(t) := \max(t, 0)$ the ReLU activation function and $v_j \in \mathbb{R}^m$, $u_{k,j} \in \mathbb{R}$, $b_j, c_k \in \mathbb{R}$. Then, denoting $\theta = (U, V, b, c)$ with $U = (u_{k,j})_{k,j} =: (u_1, \cdots, u_r) \in \mathbb{R}^{n\times r}$, $V = (v_1, \cdots, v_r) \in \mathbb{R}^{m\times r}$, $b = (b_1, \cdots, b_r)^\top \in \mathbb{R}^r$ and $c = (c_1, \cdots, c_n) \in \mathbb{R}^n$ (so that $D = (n+m+1)r + n$), we can locally factorize $g(\theta, x)$ by the reparametrization: $\phi_{\texttt{ReLU}}(\theta) = ((u_j v_j^\top, u_j b_j)_{j=1}^r, c)$. In particular, in the case without bias ($(b, c) = (0, 0)$), the reparametrization is defined by $\phi_{\texttt{ReLU}}(\theta) = (\phi_j)_{j=1}^r$ where $\phi_j = \phi_j(\theta) := u_j v_j^\top \in \mathbb{R}^{n\times m}$ (here $d = rmn$): the reparametrization $\phi_{\texttt{ReLU}}(\theta)$ contains r matrices of size $m \times n$ (each of rank at most one).

Thanks to this reparametrization and under a mild assumption on the loss ℓ, Marcotte et al. (2024, Theorem 2.14) show that for linear neural networks of any depth (resp. for two-layer ReLU neural networks), the functions that are locally conserved for any data set in a Euclidean gradient flow dynamic are entirely characterized by the *trace* $\mathcal{W}(\theta) := \text{span}\{w(\theta) : w(\cdot) \in \mathcal{W}\} \subseteq \mathbb{R}^D$ of a finite-dimensional linear space of functions, $\mathcal{W}$, determined by $\phi_{\texttt{Lin}}$ (resp. $\phi_{\texttt{ReLU}}$). For these cases, they show that for all $\theta \in \Theta$, the space W_θ defined in Appendix D satisfies $W_\theta = \mathcal{W}_\phi(\theta)$ with

$$\mathcal{W}_\phi := \text{span}\{\nabla\phi_1(\cdot), \cdots, \nabla\phi_d(\cdot)\}, \tag{7}$$

where the components $\phi_i : \mathbb{R}^D \to \mathbb{R}$, $1 \le i \le d$ of ϕ are assumed to satisfy $\nabla\phi_i \in \mathcal{C}^1(\Theta, \mathbb{R}^D)$.

Assuming that M does not depend on Z, in the following theorem, we combine (Marcotte et al., 2024, Theorem 2.14) and Proposition 2.9 to show that, under an assumption on the loss ℓ, $W_\alpha^{\texttt{mom}} \subseteq \mathbb{R}^D$ is the trace of $\mathcal{W}_\phi^{\texttt{mom}} := \text{span}\{\chi_i(\cdot) : 0 \le i \le d\} \subseteq \mathcal{C}^1(\tilde{\Theta}, \mathbb{R}^D)$ where

$$\chi_0(\alpha) := \begin{pmatrix} 1 \\ \dot\theta \\ -\tau(t)\dot\theta \end{pmatrix}; \chi_i(\alpha) := \begin{pmatrix} 0 \\ 0 \\ M(t,\theta,\dot\theta)\nabla\phi_i(\theta) \end{pmatrix}, i \ge 1, \tag{8}$$

with $\alpha = (t, \theta, \dot\theta)$. Similarly the subspace $W_\theta^{\texttt{grad}}$ of Proposition 2.12 is the trace of

$$\mathcal{W}_\phi^{\texttt{grad}} := \text{span}_i\{M(\cdot)\nabla\phi_i(\cdot)\} \subseteq \mathcal{C}^1(\Theta, \mathbb{R}^D). \tag{9}$$

Theorem 2.16. *Assume that the loss $(z, y) \mapsto \ell(z, y)$ is $\mathcal{C}^2$-differentiable with respect to $z \in \mathbb{R}^n$ for each $y \in \mathcal{Y}$, and that it satisfies the condition:*

$$\text{span}_{y\in\mathcal{Y}}\{\nabla_z \ell(z, y)\} = \mathbb{R}^n, \forall z \in \mathbb{R}^n. \tag{10}$$

Assume that $M(\cdot) \in \mathcal{C}^1(\tilde{\Theta}, \mathbb{R}^{D\times D})$ (resp. $M(\cdot) \in \mathcal{C}^1(\Theta, \mathbb{R}^{D\times D})$). Then, for linear neural networks, one has for all $\alpha \in \tilde{\Theta}$ with $\Theta := \mathbb{R}^D$ (resp. for all $\theta \in \Theta$:

$$W_\alpha^{\texttt{mom}} = \mathcal{W}_{\phi_{\texttt{Lin}}}^{\texttt{mom}}(\alpha) \quad (\textit{resp. } W_\alpha^{\texttt{mom}} = \mathcal{W}_{\phi_{\texttt{Lin}}}^{grad}(\theta)). \tag{11}$$

The same holds for two-layer ReLU networks with $\phi_{\texttt{ReLU}}$ from Example 2.15 and Θ the (open) set of all parameters θ such that hidden neurons are associated to pairwise distinct "hyperplanes" (see (Marcotte et al., 2024, Theorem 2.8)).

Condition (10) is the same as in (Marcotte et al., 2024) and is satisfied for standard losses such as the quadratic loss. Theorem 2.16 motivates the following definition.

Definition 2.17 (Conservation *law* for (3))**.** A real-valued function $h \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R})$ is a *conservation law of* ϕ for (3) if for all $\alpha := (t, \theta, \dot\theta) \in \tilde{\Omega}$, $\nabla h(\alpha) \perp \mathcal{W}_\phi^{\texttt{mom}}(\alpha)$.

Combining Proposition 2.8 and Theorem 2.16, the functions $h \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R})$ that are locally conserved on Θ for any data set are exactly the conservation laws of $\phi_{\texttt{Lin}}$ (resp. of $\phi_{\texttt{ReLU}}$ for linear (resp. ReLU) two-layer networks).

Example 2.18. Revisiting again Example 2.6, the model g is factorized by $\phi_{\texttt{Lin}}(\theta) := (u_1, u_2) \times (v_1, v_2)^\top = u_1v_1 + u_2v_2 \in \mathbb{R}$ with $\theta := (u, v) \in \Theta = \mathbb{R}^4$. We know that $h(t, \theta, \dot\theta) := u_1\dot u_2 - \dot u_1 u_2 + v_1\dot v_2 - \dot v_1 v_2$ is locally conserved during (3) (with $M = \text{Id}_4$) on $\tilde{\Theta}$ for any data set and indeed: by Example 2.11, we already have $\nabla h(\alpha) \perp \chi_0(\alpha)$ for all $\alpha := (t, \theta, \dot\theta)$ (remember that $\tau = 0$) and we also have $\nabla h(\alpha) = (0, \dot u_2, -\dot u_1, \dot v_2, -\dot v_1, -u_2, u_1, -v_2, v_1)^\top \perp (0, \ldots, 0, v_1, v_2, u_1, u_2)^\top = (0, \ldots, 0, \nabla\phi(\theta)^\top)^\top = \chi_1(\alpha)$. Thus $\nabla h(\alpha) \perp \mathcal{W}_\phi^{\texttt{mom}}(\alpha)$: h is a conservation law of $\phi_{\texttt{Lin}}$.

Again, we generalize from (Marcotte et al., 2024) the notion of conservation laws for ϕ for a non-Euclidean GF, when M does not depend on the data set Z.

Definition 2.19 (Conservation law for (2))**.** A function $h : \Theta \mapsto \mathbb{R} \in \mathcal{C}^1(\Theta, \mathbb{R})$ is a conservation law of ϕ for (2) if $\nabla h(\theta) \perp \mathcal{W}_\phi^{\texttt{grad}}(\theta)$.

Conservation laws are known for the linear case $\phi_{\texttt{Lin}}$ (Arora et al., 2019; 2018) and the ReLu case $\phi_{\texttt{ReLu}}$ (Du et al., 2018).

Example 2.20 (Conservation laws for linear and ReLu neural networks in Euclidean GF scenario). If $\theta := (U_1, \cdots, U_q)$ satisfies (2) with $M_Z(\theta) = \mathrm{Id}_D$, then for each $i = 1, \cdots, q-1$ the function $\theta \mapsto U_i^\top U_i - U_{i+1}U_{i+1}^\top$ (resp. the function $\theta \mapsto \mathrm{diag}\left(U_i^\top U_i - U_{i+1}U_{i+1}^\top\right)$) defines a set of conservation laws for $\phi_{\texttt{Lin}}$ (resp. for $\phi_{\texttt{ReLu}}$).

3. Finding Conservation Laws

In this section, we propose a constructive way to build *some* conservation laws. Then we explain how we can certify if there are conservation laws that are missing or not.

3.1. Constructing Conservation Laws

Conservation laws can be built using formal calculus, via the orthogonal relation that defines conservation laws (Definition 2.19 and Definition 2.17), as done for Euclidean-GF in (Marcotte et al., 2024). One can also exploit invariances in the spirit of Noether's theorem (Noether, 1918; Wibisono et al., 2016; Tanaka & Kunin, 2021).

Using formal calculus. By Definition 2.17 (resp. Definition 2.19), h is a conservation law for (3) (resp. (2)) if

$$\langle \nabla h(\alpha), \chi_i(\alpha)\rangle = 0,\ \forall \alpha = (t, \theta, \dot\theta) \in \tilde\Theta,\ 0 \le i \le d, \tag{12}$$

where the χ_i are the $d+1$ vector fields defined in (8) that span the linear function space $W_\alpha^{\texttt{mom}}$ (resp. if

$$\langle \nabla h(\theta), M(\theta)\nabla\phi_i(\theta)\rangle = 0,\ \forall \theta \in \Theta,\ 1 \le i \le d. \tag{13}$$

One could seek conservation laws that belong to the finite-dimensional linear space of polynomials of a given degree, as in (Marcotte et al., 2024) for the Euclidean GF scenario. However, for MF scenarios, Theorem 2.1 suggests that the conserved functions include a non-polynomial term in t, specifically $\exp(\tau t)$ when $\tau(t) = \tau \neq 0$. Thus, in the MF case when $\tau(t) = \tau$, instead of considering conserved functions under the form $h(t, \theta, \dot\theta)$, we use a change a variable $s := \exp(\tau t)$ to consider "modified" conserved functions $\tilde h$ that are under the form $\tilde h(s, \theta, \dot\theta) := h((\ln s)/\tau, \theta, \dot\theta)$. Simple calculus yields $\nabla h(t, \theta, \dot\theta) = \mathrm{diag}(\tau s, 1, \dots 1)\nabla\tilde h(s, \theta, \dot\theta)$, therefore, given any constant c and any vector field $f : \Omega \times \mathbb{R}^D \to \mathbb{R}^D \times \mathbb{R}^D$, we have: for every $(s, \theta, \dot\theta) \in \mathbb{R}_+^* \times \Omega \times \mathbb{R}^D$: $\nabla h(t, \theta, \dot\theta) \perp \left(\begin{smallmatrix} c \\ f(\theta, \dot\theta)\end{smallmatrix}\right) \Leftrightarrow \nabla\tilde h(s, \theta, \dot\theta) \perp \left(\begin{smallmatrix} \tau s c \\ f(\theta, \dot\theta)\end{smallmatrix}\right)$. Recalling that we consider the case $\tau(t) = \tau$, all the vector fields $\chi_i(t, \theta, \dot\theta)$ from (8) involved in the definition of $W_\alpha^{\texttt{mom}}$ do not depend on time, and their first coordinates are constant. Consequently, defining $\tilde\chi_i(s, \theta, \dot\theta) := \mathrm{diag}(\tau s, 1, \dots 1)\chi_i(0, \theta, \dot\theta)$, we can solve[2] (12) with the new vector fields $\tilde\chi_i$ and determine all polynomial $\tilde h$ of a given degree such that $\nabla\tilde h(s, \theta, \dot\theta) \perp \tilde\chi_i(s, \theta, \dot\theta)$ for all i. Finally, since $s = \exp(\tau t)$, we can determine all conservation laws that are polynomial in $\theta, \dot\theta$ and in $\exp(\tau t)$.

In the case where $\tau(t)$ is *not* constant, it remains possible to try directly to solve (12) with the initial vector fields χ_i in $\mathcal{W}_\phi^{\texttt{mom}}$ and see if there are polynomial conservation laws h at a given degree. In particular, for a Nesterov flow with Euclidean metric, $\tau(t) = 3/t$, it turns out that there are polynomial conservation laws (see Section 4.1). Our code to compute them is available at https://github.com/sibyllema/Conservation_laws_ICML.

Using invariances – gradient flows. For gradient flows, invariances of the cost (1) directly lead to conservation laws.

Definition 3.1 (Invariant transformation on the cost (1))**.** A (one-parameter) transformation on an open set $\Omega \subseteq \Theta \subseteq \mathbb{R}^D$ is a map $T : \mathbb{R} \times \Omega \to \mathbb{R}^D$ such that $T(\cdot, \theta)$ is differentiable for each $\theta \in \Omega$ and $T(0, \cdot) = \texttt{id}$. This transformation leaves invariant the cost (1) if for all $\theta \in \Omega$ and for all $\epsilon \ge 0$, $\mathcal{E}_Z(T(\epsilon, \theta)) = \mathcal{E}_Z(T(0, \theta)) = \mathcal{E}_Z(\theta)$. When this holds, simple calculus yields for every $\theta \in \Omega$:

$$\left\langle \nabla\mathcal{E}_Z(\theta), \frac{\partial}{\partial\epsilon}T(\epsilon, \theta)\Big|_{\epsilon=0}\right\rangle = 0. \tag{14}$$

We denote $\Delta_T(\cdot) := \frac{\partial}{\partial\epsilon}T(\epsilon, \cdot)\Big|_{\epsilon=0}$.

Example 3.2 (Fondamental example of linear transformation). Let us consider $\theta := (U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. We define T^A the linear transformation where for all $\epsilon \ge 0$:

$$T^A(\epsilon, U, V) = \big(U\exp(\epsilon A), V\exp(-\epsilon A^\top)\big). \tag{15}$$

Simple calculus yields $\Delta_{T^A}(U, V) = (UA, -VA^\top)$. Considering $g(\theta, x) := UV^\top x$ and any $A \in \mathbb{R}^{r\times r}$, T^A is an invariant transformation on (1): for all ϵ and θ, $g(T^A(\epsilon, \theta), \cdot) = g(\theta, \cdot)$ hence $\mathcal{E}_Z(T^A(\epsilon, \theta)) = \mathcal{E}_Z(\theta)$. Considering now $g(\theta, x) := U\sigma(V^\top x)$ with σ the ReLU activation function, a similar reasoning shows that T^A is an invariant transformation on (1) if $A \in \mathbb{R}^{r\times r}$ is *diagonal*.

From loss invariance to conservation laws for GF. In the context of a gradient flow (2), the consequence (14) of invariance rewrites as

$$\langle M_Z(\theta(t))^{-1}\dot\theta(t), \Delta_T(\theta(t))\rangle = 0 \tag{16}$$

as soon as $M_Z(\theta)$ is invertible for every θ.

As a particular consequence, for gradient flows with linear networks in a Euclidean setting ($M \equiv \mathrm{Id}_D$), since (16) holds for $T = T^A$ with *any* matrix $A \in \mathbb{R}^{r\times r}$, denoting

[2] Concretely this is expressed as a linear system.

$\langle M, N\rangle := \mathtt{Tr}(M^\top N)$ we obtain $0 = \langle \dot{U}(t), U(t)A\rangle - \langle \dot{V}(t), V(t)A\rangle$ at each time and for any A. Specializing this to any *symmetric* matrix, we obtain that $0 = \langle \dot{U}, UA\rangle - \langle \dot{V}, VA\rangle = 1/2\frac{\mathrm{d}}{\mathrm{d}t}(\langle U, UA\rangle - \langle V, VA\rangle)$. Thus for every symmetric matrix A, $\langle U, UA\rangle - \langle V, VA\rangle$ is conserved, which coincides with all conservation laws in that case (Arora et al., 2019; 2018; Marcotte et al., 2024).

Similarly, for ReLU neural networks in a Euclidean setting (without bias for simplicity, this holds with bias too as detailed in the proof of Theorem 4.6), by restricting ourselves to *elementary diagonal matrices* $A = E_{i,i}$ where $E_{i,i}$ is the one-hot matrix in $\mathbb{R}^{r\times r}$ with the (i,i)-th entry being 1, $i = 1, \cdots, r$, we obtain that for all i, $0 = \langle \dot{U}_i(t), U_i(t)\rangle - \langle \dot{V}_i(t), V_i(t)\rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\langle U_i, U_i\rangle - \langle V_i, V_i\rangle)$. Thus for all i, $\|U_i\|^2 - \|V_i\|^2$ is conserved, recovering all conservation laws (Du et al., 2018; Marcotte et al., 2024).

Using invariances – momentum flows. Invariances of the cost (1) are replaced in Noether's theorem by invariances of a *Lagrangian* compatible with the flow (3).

Definition 3.3 (Lagrangian)**.** A Lagrangian (Noether, 1918; Wibisono et al., 2016) is a function of three variables $\mathcal{L}(t, \theta, v)$. The second order dynamical system (3) is compatible with this Lagrangian if for every solution $\theta(t)$ of (3) the Euler-Lagrange equation of $\mathcal{L}$ holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_v \mathcal{L}\big(t, \theta(t), \dot{\theta}(t)\big)\Big) = \partial_\theta \mathcal{L}\big(t, \theta(t), \dot{\theta}(t)\big).$$

Example 3.4 (Euclidean momentum (Wibisono et al., 2016)). In the case $M = \mathrm{Id}_D$ and $\tau(t) = \tau$ the map

$$\mathcal{L}(t, \theta, v) := \exp(\tau t)\big(\tfrac{\|v\|^2}{2} - \mathcal{E}_Z(\theta)\big) \tag{17}$$

is a *Lagrangian* of the second-order dynamical system (3). In the Euclidean case ($M = \mathrm{Id}_d$) with a Nesterov flow ($\tau(t) = 3/t$),

$$\mathcal{L}(t, \theta, v) := \tfrac{t^3}{2}\big(\tfrac{\|v\|^2}{2} - \mathcal{E}_Z(\theta)\big) \tag{18}$$

is a *Lagrangian* of the second-order dynamical system (3).

More generally, (Wibisono et al., 2016) gives a Lagrangian associated to (3), subject to: a) an hypothesis of an "ideal scaling condition" (Wibisono et al., 2016, Equation 2.2b)); and b) the assumption that the matrix M is the inverse of the Hessian of an explicitly known metric ψ.

In general, a transformation leaving the cost invariant *does not necessarily* leave invariant the Lagrangian of the associated dynamic (e.g., revisiting Example 3.2, T^A leaves the cost invariant for *any* $A \in \mathbb{R}^{r\times r}$, yet only *skew* matrices A leave the Lagrangian (17) invariant). However, to obtain a conserved function with Noether theorem[3], one needs to consider transformations leaving invariant the Lagrangian: in that case the function $\partial_v \mathcal{L}(t, \theta(t), \dot{\theta}(t)) \cdot \Delta_T(\theta(t))$ is conserved.

Theorem 3.5 (Noether theorem)**.** *Let $(\theta(t), \dot{\theta}(t))$ be a solution of (3). Then for each transformation T that leaves $\mathcal{L}$ invariant, $\frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_v \mathcal{L}(t, \theta(t), \dot{\theta}(t)) \cdot \Delta_T(\theta(t))\Big) = 0$.*

3.2. Finding the number of conservation laws

While being able to build conservation laws is beneficial, the question remains: how to ensure that we have derived "all" possible laws? This first requires to recall the definition of independent conserved functions (Marcotte et al., 2024, Definition 2.18), to avoid functional redundancies.

Definition 3.6. A family of N functions $(h_1, \cdots, h_N)$ in $\mathcal{C}^1(\tilde{\Omega}, \mathbb{R})$ is *independent* if for all $\alpha \in \tilde{\Omega}$ the vectors $(\nabla h_1(\alpha), \cdots, \nabla h_N(\alpha))$ are linearly independent.

(Marcotte et al., 2024, Theorem 3.3) link the number of independent conservation laws to the dimension of a space involving Lie algebras. Knowledge about Lie algebras is not mandatory in the main body of this paper, basics are recalled in Appendix H.1 (See (Marcotte et al., 2024)) to support the proofs. The space $\mathrm{Lie}(\mathcal{W})$ is *the generated Lie algebra of* $\mathcal{W}$ (See Appendix H.1), it is entirely characterized by $\mathcal{W}$.

Theorem 3.7. *If $\dim(\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha))$ is locally constant then each $\alpha \in \tilde{\Omega} \subseteq \mathbb{R}^{2D+1}$ admits a neighborhood $\tilde{\Omega}'$ such that there are $2D + 1 - \dim(\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha))$ independent conservation laws of ϕ for (3) on $\tilde{\Omega}'$.*

The same theorem holds for $\mathcal{W}_\phi^{\mathtt{grad}}$ by replacing $2D + 1$ by D. Besides, given a finite set of vector fields, Marcotte et al. (2024) provide a code (detailed in Section 3.3) that computes the dimension of the trace of their generated Lie algebra: in their case, they apply it to the vector fields $(\nabla\phi_i(\cdot))_i$ that span $\mathcal{W}_\phi$ (as defined in (7)). We can directly use this code with the finite set of vector fields that span $\mathcal{W}_\phi^{\mathtt{mom}}$ as defined in (8) (resp. on the fields $(M(\cdot)\nabla\phi_i(\cdot))_i$ that span $\mathcal{W}_\phi^{\mathtt{grad}}$ for the non-Euclidean GF case, see (9)).

By computing the number of independent conservation laws with this code in the Euclidean GF scenario, Marcotte et al. (2024) (in their Section 3.3) establish that the set of known conservation laws for linear and ReLu neural networks (recalled in Example 2.20) is complete. In particular, they fully work out theoretically the 2-layer case and show (Marcotte et al., 2024, Proposition 4.2, Corollary 4.4):

Proposition 3.8. *If $(U; V)$ has full rank noted $\mathtt{rk}$, then in a neighborhood of (U, V), the set of $\mathtt{rk}/2 \cdot (2r + 1 - \mathtt{rk})$ independent conservation laws given by Example 2.20 is complete: there exists no other conservation law.*

[3]The full version of Noether theorem with a non-zero right-hand side is recalled in (Tanaka & Kunin, 2021).

4. Conservation for Neural Networks

We now exemplify our results in different settings. We study the conservation laws for neural networks with q layers, and either a linear or ReLU activation. We write $\theta := (U_1, \cdots, U_q)$ with $U_i \in \mathbb{R}^{n_{i-1} \times n_i}$ the weight matrices. We also consider the impact of the choice of metric. The striking outcome of this analysis is that momentum flows radically change the structure of the conservation laws. There are fewer (or even none) conserved quantities when using momentum flows, and this phenomenon appears both for Euclidean and non-Euclidean geometries.

4.1. PCA: linear networks with Euclidean geometry

The following theorem (proved in Appendix E) gives a set of conserved functions when $g(\theta, x) := U_1 \dots U_q x$. Here $\mathcal{A}_n$ denotes the set of skew-symmetric matrices in $\mathbb{R}^{n \times n}$.

Theorem 4.1. *Consider the model $g(\theta, x) := U_1 \dots U_q x$. For all $i = 1, \cdots q-1$ and for all $A \in \mathcal{A}_{n_i}$, the function*

$$\exp\big(\textstyle\int^t \tau(s)\mathrm{d}s\big)\big(\langle \dot{U}_i, U_i A\rangle + \langle \dot{U}_{i+1}, A^\top U_{i+1}\rangle\big) \quad (19)$$

is a conservation law for (3) with $M = \mathrm{Id}_D$, where $\int^t \tau(s)\mathrm{d}s$ is a primitive of $\tau(\cdot)$. Moreover, for each i such that $n_{i-1} = n_{i+1} = 1$ and all $A \in \mathcal{A}_{n_i}$ the function $\exp\big(\int^t \tau(s)\mathrm{d}s\big)\big(\langle \dot{U}_i, U_{i+1}^\top A\rangle + \langle \dot{U}_{i+1}^\top, U_i A\rangle\big)$ is an additional conservation law.

In particular, for the heavy ball case $\tau(t) = \tau$, these functions exactly correspond to the ones obtained by solving (12) with the change of variable explained in Section 3.1. For the Euclidean Nesterov case $\tau(t) = 3/t$, these functions are also directly obtained by solving (12), with a polynomial term t^3 in t. As discussed in Appendix F, in the case $(n_{i-1}, n_{i+1}) \neq (1,1)$, the associated conserved functions can also be found using Noether theorem with T^A (defined in Example 3.2) for every skew matrix A, and for the case $(n_{i-1}, n_{i+1}) = (1,1)$, the new associated conserved functions can also be found using Noether theorem with a new linear transformation T'^A (defined in (29)).

The above theorem gives a set of conserved functions. A priori, they are not all independent. The following proposition (proved in Appendix G) gives the number of independent conserved functions given by Theorem 4.1 in the case $q = 2$. We rewrite $\theta = (U, V)$ as a vertical matrix concatenation denoted $(U; V) \in \mathbb{R}^{(n+m)\times r}$.

Proposition 4.2. *Consider $\widetilde{\Theta}$ the set of $\alpha = (t, \theta, \dot{\theta})$ such that $(\theta, \dot{\theta}) = (U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ has full rank, and assume $(n, m) \neq (1,1)$. Then if $2(n+m) \leq r$, Theorem 4.1 gives exactly $(n+m)[(r - 2(n+m)) + r - 1]$ independent conserved functions. If $2(n+m) > r$, Theorem 4.1 gives exactly $r(r-1)/2$ independent conserved functions.*

Now we want to establish if there are other conservation laws independent from the ones obtained in Theorem 4.1. By using Theorem 3.7, we only need to determine the dimension of the trace of a Lie algebra characterized by $\mathcal{W}^{\mathtt{mom}}_{\phi_{\mathtt{Lin}}}$. We fully work out theoretically the case $q = 2$ as detailed in the following theorem and show that there no other conservation laws. See Appendix H for a proof.

Theorem 4.3. *Consider $g(\theta, x) := UV^\top x$ with $(U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m \times r}$. If $(U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ has full rank noted $\mathtt{rk}$ then, in a neighborhood of $(t, U, V, \dot{U}, \dot{V})$:*

- *If $(n, m) \neq (1,1)$, there are exactly $\mathtt{rk}/2 \cdot (2r - 1 - \mathtt{rk})$ independent conservation laws of $\phi_{\mathtt{Lin}}$.*
- *If $n = m = 1$ and if $r \geq 4$, there are exactly $4r - 6$ independent conservation laws of $\phi_{\mathtt{Lin}}$.*

For deeper cases, we computed $\dim[\mathrm{Lie}(\mathcal{W}^{\mathtt{mom}}_{\phi_{\mathtt{Lin}}})(\alpha)]$ using the method explained in Section 3.2 with the vector fields that generate $W^{\mathtt{mom}}_{\phi_{\mathtt{Lin}}}$ on a sample of depths/widths of small size. This empirically confirmed (see Appendix M.1) that there are no other conservation laws for deeper cases too.

Comparing the number of independent conservation laws in the GF scenario given in Proposition 3.8 with the one in the MF scenario given in the last Theorem 4.3, we obtain as highlighted next a *loss of conservation* when transitioning from Euclidean gradient flows to the Euclidean momentum setting and when $r \leq n + m$. Notice that this case includes the factorization by matrices up to full rank and even in mildly over-parameterized regimes. By contrast, in the more over-parameterized regime $r > n + m$, we obtain *a gain in conservation*. A proof can be found in Appendix I.

Proposition 4.4. *Consider $g(\theta, x) := UV^\top x$ with $(U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. Assume both $(U; V) \in \mathbb{R}^{(n+m)\times r}$ and $(U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ have full rank and $(n, m) \neq (1,1)$. Denote N_{GF} (resp. N_{MF}) the number of independent conservation laws of $\phi_{\mathtt{Lin}}$ for (2) (resp. for (3)) in a neighborhood of (U, V) (resp. of $(t, U, V, \dot{U}, \dot{V})$, then if $r \leq n+m$: $N_{GF} - N_{MF} = r > 0$, else: $N_{GF} - N_{MF} \leq 0$.*

4.2. MLP: ReLU Networks with Euclidean geometry

For the case $q = 2$ without bias, since $\phi_{\mathtt{ReLU}}(\theta) = (\phi_j(\theta))_{j=1}^r$ is decoupled into r functions $\phi_j(\theta) := u_j v_j^\top \in \mathbb{R}^{n\times m}$ each depending on a separate block of coordinates, Jacobian matrices and Hessian matrices are block-diagonal. Therefore Lie brackets computations can be done separately for each block, using Theorem 4.3 with $r = 1$. We obtain that there is no conserved function for Euclidean momentum flow with a two-layer ReLU network without bias. For deeper ReLU networks including with bias, by computing on a sample of depths/widths of small size the number of conservation laws as explained in Section 3.2 using the finite set of vector fields that generates $\mathcal{W}^{\mathtt{mom}}_{\phi_{\mathtt{ReLU}}}$, we obtain that there is no conservation law (see details in Appendix M.1).

4.3. NMF: Linear Networks with Mirror geometry

Non-negative matrix factorization (NMF) is an example of a two-layer linear neural network, and we use the mirror geometry associated to the Shannon entropy potential both on U and V to enforce non-negativity. The following theorem (proved in Appendix J) gives a set of conserved functions in the gradient flow case (2). We denote $1_\ell \in \mathbb{R}^\ell$ the vector with all coordinates equal to 1.

Theorem 4.5. *Consider $g(\theta, x) := UV^\top x$ with $\theta = (U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. Then*

$$1_n^\top U - 1_m^\top V \tag{20}$$

defines r independent conservation laws for (2) with $M(\theta) := \mathrm{diag}(\theta)$.

These functions can be found by solving (13) with formal calculus with the vector fields that generate $\mathcal{W}^{\mathtt{grad}}_{\phi_{\mathtt{Lin}}}$ in this non-Euclidean gradient flow case. As for the Euclidean GF case, these functions can also be linked to invariance on the cost as they coincide with the time integration of (16): $\langle \mathrm{diag}(\theta)^{-1}\dot{\theta}, \Delta_{T^A}(\theta)\rangle = 0$ when $A = E_{i,i}$ for each $i = 1, \cdots, r$. By computing the number of independent conservation laws with the method explained in Section 3.2, we obtain that there are no other conservation laws (see Appendix M.2). In contrast, by computing the number of independent conservation laws for the vector fields that generate $\mathcal{W}^{\mathtt{mom}}_{\phi}$ for the *non-Euclidean* momentum flow case, we found that there is no conservation law at all in that case. Once again there are fewer conservation laws for MF than GF.

4.4. ICNN: mixing Mirror/Euclidean geometries

In this section, we only treat the two-layer ReLU case of ICNNs (Amos et al., 2017), corresponding to $g(\theta, x) = U\sigma(V^\top x + b^\top)$, with $\theta = (U; V) \in \mathbb{R}^{(m+n)\times r}$ and $\sigma(t) := \max(t, 0)$. We employ the mirror geometry associated to the Shannon entropy potential on U to enforce non-negativity ($U \geq 0$) and the Euclidean metric on V and b. The following theorem (proved in Appendix K) gives a set of conservation laws for GF.

Theorem 4.6. *Consider $g(\theta, x) = U\sigma(V^\top x + b^\top)$. Denote U_j (resp. V_j / b_j) the j-th column of U (resp column of V / entry of b). For all $j = 1, \cdots, r$, the function*

$$1_n^\top U_j - \frac{1}{2}\left(\|V_j\|^2 + b_j^2\right) \tag{21}$$

is a conservation law for (2) with $M(U, V, b) := \mathrm{diag}[(U, 1_{m\times r}, 1_{1\times r})]$.

By computing the number of independent conservation laws with the method explained in Section 3.2, we obtain that there are no other conservation laws. In contrast, by computing the number of independent conservation laws for the vector fields that generate $\mathcal{W}^{\mathtt{mom}}_{\phi_{\mathtt{ReLU}}}$ for the non-Euclidean momentum flow case, we found that there is no conservation law at all in that case (see Appendix M.3). Thus, once again we observe a loss of conservation from GF to MF regimes.

4.5. Natural gradient

Our main result on conservation laws for momentum, Proposition 2.9, only holds when $M = M_Z$ is *independent* of the dataset Z. While we cannot apply it to the natural gradient setting, we can still conduct (a part of) our study for the non-Euclidean GF. We obtain that all conservation laws known for the Euclidean gradient flow case recalled in Example 2.20 are conservation laws for the Natural gradient flow. See Appendix L for a proof.

Conclusion

In this paper, we examined conservation laws for gradient or momentum flows within Euclidean as well as non-Euclidean geometries. One notable constraint of this theory is its limitation to continuous-time flow. Appendix N studies numerically the impact of the time discretization on MLP and NMF problems. It also studies the impact of momentum on quantities preserved by GF. Understanding these approximate conservations is an important open problem. We also anticipate that our theory will be adaptable to the study of the SGD case.

Acknowledgements

The work of G. Peyré was supported by the European Research Council (ERC project NORIA) and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). The work of R. Gribonval was partially supported by the AllegroAssai ANR project ANR-19-CHIA-0009 and the SHARP ANR Project ANR-23-PEIA-0008 of the PEPR IA, funded in the framework of the France 2030 program.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*,

pp. 146–155. PMLR, 2017.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

Bah, B., Rauhut, H., Terstiege, U., and Westdickenberg, M. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1): 307–353, 2022.

Bonnard, B., Chyba, M., and Rouot, J. *Geometric and Numerical Optimal Control - Application to Swimming at Low Reynolds Number and Magnetic Resonance Imaging*. SpringerBriefs in Mathematics. Springer Int. Publishing, 2018.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3): 200–217, 1967.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conf. on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Du, S. S., Hu, W., and Lee, J. D. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31, 2018.

Głuch, G. and Urbanke, R. Noether: The more things change, the more stay the same. *arXiv preprint arXiv:2104.05508*, 2021.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Kunin, D., Sagastuy-Brena, J., Ganguli, S., Yamins, D. L., and Tanaka, H. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, 1999.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Marcotte, S., Gribonval, R., and Peyré, G. Abide by the law and follow the flow: Conservation laws for gradient flows. *Advances in Neural Information Processing Systems*, 36, 2024.

Martens, J. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 735–742, 2010.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Martens, J. and Sutskever, I. *Training Deep and Recurrent Networks with Hessian-Free Optimization*. Springer, 2012.

Min, H., Tarmoun, S., Vidal, R., and Mallada, E. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pp. 7760–7768. PMLR, 2021.

Nemirovskij, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Nesterov, Y. E. A method of solving a convex programming problem with convergence rate $o(k^2)$. In *Doklady Akademii Nauk*, volume 269, pp. 543–547. Russian Academy of Sciences, 1983.

Noether, E. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918:235–257, 1918.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Su, W., Boyd, S., and Candes, E. J. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Tanaka, H. and Kunin, D. Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Tarmoun, S., Franca, G., Haeffele, B. D., and Vidal, R. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pp. 10153–10161. PMLR, 2021.

The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 9.7)*, 2022. `https://www.sagemath.org`.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

Zhao, B., Ganev, I., Walters, R., Yu, R., and Dehmamy, N. Symmetries, flat minima, and the conserved quantities of gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023.

A. Proof of Theorem 2.1

Theorem 2.1 (Structure theorem)**.** *Let $h(t,\theta)$ (resp. $h(t,\theta,\dot\theta)$) be a conserved function for the ODE (2) (resp. (3) with $\tau(t)=\tau$) when its right-hand side is zero. For all t and θ, one has $h(t,\theta)=h(0,\theta)$ (resp. $h(t,\theta,\dot\theta)=H(\theta+\frac{\dot\theta}{\tau},\dot\theta\exp(\tau t))$, where $H(a,b)\coloneqq h(0,a-\frac{b}{\tau},b)$). In particular, conserved functions in the MF scenario can be expressed with $2D$ variables (instead of $2D+1$).*

Proof. **GF case:** Let $h(t,\theta)$ be a conserved function for the ODE (2) with a right-hand side equal to zero with any initial condition θ_0. In that case, (2) rewrites $\dot\theta=0$. For each initialization θ_0, the solution of this ODE is $\theta(t)=\theta_0$ for $t\in\mathbb{R}$ and by definition of a conserved function, one has $h(t,\theta_0)=h(t,\theta(t))=h(0,\theta(0))=h(0,\theta_0)$.

MF case: Let $h(t,\theta,\dot\theta)$ be a conserved function for the ODE (3) with a right-hand side equal to zero, an arbitrary initial condition $(t_0,\theta_0,\dot\theta_0)$, and $\tau(t)\coloneqq\tau$. In that case, (3) rewrites $\ddot\theta+\tau\dot\theta=0$. For each initialization $(\theta_0,\dot\theta_0)$, the solution of this ODE is: $\theta(t)=\theta_0+\dot\theta_0/\tau\,(1-\exp(-t\tau))$, for every $t\in\mathbb{R}$, and it satisfies $(\theta(0),\dot\theta(0))=(\theta_0,\dot\theta_0)$.

Since h is conserved, $h(0,\theta_0,\dot\theta_0)=h(t,\theta(t),\dot\theta(t))=h\left(t,\theta_0+\dot\theta_0/\tau\,(1-\exp(-t\tau))\,,\dot\theta_0\exp(-t\tau)\right)$ for each $t\in\mathbb{R}$. This holds for any $\theta_0,\dot\theta_0$, and given any $\theta,\dot\theta$ one can find $\theta_0,\dot\theta_0$ such that $(\theta,\dot\theta)=(\theta_0+\dot\theta_0/\tau\,(1-\exp(-t\tau))\,,\dot\theta_0\exp(-t\tau))$. Expliciting the expression of such $\theta_0,\dot\theta_0$ in terms of $\theta,\dot\theta$ yields

$$h(t,\theta,\dot\theta)=h(0,\theta-\dot\theta/\tau(\exp(t\tau)-1),\dot\theta\exp(\tau t))=H\left(\theta+\frac{\dot\theta}{\tau},\dot\theta\exp(\tau t)\right).\quad\square$$

B. Proof of Proposition 2.7

Proposition 2.7 (Smooth functions conserved through a given flow)**.** *Given $F\in\mathcal{C}^1(\tilde\Omega,\mathbb{R}^{2D})$, a function $h\in\mathcal{C}^1(\tilde\Omega,\mathbb{R})$ is conserved through the flow induced by F if and only if $\langle\nabla h(\alpha),(1,F(\alpha)^\top)^\top\rangle=0$ for all $\alpha\in\tilde\Omega$.*

Proof. We will use that $\langle\nabla h(\alpha),(1,F(\alpha)^\top)^\top\rangle=\partial h(\alpha)(1,F(\alpha)^\top)^\top$ with ∂h the Jacobian of h.

Assume that $\partial h(\alpha)(1,F(\alpha)^\top)^\top=0$ for all $\alpha\in\tilde\Omega$. Then for all $\texttt{init}\coloneqq(t_{\text{init}},\omega_{\text{init}})\in\tilde\Omega$ and for all $t\in(t_{\text{init}}-\eta_{\text{init}},t_{\text{init}}+\eta_{\text{init}})$, denoting $\dot\omega(t,\texttt{init})\coloneqq\frac{\mathrm{d}}{\mathrm{d}t}\omega(t,\texttt{init})$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}h(t,\omega(t,\texttt{init}))=\partial h(t,\omega(t,\texttt{init}))(1,\dot\omega(t,\texttt{init})^\top)^\top=\partial h(t,\omega(t,\texttt{init}))(1,F(t,\omega(t,\texttt{init}))^\top)^\top=0.$$

Thus: $h(t,\omega(t,\texttt{init}))=h(\texttt{init})$, *i.e.*, h is conserved through χ.

Conversely, assume that there exists $(t_0,\omega_0)\in\tilde\Omega$ such that $\partial h(t_0,\omega_0)(1,F(t_0,\omega_0)^\top)^\top\neq0$. Then by continuity of $z\in\tilde\Omega\mapsto\partial h(z)(1,F(z)^\top)^\top$, there exists $r>0$ such that $\partial h(z)(1,F(z)^\top)^\top\neq0$ on $B((t_0,\omega_0),r)$. With $\texttt{init}=(t_0,\omega_0)$ by continuity of $t\mapsto\omega(t,\texttt{init})$, there exists $\epsilon>0$, such that for all $t\in(t_0,t_0+\epsilon)$, $\omega(t,\texttt{init})\in B((t_0,\omega_0),r)$. Then for all $t\in(t_0,t_0+\epsilon)$: $\frac{\mathrm{d}}{\mathrm{d}t}h(t,\omega(t,\texttt{init}))=\partial h(t,\omega(t,\texttt{init}))(1,F(t,\omega(t,\texttt{init}))^\top)^\top\neq0$, hence h is not conserved through the flow induced by F. $\square$

C. Proof of Proposition 2.8

Recall that F_Z is defined in (4).

Proposition 2.8. *A function $h:\tilde\Theta\mapsto\mathbb{R}\in\mathcal{C}^1(\tilde\Theta,\mathbb{R})$ is locally conserved on Θ for any data set if and only if $\nabla h(\alpha)\perp W_\alpha^{\texttt{mom}}$ for all $\alpha\in\tilde\Theta$, where for all $\alpha\coloneqq(t,\theta,\dot\theta)\in\tilde\Theta$:*

$$W_\alpha^{\texttt{mom}}\coloneqq\operatorname*{span}_{Z\in\mathcal{Z}_\theta}\left\{(1,F_Z(\alpha)^\top)^\top\right\},$$

with $\mathcal{Z}_\theta$ the set of all data sets $Z=(x_i,y_i)_i$ such that there exists a neighborhood Ω of θ such that for all i, $g(\cdot,x_i)\in\mathcal{C}^2(\Omega,\mathbb{R}^n)$ and such that $M_Z(\cdot)\in\mathcal{C}^1(\tilde\Omega,\mathbb{R}^{D\times D})$.

Proof. Let us consider $h:\tilde\Theta\mapsto\mathbb{R}\in\mathcal{C}^1(\tilde\Theta,\mathbb{R})$. We first show the direct implication. We assume that h is locally conserved on Θ for any data set. Let $\alpha\in\tilde\Theta$ and let $Z=(x_i,y_i)_i\in\mathcal{Z}_\theta$. By definition of $\mathcal{Z}_\theta$, there exists a neighborhood of θ

$\Omega \subseteq \Theta$ such that for all i, $g(\cdot, x_i) \in \mathcal{C}^2(\Omega, \mathbb{R}^n)$ and such that $M_Z(\cdot) \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^{D\times D})$. Then by definition of being locally conserved on Θ for any data set, h is in particular conserved on Ω for any data set. Thus in particular h is conserved on Ω during the flow F_Z. By Proposition 2.7, $\nabla h(\alpha) \perp (1, F_Z(\alpha)^\top)^\top$. This holds for any $Z \in \mathcal{Z}_\theta$, thus $\nabla h(\alpha) \perp W_\alpha^{\mathtt{mom}}$. As it is true for any $\theta \in \tilde{\Theta}$, we have the direct implication.

We now show the converse implication. We assume that $\nabla h(\alpha) \perp W_\alpha^{\mathtt{mom}}$ for all $\alpha \in \tilde{\Theta}$. Let us consider an open subset $\Omega \subseteq \Theta$, and let us consider $Z = (x_i, y_i)_i$ a data set such that $g(\cdot, x_i) \in \mathcal{C}^2(\Omega, \mathbb{R})$ for each i and $(t, \theta, \dot{\theta}) \mapsto M_Z(t, \theta, \dot{\theta}) \in \mathcal{C}^1(\tilde{\Omega}, \mathbb{R}^{D\times D})$. In particular, $Z \in \mathcal{Z}_\theta$. Thus, for any $\alpha \in \tilde{\Omega}$, as $\nabla h(\alpha) \perp W_\alpha^{\mathtt{mom}}$, one has $\nabla h(\alpha) \perp (1, F_Z(\alpha)^\top)^\top$, and by Proposition 2.7, h is conserved on Ω during the flow F_Z. As this holds for any $\Omega \subseteq \Theta$, h is locally conserved on Θ for any data set. □

D. Proof of Proposition 2.9

Proposition 2.9. *Assume that $M_Z(t, \theta, \dot{\theta}) = M(t, \theta, \dot{\theta})$ does not depend on the data set Z and that $M(\cdot) \in \mathcal{C}^1(\tilde{\Theta}, \mathbb{R}^{D\times D})$. Assume that for each $y \in \mathcal{Y}$ the loss $\ell(z, y)$ is $\mathcal{C}^2$-differentiable with respect to $z \in \mathbb{R}^n$ and that for each $\theta \in \Theta$, there exists a training dataset $Z = (x_i, y_i)_i$ such that $\nabla \mathcal{E}_Z(\theta) = 0$ and such that for all i, $\theta \mapsto g(\theta, x_i)$ is $\mathcal{C}^2$ in a neighborhood of θ. Then, for each $\alpha \coloneqq (t, \theta, \dot{\theta}) \in \tilde{\Theta}$:*

$$W_\alpha^{\mathtt{mom}} = \mathbb{R}\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ M(t,\theta,\dot{\theta})W_\theta \end{pmatrix}. \tag{5}$$

Proof. Let $z \coloneqq (t, \theta, \dot{\theta}) \in \tilde{\Theta}$. Let us denote $\mathcal{Z}'_\theta$ the collection of all data set $Z = (x_i, y_i)_i$ such that for all i, $g(\cdot, x_i)$ is $\mathcal{C}^2$-differentiable in the neighborhood of θ. Let $z \coloneqq (t, \theta, \dot{\theta}) \in \tilde{\Theta}$. First, let us recall that (See Proposition 2.7 of (Marcotte et al., 2024)) we have:

$$W_\theta = \operatorname*{span}_{Z \in \mathcal{Z}'_\theta} \{\nabla \mathcal{E}_Z(\theta)\}. \tag{22}$$

Then as by assumption $M(\cdot) \in \mathcal{C}^1(\tilde{\Theta}, \mathbb{R}^{D\times D})$, we have that $\mathcal{Z}'_\theta = \mathcal{Z}_\theta$ with $\mathcal{Z}_\theta$ defined as in Proposition 2.8. Thus, we can rewrite:

$$W_\alpha^{\mathtt{mom}} = \operatorname*{span}_{Z \in \mathcal{Z}'_\theta} \left\{(1, F_Z(\alpha)^\top)^\top\right\} \overset{(4)}{=} \operatorname*{span}_{Z \in \mathcal{Z}'_\theta} \left\{ \begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})\nabla \mathcal{E}_Z(\theta) \end{pmatrix} \right\}.$$

Thus: $W_\alpha^{\mathtt{mom}} \subseteq \mathbb{R}\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \operatorname*{span}_{Z \in \mathcal{Z}'_\theta} \begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})\nabla \mathcal{E}_Z(\theta) \end{pmatrix} = \mathbb{R}\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})W_\theta \end{pmatrix}$, which gives the direct inclusion.

Let us show the converse inclusion. By assumption, there exists $Z_0 \in \mathcal{Z}'_\theta$ such that $\nabla \mathcal{E}_{Z_0}(\theta) = 0$, so that $W_\alpha^{\mathtt{mom}} \ni (1, F_{Z_0}(\alpha)^\top)^\top = \begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix}$, and thus $\mathbb{R}\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} \subseteq W_\alpha^{\mathtt{mom}}$. Then for all $Z \in \mathcal{Z}'_\theta$,

$$\begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})\nabla \mathcal{E}_Z(\theta) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})\nabla \mathcal{E}_Z(\theta) \end{pmatrix}}_{=(1,F_Z(\alpha)^\top)^\top \in W_\alpha^{\mathtt{mom}}} - \underbrace{\begin{pmatrix} 1 \\ \dot{\theta} \\ -\tau(t)\dot{\theta} \end{pmatrix}}_{\in W_\alpha^{\mathtt{mom}}} \in W_\alpha^{\mathtt{mom}},$$

and thus $\begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})W_\theta \end{pmatrix} = \operatorname*{span}_{Z \in \mathcal{Z}'_\theta} \begin{pmatrix} 0 \\ 0 \\ -M(t,\theta,\dot{\theta})\nabla \mathcal{E}_Z(\theta) \end{pmatrix} \subseteq W_\alpha^{\mathtt{mom}}$, which gives the converse inclusion. □

We now show that the following assumption holds for classical losses in machine learning.

Assumption D.1. For all $\theta \in \Theta$, there exists $Z \in \mathcal{Z}'_\theta$ such that $\nabla \mathcal{E}_Z(\theta) = 0$.

Lemma D.2. *Assumption D.1 holds for the mean-square error loss and the cross-entropy loss.*

Proof. Let $\theta \in \Theta$.

Mean-square error loss. The mean-squared error loss is defined by $(z, y) \mapsto \ell_2(z, y) \coloneqq \|y - z\|^2$. Let consider x such that $g(\cdot, x)$ is $\mathcal{C}^2$ on a neighborhood of θ. Then consider $y \coloneqq g(\theta, x)$. By definition $Z = (x, y) \in \mathcal{Z}'_\theta$. Moreover one has in a neighborhood of θ: $\nabla \mathcal{E}_Z(\theta') = \partial g(\theta', x)^\top \nabla_z \ell_2(g(\theta', x), y)) = 2\partial g(\theta', x)^\top (g(\theta', x) - y))$. Thus $\nabla \mathcal{E}_Z(\theta) = 0$.

Cross-entropy loss. The cross-entropy loss is defined by $(z \in \mathbb{R}^n, y \in \Sigma_n) \mapsto \ell_{\texttt{cross}}(z, y) := \mathrm{KL}(\mathrm{softmax}(z), y)$, where $\Sigma_n := \{(y_1, \cdots, y_n) \in \mathbb{R}^n_+ : \sum_{i=1}^n y_i = 1\}$ is the simplex and KL is the Kullback–Leibler divergence defined on $\Sigma_n \times \Sigma_n$ by $\mathrm{KL}(r, q) := \sum_{i=1}^n r_i \log(r_i/q_i)$. In particular by taking x such that $g(\cdot, x)$ is $\mathcal{C}^2$ on a neighborhood of θ and by taking $y = \mathrm{softmax}(g(\theta, x))$ and $Z = (x, y)$ (in $\mathcal{Z}'_\theta$), we obtain $\nabla \mathcal{E}_Z(\theta) = 0$. □

E. Proof of Theorem 4.1

We consider linear networks with $\theta := (U_1, \dots, U_q)$, $g(\theta, x) := U_1 \dots U_q x$.

Theorem 4.1. *Consider the model $g(\theta, x) := U_1 \dots U_q x$. For all $i = 1, \cdots q-1$ and for all $A \in \mathcal{A}_{n_i}$, the function*

$$\exp\big(\textstyle\int^t \tau(s)\mathrm{d}s\big)\big(\langle \dot{U}_i, U_i A\rangle + \langle \dot{U}_{i+1}, A^\top U_{i+1}\rangle\big) \tag{19}$$

is a conservation law for (3) with $M = \mathrm{Id}_D$, where $\int^t \tau(s)\mathrm{d}s$ is a primitive of $\tau(\cdot)$. Moreover, for each i such that $n_{i-1} = n_{i+1} = 1$ and all $A \in \mathcal{A}_{n_i}$ the function $\exp\big(\int^t \tau(s)\mathrm{d}s\big)\big(\langle \dot{U}_i, U_{i+1}^\top A\rangle + \langle \dot{U}_{i+1}^\top, U_i A\rangle\big)$ is an additional conservation law.

Proof. We first treat the general case. The specific case $(n_{i-1}, n_{i+1}) = (1, 1)$ will come next. Let us denote for $A \in \mathcal{A}_{n_i}$:

$$h_A(t, \theta, \dot{\theta}) := \exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\Big(\Big\langle \dot{U}_i, U_i A\Big\rangle + \Big\langle \dot{U}_{i+1}, A^\top U_{i+1}\Big\rangle\Big). \tag{23}$$

Observe that (3) implies

$$\ddot{U}_i(t) + \tau(t)\dot{U}_i(t) = -\nabla_{U_i}\mathcal{E}(\theta(t)) \tag{24}$$

for every $1 \le i \le q$, where to ease further computation each ∇_{U_i} is reshaped as the $n_{i-1} \times n_i$ matrix U_i. Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} h_A(t, \theta(t), \dot{\theta}(t)) &= \exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\Bigg(\Big\langle \ddot{U}_i, U_i A\Big\rangle + \underbrace{\Big\langle \dot{U}_i, \dot{U}_i A\Big\rangle}_{=0 \text{ as } A \text{ is a skew matrix}} + \Big\langle \ddot{U}_{i+1}, A^\top U_{i+1}\Big\rangle + \underbrace{\Big\langle \dot{U}_{i+1}, A^\top \dot{U}_{i+1}\Big\rangle}_{=0 \text{ as } A \text{ is a skew matrix}}\Bigg) \\
&\quad + \tau(t) h_A(t, \theta(t), \dot{\theta}(t)) \\
&\overset{(24)}{=} -\exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\Bigg(\langle \nabla_{U_i}\mathcal{E}_Z(\theta), U_i A\rangle + \langle \nabla_{U_{i+1}}\mathcal{E}_Z(\theta), \underbrace{A^\top}_{=-A} U_{i+1}\rangle\Bigg) \\
&= 0 \text{ as } T^A \text{ leaves } \mathcal{E}_Z \text{invariant (See (14) and Example 3.2)},
\end{aligned}$$

where here $T^A(\epsilon, U_i, U_{i+1}) = \big(U_i \exp(\epsilon A), \exp(-\epsilon A) U_{i+1}\big)$, and $\Delta_{T^A}(U_i, U_{i+1}) = (U_i A, -A U_{i+1})$.

Special case where $(n_{i-1} = n_{i+1} = 1)$. For any $A \in \mathcal{A}_{n_i}$, we denote:

$$g_A(t, \theta, \dot{\theta}) := \exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\Big(\Big\langle \dot{U}_i, U_{i+1}^\top A\Big\rangle + \Big\langle \dot{U}_{i+1}^\top, U_i A\Big\rangle\Big)$$

By our assumption on the dimensions, $U_{i+1}^\top A$ is an $n_{i+1} \times n_i = 1 \times n_i$ matrix, just as $\dot{U}_i$, and similarly for $U_i A$ and $\dot{U}_{i+1}^\top$, so g_A is indeed well-defined. We will prove below that (again with gradients properly reshaped as matrices)

$$\nabla_{U_i}\mathcal{E}_Z(\theta) \propto U_{i+1}^\top \text{ and } \nabla_{U_{i+1}}\mathcal{E}_Z(\theta) \propto U_i^\top. \tag{25}$$

As a result, using again (24) we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} g_A(t,\theta(t),\dot\theta(t)) &= \exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\left(\left\langle \ddot U_i, U_{i+1}^\top A\right\rangle + \left\langle \ddot U_{i+1}^\top, U_i A\right\rangle + \underbrace{\left\langle \dot U_i, \dot U_{i+1}^\top A\right\rangle + \left\langle \dot U_{i+1}^\top, \dot U_i A\right\rangle}_{=\langle \dot U_{i+1}\dot U_i + \dot U_i^\top \dot U_{i+1}^\top, A\rangle = 0 \text{ as } A \text{ is a skew matrix}}\right)\\ &\quad + \tau(t) g_A(t,\theta(t),\dot\theta(t))\\ &= -\exp\Big(\int^t \tau(s)\mathrm{d}s\Big)\left(\underbrace{\langle \underbrace{\nabla_{U_i}\mathcal{E}_Z(\theta)}_{\propto U_{i+1}^\top \text{ by (25)}}, U_{i+1}^\top A\rangle}_{=0 \text{ as } A \text{ is a skew matrix}} + \underbrace{\langle \underbrace{[\nabla_{U_{i+1}}\mathcal{E}_Z(\theta)]^\top}_{\propto U_i \text{ by (25)}}, U_i A\rangle}_{=0 \text{ as } A \text{ is a skew matrix}}\right)\\ &= 0.\end{aligned}$$

Let us now prove (25) as claimed. First, to simplify the notations, let us define:

$$\widehat{V_j} = \begin{cases} U_1 ... U_{j-1} \text{ for all } j = 2, ..., q, \\ \mathrm{Id}_{n_0} \text{ for } j = 1. \end{cases}$$

and

$$\widehat{W_j} = \begin{cases} U_{j+1} ... U_q \text{ for all } j = 1, ..., q-1, \\ \mathrm{Id}_{n_q} \text{ for j=q.} \end{cases}$$

Let us now derive an expression $\nabla_{U_i}\mathcal{E}_Z$. Given that $g(\theta, x) = U_1 \dots U_q x$,we can factorize $\mathcal{E}_Z(\theta) = F \circ \phi(\theta)$ with $\phi(\theta) := U_1 \dots U_q \in \mathbb{R}^{n_0 \times n_q}$ (identified with $\mathbb{R}^d$, $d = n_0 n_q$), and $F : \mathbb{R}^d \to \mathbb{R}$. The chain rule for Jacobians yields

$$\partial \mathcal{E}_Z(\theta) = \partial F(\phi(\theta)) \cdot \partial \phi(\theta) \tag{26}$$

hence for all $(H_1, \cdots, H_q) \in \mathbb{R}^D$ (with components $H_i \in \mathbb{R}^{n_{i-1} \times n_i}$ seen as vectors in $\mathbb{R}^{n_{i-1} n_i}$):

$$\begin{aligned}\langle \nabla \mathcal{E}_Z(\theta), (H_1, \cdots, H_q)\rangle &= \partial \mathcal{E}_Z(\theta) \cdot (H_1, \cdots, H_q) \overset{(26)}{=} \langle \nabla F(\phi(\theta)), \partial\phi(\theta) \cdot (H_1, \cdots, H_q)\rangle\\ &= \left\langle \nabla F(\phi(\theta)), \sum_{j=1}^q U_1 ... U_{j-1} H_j U_{j+1} ... U_q \right\rangle = \sum_{j=1}^q \left\langle \nabla F(\phi(\theta)), \widehat{V_j} H_j \widehat{W_j}\right\rangle\\ &= \sum_{j=1}^q \mathtt{Tr}\left([\nabla F(\phi(\theta))]^\top \widehat{V_j} H_j \widehat{W_j}\right) = \sum_{j=1}^q \mathtt{Tr}\left(H_j \widehat{W_j} [\nabla F(\phi(\theta))]^\top \widehat{V_j}\right)\\ &= \sum_{j=1}^q \left\langle H_j, \widehat{V_j}^\top \nabla F(\phi(\theta)) \widehat{W_j}^\top\right\rangle.\end{aligned}$$

Then by Riez theorem, viewing again for convenience ∇F as its matrix reshaped version, we get that for each $1 \le j \le q$:

$$\nabla_{U_j}\mathcal{E}_Z(\theta) = \underbrace{\widehat{V_j}^\top}_{n_{j-1} \times n_0} \underbrace{\nabla F(\phi(\theta))}_{n_0 \times n_q} \underbrace{\widehat{W_j}^\top}_{n_q \times n_j}. \tag{27}$$

Thus by specializing to $j \in \{i, i+1\}$ and by developing (27), one has:

$$\nabla_{U_i}\mathcal{E}_Z(\theta) = \underbrace{a}_{\text{size } n_{i-1} \times n_{i+1}} U_{i+1}^\top, \quad \text{and} \quad \nabla_{U_{i+1}}\mathcal{E}_Z(\theta) = U_i^\top \underbrace{b}_{\text{size } n_{i-1} \times n_{i+1}},$$

thus establishing (25). □

F. Application of Noether Theorem

Invariances of $\mathcal{E}_Z$ with respect to certain transformations have been used in the proof of Theorem 4.1. Here we establish more direct connections with Noether's theorem (see Theorem 3.5). For the sake of brevity we describe the case $q = 2$ but the same reasoning can easily be adapted to any $q \geq 2$ by considering the invariances associated to each pair U_i, U_{i+1}.

Invariances valid for any dimension. We consider the flow (3) in the Euclidean case ($M = \mathrm{Id}_d$), with $\theta := (U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ and $g(\theta, x) := UV^\top x$. We recall that for all $A \in \mathbb{R}^{r\times r}$, the linear transformation T^A from (15) leaves the cost (1) invariant, and that (see Example 3.2) that: $\Delta_{T^A}(U, V) = (UA, -VA^\top)$. Moreover for any A in $\mathcal{A}_r$ the space of skew-symmetric matrices in $\mathbb{R}^{r\times r}$, T^A leaves invariant the following Lagrangian (See for example (Głuch & Urbanke, 2021)), with which the Euclidian MF is compatible:

$$\mathcal{L}(t, \theta, v) := \exp\left(\int^t \tau(s)\mathrm{d}s\right)\left(1/2\|v\|^2 - \mathcal{E}_Z(\theta)\right). \tag{28}$$

Thus by Noether theorem (Theorem 3.5), for any $A \in \mathcal{A}_r$, a conserved function is

$$\begin{aligned} h(t, \theta, \dot{x}) :=& \partial_v \mathcal{L}(t, \theta, \dot{\theta}) \cdot \Delta_T^A(\theta(t)) = \exp\left(\int^t \tau(s)\mathrm{d}s\right)\left(\langle \dot{U}, UA\rangle - \langle \dot{V}, VA^\top\rangle\right) \\ &= \exp\left(\int^t \tau(s)\mathrm{d}s\right)\left(\langle \dot{U}, UA\rangle + \langle \dot{V}, VA\rangle\right). \end{aligned}$$

A supplementary invariance when $(n, m) = (1, 1)$. Assume $(n, m) = (1, 1)$, so that $U, V \in \mathbb{R}^{1\times r}$ and $\theta \in \mathbb{R}^{1\times 2r}$, and consider T'^A the linear transformation where for all $\epsilon \geq 0$:

$$T'^A(\epsilon, U, V) = \Big(V \exp(-\epsilon A^\top), U \exp(\epsilon A)\Big). \tag{29}$$

One has: $\Delta_{T'^A}(U, V) = (-VA^\top, UA)$. It is easy to check that since $g(\theta, x) := UV^\top x$, we have $g(T'^A(\epsilon, \cdot)) = g(\theta, \cdot)$ hence T'^A leaves invariant the cost (1). Moreover, one can also easily check that for any skew matrix $A \in \mathcal{A}_r$, the transformation T'^A also leaves the Lagrangian (28) invariant. Thus, by Noether theorem, for any $A \in \mathcal{A}_r$, the quantity

$$\begin{aligned} h(t, \theta, \dot{x}) :=& \partial_v \mathcal{L}(t, \theta, \dot{\theta}) \cdot \Delta_{T'^A}(\theta) = \exp\left(\int^t \tau(s)\mathrm{d}s\right)\left(\langle \dot{U}, -VA^\top\rangle + \langle \dot{V}, UA\rangle\right) \\ &= \exp\left(\int^t \tau(s)\mathrm{d}s\right)\left(\langle \dot{U}, VA\rangle + \langle \dot{V}, UA\rangle\right) \end{aligned}$$

is conserved.

G. Proof of Proposition 4.2

Proposition 4.2. *Consider $\widetilde{\Theta}$ the set of $\alpha = (t, \theta, \dot{\theta})$ such that $(\theta, \dot{\theta}) = (U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ has full rank, and assume $(n, m) \neq (1, 1)$. Then if $2(n + m) \leq r$, Theorem 4.1 gives exactly $(n + m)[(r - 2(n + m)) + r - 1]$ independent conserved functions. If $2(n + m) > r$, Theorem 4.1 gives exactly $r(r - 1)/2$ independent conserved functions.*

Proof. First, observe that if $r = 1$, then Theorem 4.1 gives zero conserved functions as $\mathcal{A}_1 = \{0\}$, and indeed here $2(n + m) > r = 1$ and $r(r - 1) = 0$.

We now focus on the case where $r > 1$ and denote $U = (u_1, \cdots, u_r)$, $V = (v_1, \cdots, v_r)$, $\dot{U} = (\dot{u}_1, \cdots, \dot{u}_r)$, $\dot{V} = (\dot{v}_1, \cdots, \dot{v}_r)$, with $u_i, \dot{u}_i \in \mathbb{R}^n$ and $v_i, \dot{v}_i \in \mathbb{R}^m$. For every $1 \leq i, j \leq r$, by Theorem 4.1 the elementary skew matrix $A_{i,j} := E_{i,j} - E_{j,i} \in \mathcal{A}_r$ leads to a conserved function $h_{i,j} := h_{A_{i,j}}$ (see (23)). Since $(n, m) \neq (1, 1)$, each conserved function predicted by Theorem 4.1 is of the form h_A, $A \in \mathcal{A}_r$, hence is a linear combination of $h_{i,j}$, $1 \leq i, j \leq r$. As a result, conserved function predicted by Theorem 4.1 satisfy $\nabla h_A(\alpha) \in \mathrm{span}\{\nabla h_{i,j}(\alpha) : 1 \leq i, j \leq r\}$ for every $\alpha = (t, \theta, \dot{\theta})$. We will show below that for any $\alpha \in \widetilde{\Theta}$, there is a set of indices $S \subseteq \{1, \ldots r\}$ of cardinality $R = \min(r, 2(n + m))$ and a neighborhood $\widetilde{\Omega}$ of α such that, with $T := \{(i, j) \in S \times S, i < j\} \cup (S \times S^c)$, we have for every $\alpha' \in \widetilde{\Omega}$:

- $\nabla h_{k,\ell}(\alpha') \in \mathrm{span}\{\nabla h_{i,j}(\alpha') : (i,j) \in T\}$ for every $(k,\ell) \notin T$;
- the vectors $\nabla h_{i,j}(\alpha')$, $(i,j) \in T$ are linearly independent.

This will conclude since

$$\sharp T = R(R-1)/2 + R(r-R) = \begin{cases} r(r-1)/2 & \text{if } R = r < 2(n+m) \\ (n+m)[(r-2(n+m)) + r - 1] & \text{if } R = 2(n+m) \leq r. \end{cases}$$

To proceed, specialize the definition (23) of h_A to our context: $q = 2$, $U_1 = U$ and $U_2 = V^\top$ yields $h_A(\alpha) = h_A(t,\theta,\dot\theta) = \exp\Big(\int^t \tau(s)\mathrm{d}s\Big) H_A(\theta,\dot\theta)$ with $H_A(\theta,\dot\theta) := \langle \dot U, UA\rangle + \langle \dot V^\top, V^\top A^\top\rangle = \langle \dot U, UA\rangle + \langle \dot V, VA\rangle = \langle \dot U A^\top, U\rangle + \langle \dot V A^\top, V\rangle$. As a result

$$\nabla h_{i,j}(t,\theta,\dot\theta) = (\tau(t) h_{i,j}(t,\theta,\dot\theta), \exp(\int^t \tau(s)\mathrm{d}s) \times [\nabla H_{A_{i,j}}(\theta,\dot\theta)]^\top)^\top \tag{30}$$

where, up to proper reshaping

$$\nabla H_A(\theta,\dot\theta) = \begin{pmatrix} \nabla_U H_A \\ \nabla_V H_A \\ \nabla_{\dot U} H_A \\ \nabla_{\dot V} H_A \end{pmatrix} = \begin{pmatrix} \dot U A^\top \\ \dot V A^\top \\ UA \\ VA \end{pmatrix} \overset{A^\top = -A}{=} \begin{pmatrix} -\dot U A \\ -\dot V A \\ UA \\ VA \end{pmatrix} = \underbrace{(-\dot U; -\dot V; U; V)}_{=:W} A.$$

hence the $2(n+m) \times r$ matrix $W = W(\alpha)$ will play a special role. Denote $w_j(\alpha)$, $1 \leq j \leq r$ its columns.

Observe that given $\alpha = (t,\theta,\dot\theta) \in \widetilde\Theta$, $R := \min(r, 2(n+m))$ is both the rank of $(\theta,\dot\theta) = (U;V;\dot U;\dot V)$ and the rank of W. Hence, given $\alpha \in \widetilde\Theta$, there is a subset of indices $S \subseteq \{1,\dots r\}$ of cardinality R such that the vectors $w_j(\alpha)$, $j \in S$ are linearly independent, while for $k \notin S$ we have $w_k(\alpha) \in \mathrm{span}(w_j(\theta,\dot\theta), j \in S)$. By standard calculus, there is a neighborhood $\widetilde\Omega \subseteq \widetilde\Theta$ of α such that these properties remain valid (with the same S) for every $\alpha' \in \widetilde\Omega$. We show below that this implies the claimed linear (in)dependence properties of the vectors $\nabla h_{i,j}(\alpha')$. From now on we omit the dependence in α' for the sake of brevity.

By (30), to show that the vectors $\nabla h_{i,j}$, $(i,j) \in T$ are linearly independent, it is sufficient to show the linear independence of $\nabla H_{A_{i,j}} := WA_{i,j}$, $(i,j) \in T$. Assume that $0_{2(n+m)\times r} = \sum_{(i,j)\in T} \lambda_{i,j} WA_{i,j}$. Our goal is now to show that $\lambda_{k,\ell} = 0$ for every $(k,\ell) \in T = (T \cap (S \times S)) \cup (S \times S^c)$. We first prove it for every $(k,\ell) \in S \times S^c$, then on $T \cap (S \times S)$.

First consider $(k,\ell) \in S \times S^c$. For any $(i,j) \in T$, by the definition of T we have $i \in S$, hence $\ell \neq i$ since $\ell \notin S$. Using the standard notation for canonical vectors and Kronecker deltas, we thus have $e_i^\top e_\ell = \delta_{i,\ell} = 0$, and since $A_{i,j} = e_i e_j^\top - e_j e_i^\top$ we obtain $A_{i,j} e_\ell = e_i \delta_{j,\ell}$, so that $WA_{i,j} e_\ell = \delta_{j,\ell} w_i$ and

$$0 = \sum_{(i,j)\in T} \lambda_{i,j} WA_{i,j} e_\ell = \sum_{(i,j)\in T} \lambda_{i,j} \delta_{j,\ell} w_i = \sum_{i\in S} \Big(\sum_{j:(i,j)\in T} \lambda_{i,j}\delta_{j,\ell} \Big) w_i.$$

where by convention an empty sum is zero. By the linear independence of w_i, $i \in S$ we get for each $i \in S$ that $0 = \sum_{j:(i,j)\in T} \lambda_{i,j}\delta_{j,\ell}$. Since $(k,\ell) \in T$, specializing to $i := k \in S$, we obtain $0 = \sum_{j:(k,j)\in T} \lambda_{k,j}\delta_{j,\ell} = \lambda_{k,\ell}$ as claimed.

Since the above holds for any $(k,\ell) \in S \times S^c$, and given the definition of T, we have established that in fact

$$0_{2(n+m)\times r} = \sum_{(i,j)\in T\cap(S\times S)} \lambda_{i,j} WA_{i,j} = W\Big(\sum_{(i,j)\in T\cap(S\times S)} \lambda_{i,j} A_{i,j} \Big) = WB,$$

where $B := \sum_{(i,j)\in T\cap(S\times S)} \lambda_{i,j} A_{i,j}$. Observe that, by definition of $A_{i,j}$, we have $B(S^c,:) = 0$ (the rows of B indexed by S^c are zero), hence $0 = WB = W(:,S)B(S,:)$. By the linear independence of the columns w_j, $j \in S$ of $W(:,S)$, we conclude that $B(S,:) = 0$, hence $B = 0$, that is to say $\sum_{(i,j)\in T\cap(S\times S)} \lambda_{i,j} A_{i,j} = 0$. Since $(i,j) \in T$ implies $i \neq j$ and as the matrices $A_{i,j}$, $i \neq j$, are linearly independent, we conclude that $\lambda_{i,j} = 0$ for every $(i,j) \in T \cap (S \times S)$.

This establishes that $\nabla h_{i,j}(\alpha')$, $(i,j) \in T$, are linearly independent.

To conclude the proof, there remains to show that $\nabla h_{k,\ell}(\alpha') \in \operatorname{span}\{\nabla h_{i,j}(\alpha') : (i,j) \in T\}$ for every $(k,\ell) \notin T$.

Consider $(k,\ell) \notin T$. As w_k and w_ℓ are linear combinations of $\{w_j\}_{j\in S}$ there exists $(\alpha_j)_j \neq (0)$ and $(\beta_j)_j \neq (0)$ such that $w_k = \sum_{j\in S} \alpha_j w_j$ and $w_\ell = \sum_{j\in S} \beta_j w_j$. As a result

$$\nabla h_{k,\ell} = \sum_{j\in S} (\alpha_j \nabla h_{j,\ell} - \beta_j \nabla h_{j,k}) - \sum_{i,j\in S} \alpha_i \beta_j \nabla h_{i,j}.$$

□

H. Proof of Theorem 4.3

First, we recall some definitions/results about Lie algebra, directly taken from (Marcotte et al., 2024) and we add a supplementary lemma (Lemma H.2) that states useful results.

H.1. Background on Lie algebra

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

Typically, the Lie algebra of interest is the set of infinitely smooth vector fields $\mathcal{A} := \mathcal{C}^\infty(\tilde{\Theta}, \mathbb{R}^{2D+1})$, endowed with the Lie bracket $[\cdot,\cdot]$ defined by

$$[\chi_1, \chi_2] : \quad \alpha \in \tilde{\Theta} \mapsto [\chi_1, \chi_2](\alpha) := \partial\chi_1(\alpha)\chi_2(\alpha) - \partial\chi_2(\alpha)\chi_1(\alpha), \tag{31}$$

with $\partial\chi(\alpha) \in \mathbb{R}^{(2D+1)\times(2D+1)}$ the jacobian of χ at α. The space $\mathbb{R}^{n\times n}$ of matrices is also a Lie algebra endowed with the Lie bracket $[A, B] := AB - BA$. This can be seen as a special case of (31) in the case of *linear* vector fields, i.e. $\chi(\alpha) = A\alpha$.

Generated Lie algebra Let $\mathcal{A}$ be a Lie algebra and let $\mathcal{W} \subset \mathcal{A}$ be a vector subspace of $\mathcal{A}$. There exists a smallest Lie algebra that contains $\mathcal{W}$. It is denoted $\operatorname{Lie}(\mathcal{W})$ and called the generated Lie algebra of $\mathcal{W}$. The following proposition (See Definition 20 of (Bonnard et al., 2018)) constructively characterizes $\operatorname{Lie}(\mathcal{W})$, where for vector subspaces $[\mathcal{W}, \mathcal{W}'] := \{[\chi_1, \chi_2] : \chi_1 \in \mathcal{W}, \chi_2 \in \mathcal{W}'\}$, and $\mathcal{W} + \mathcal{W}' = \{\chi_1 + \chi_2 : \chi_1 \in \mathcal{W}, \chi_2 \in \mathcal{W}'\}$.

Proposition H.1. *Given any vector subspace $\mathcal{W} \subseteq \mathcal{A}$ we have $\operatorname{Lie}(\mathcal{W}) = \bigcup_k \mathcal{W}_k$ where:*

$$\begin{cases} \mathcal{W}_0 & := \mathcal{W} \\ \mathcal{W}_k & := \mathcal{W}_{k-1} + [\mathcal{W}_0, \mathcal{W}_{k-1}] \ \textit{for} \ \ k \geq 1. \end{cases}$$

Lemma H.2. *let $\mathcal{W} \subseteq \mathcal{X}\left(\tilde{\Omega}\right)$ be a vector space. Then by considering*

$$\mathcal{D}(\mathcal{W}) := \operatorname{span}\{a(\cdot)\chi(\cdot) : a \in \mathcal{C}^\infty(\tilde{\Omega}, \mathbb{R}), \chi \in \mathcal{W}\}, \tag{32}$$

one has $\mathcal{D}(\operatorname{Lie}(\mathcal{D}(\mathcal{W}))) = \operatorname{Lie}(\mathcal{D}(\mathcal{W}))$ and for all $\alpha \in \tilde{\Omega}$, $[\operatorname{Lie}(\mathcal{W})](\alpha) = [\operatorname{Lie}(\mathcal{D}(\mathcal{W}))](\alpha)$.

Proof. **We first show by recursion that for all** $k \in \mathbb{N}$:

$$\mathcal{D}(\mathcal{W})_k = \mathcal{D}\left(\mathcal{W}_k\right). \tag{33}$$

where the iterates are defined in Proposition H.1.

Initialisation. One has: $\mathcal{D}(\mathcal{W})_0 = \mathcal{D}(\mathcal{W}) = \mathcal{D}(\mathcal{W}_0)$ by definition of the first iterate.

Recursion. Let $k \in \mathbb{N}$ and assume that k satifies (33). Let us show that $\mathcal{D}(\mathcal{W})_{k+1} = \mathcal{D}\left(\mathcal{W}_{k+1}\right)$. By definition (Proposition H.1), one has

$$\mathcal{D}\left(\mathcal{W}\right)_{k+1} = \mathcal{D}\left(\mathcal{W}\right)_k + [\mathcal{D}\left(\mathcal{W}\right), \mathcal{D}\left(\mathcal{W}\right)_k]. \tag{34}$$

Thus $\mathcal{D}\left(\mathcal{W}\right)_{k+1} \overset{(33)}{=} \mathcal{D}\left(\mathcal{W}_k\right) + [\mathcal{D}\left(\mathcal{W}\right), \mathcal{D}\left(\mathcal{W}_k\right)]$. We first show the direct inclusion $\mathcal{D}(\mathcal{W})_{k+1} \subseteq \mathcal{D}(\mathcal{W}_{k+1})$. Since $\mathcal{W}_k \subseteq \mathcal{W}_{k+1}$ by construction, we have $\mathcal{D}(\mathcal{W}_k) \subseteq \mathcal{D}(\mathcal{W}_{k+1})$, hence it is enough to show that $[\mathcal{D}\left(\mathcal{W}\right), \mathcal{D}\left(\mathcal{W}_k\right)] \subseteq \mathcal{D}\left(\mathcal{W}_{k+1}\right)$.

Let $X \in \mathcal{D}(\mathcal{W})$ and $Y \in \mathcal{D}(\mathcal{W}_k)$. By definition of the operator $\mathcal{D}$, there are smooth real-valued functions a_i, b_j and $\chi_i \in \mathcal{W}, \mu_j \in \mathcal{W}_k$ such that $X(\cdot) = \sum_1^{m_1} a_i(\cdot)\chi_i(\cdot)$ and $Y(\cdot) = \sum_1^{m_2} b_j(\cdot)\mu_j(\cdot)$ on $\tilde{\Omega}$ and we deduce by bilinearity of the Lie brackets that $[X, Y](\cdot) = \sum_{i,j}[a_i\chi_i, b_j\mu_j](\cdot)$ on $\tilde{\Omega}$. Moreover, one has:

$$[a_i\chi_i, b_j\mu_j] = \underbrace{a_i b_j[\chi_i, \mu_j]}_{\in\mathcal{D}([\mathcal{W},\mathcal{W}_k])} + \underbrace{b_j[(\partial a_i)\mu_j]\chi_i}_{\in\mathcal{D}(\mathcal{W})} - \underbrace{a_i[(\partial b_j)\chi_i]\mu_j}_{\in\mathcal{D}(\mathcal{W}_k))}, \tag{35}$$

where, due to dimensions, both $(\partial a_i)\mu_j$ and $(\partial b_j)\chi_i$ are smooth scalar-valued functions. Thus $[X, Y] \in \mathcal{D}([\mathcal{W}, \mathcal{W}_k] + \mathcal{W}_k) = \mathcal{D}(\mathcal{W}_{k+1})$. Finally, $[\mathcal{D}(\mathcal{W}), \mathcal{D}(\mathcal{W}_k)] \subseteq \mathcal{D}(\mathcal{W}_{k+1})$ and thus $\mathcal{D}(\mathcal{W})_{k+1} \subseteq \mathcal{D}(\mathcal{W}_{k+1})$. We now show the converse inclusion. Let $X \in \mathcal{D}(\mathcal{W}_{k+1})$. There are smooth real-valued functions a_i and $Y_i \in \mathcal{W}_{k+1}$, such that $X = \sum_i a_i Y_i$. As by definition: $\mathcal{W}_{k+1} = \mathcal{W}_k + [\mathcal{W}_k, \mathcal{W}]$, there exists $\chi_i, v_i \in \mathcal{W}_k, \mu_i \in \mathcal{W}$ such that: $Y_i = v_i + [\chi_i, \mu_i]$. Thus $X = \sum_i a_i(v_i + [\chi_i, \mu_i])$. As $\mathcal{D}(\mathcal{W}_{k+1})$ is a linear space, it is enough to show that both $a_i v_i$ and $a_i[\chi_i, \mu_i]$ are in $\mathcal{D}(\mathcal{W}_{k+1})$. As $v_i \in \mathcal{W}_k \subseteq \mathcal{W}_{k+1}$, one has directly that $a_i v_i \in \mathcal{D}(\mathcal{W}_{k+1})$. Finally by using again equality (35) (with $b_i \equiv 1$ and $j = i$), one has:

$$a_i[\chi_i, \mu_i] = [a_i\chi_i, \mu_i] - (\partial a_i \mu_i)\chi_i.$$

Both $a_i\chi_i$ and $(\partial a_i\mu_i)\chi_i$ are in $\mathcal{D}(\mathcal{W}_k) \overset{(33)}{=} \mathcal{D}(\mathcal{W})_k$ since a_i and $\partial a_i\mu_i$ are smooth real-valued functions. We also have $\mu_i \in \mathcal{W} \subseteq \mathcal{D}(\mathcal{W})$ hence, using the characterization (34), one has $a_i[\chi_i, \mu_i] \in \mathcal{D}(\mathcal{W})_{k+1}$ and thus $X \in \mathcal{D}(\mathcal{W})_{k+1}$, which concludes the recursion.

We now prove that $\mathcal{D}(\mathrm{Lie}(\mathcal{D}(\mathcal{W}))) = \mathrm{Lie}(\mathcal{D}(\mathcal{W}))$. One inclusion is trivial so we only need to prove the other one. Let $X \in \mathrm{Lie}(\mathcal{D}(\mathcal{W}))$ and let us consider a a smooth real-valued function. To conclude, we need to show that $aX \in \mathrm{Lie}(\mathcal{D}(\mathcal{W}))$. By definition of the generated Lie algebra Proposition H.1, there exists k such that $X \in \mathcal{D}(\mathcal{W})_k$. Then, by using (33), one has $X \in \mathcal{D}(\mathcal{W}_k)$ and thus $aX \in \mathcal{D}(\mathcal{W}_k) \overset{(33)}{=} \mathcal{D}(\mathcal{W})_k \subseteq \mathrm{Lie}(\mathcal{D}(\mathcal{W}))$.

Finally, we now prove that for all $\alpha \in \tilde{\Omega}$**,** $[\mathrm{Lie}(\mathcal{W})](\alpha) = [\mathrm{Lie}(\mathcal{D}(\mathcal{W}))](\alpha)$**.** Let $\alpha \in \tilde{\Omega}$. By using (33), one has for all $k \in \mathbb{N}$, $[\mathcal{D}(\mathcal{W})_k](\alpha) = [\mathcal{D}(\mathcal{W}_k)](\alpha) = [\mathcal{W}_k](\alpha)$. Then $[\mathrm{Lie}(\mathcal{D}(\mathcal{W}))](\alpha) = [\mathrm{Lie}(\mathcal{W})](\alpha)$, which concludes the proof. □

H.2. Proof of Theorem 4.3

First, we recall the statement of the theorem for the reader's convenience.

Theorem 4.3. *Consider $g(\theta, x) := UV^\top x$ with $(U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. If $(U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ has full rank noted* `rk` *then, in a neighborhood of $(t, U, V, \dot{U}, \dot{V})$:*

- *If $(n, m) \neq (1, 1)$, there are exactly $\mathtt{rk}/2 \cdot (2r - 1 - \mathtt{rk})$ independent conservation laws of $\phi_{\mathtt{Lin}}$.*
- *If $n = m = 1$ and if $r \geq 4$, there are exactly $4r - 6$ independent conservation laws of $\phi_{\mathtt{Lin}}$.*

The proof of this theorem relies on Theorem 3.7: we compute $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha)$ with $\phi = \phi_{\mathtt{Lin}}$ and show that its dimension is locally constant around $\alpha = (t, \theta, \dot{\theta})$ where, $\theta = \mathtt{vec}((U; V))$ and $\dot{\theta} = \mathtt{vec}((\dot{U}; \dot{V}))$ are vectorized versions of two $(n + m) \times r$ matrices such that $(U; V; \dot{U}; \dot{V}) \in \mathbb{R}^{2(n+m)\times r}$ has full rank.

First we make $\mathcal{W}_\phi^{\mathtt{mom}}$ more explicit. Recall that $\mathcal{W}_\phi^{\mathtt{mom}}$ is the functional linear space spanned by the functions $\alpha \mapsto \chi_i(\alpha)$ defined in (8). Since we consider the Euclidean geometry, we have $M = \mathrm{Id}_D$ so that any function $\chi \in \mathcal{W}_\phi^{\mathtt{mom}}$ is a linear combination of $\chi_i(\cdot)$, $0 \leq i \leq d$, hence it satisfies

$$\chi(\alpha) = \beta\chi_0(\alpha) + \sum_{i=1}^{d} \gamma_i\chi_i(\alpha) = \begin{pmatrix} \beta \\ \beta\dot{\theta} \\ [\partial\phi(\theta)]^\top\gamma - \tau(t)\beta\dot{\theta} \end{pmatrix}$$

for some $\beta \in \mathbb{R}$ and $\gamma := (\gamma_i)_{i=1}^d \in \mathbb{R}^d$. Since $d = nm$, we can write $\gamma = \mathtt{vec}(\Delta)$ with $\Delta \in \mathbb{R}^{n\times m}$, and leveraging (Marcotte et al., 2024, Proposition H.2), we obtain $\partial\phi(\theta)^\top\gamma = \partial\phi(\theta)^\top\mathtt{vec}(\Delta) = \mathtt{vec}(S_\Delta(U; V))$ where $S_\Delta := \begin{pmatrix} 0 & \Delta \\ \Delta^\top & 0 \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$. Using a basic property[4] of Kronecker products ($(\mathbf{A} \otimes \mathbf{B})\mathtt{vec}(\mathbf{X}) = \mathtt{vec}(\mathbf{B}\mathbf{X}\mathbf{A}^\top)$), this is further rewritten as $\partial\phi(\theta)^\top\gamma = \mathtt{vec}(S_\Delta(U; V)\mathrm{Id}_r) = (\mathrm{Id}_r \otimes S_\Delta)\theta$. Overall we obtain that $\mathcal{W}_\phi^{\mathtt{mom}}$ is the collection

[4] this property can indeed serve as an *operational definition* of the Kronecker product between matrices.

of all vector fields

$$\chi_{\beta,\Delta}(\alpha) := \begin{pmatrix} \beta \\ \beta\dot{\theta} \\ (\mathrm{Id}_r \otimes S_\Delta)\theta - \tau(t)\beta\dot{\theta} \end{pmatrix}.$$

Second, we express $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha)$. We highlight in purple the results and reasoning steps specific to $n = m = 1$.

Proposition H.3. *Denote $\mathcal{S}_\ell \subset \mathbb{R}^{\ell\times\ell}$ the space of symmetric matrices, $H := \begin{pmatrix} 0 & \mathrm{Id}_{n+m} \\ -\mathrm{Id}_{n+m} & 0 \end{pmatrix}$, and for any square matrix M of size $2(n+m)$ and $\beta \in \mathbb{R}$ denote*

$$\eta_{\beta,M}(\alpha) := \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HM)] \cdot \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 0_{(n+m)r} \\ -\tau(t)\dot{\theta} \end{pmatrix}.$$

We have $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = \{\eta_{\beta,M}(\alpha) : \beta \in \mathbb{R}, M \in \mathcal{S}\}$ where $\mathcal{S} := \mathcal{S}_{2(n+m)}$ when $(n,m) \neq (1,1)$, while for $n = m = 1$, $\mathcal{S} := \tilde{\mathcal{S}}_4 := \left\{ \begin{pmatrix} S_1 & S \\ S & S_2 \end{pmatrix} : S, S_1, S_2 \in \mathcal{S}_2' \right\} \subsetneq \mathcal{S}_4$, where $\mathcal{S}_2' \subsetneq \mathcal{S}_2$ is the set of symmetric matrices on the form $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$.

Proof. **First we show that $\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}) \subseteq \mathcal{D}(\mathcal{W})$,** where $\mathcal{W} := \{\eta_{\beta,M} : \beta \in \mathbb{R}, M \in \mathcal{S}\}$ and with the operator $\mathcal{D}$ defined by (32). One has (see Appendix H.3 for more details):

$$\chi_{\beta,\Delta} = \eta_{\beta,M}, \text{ where } M := \begin{pmatrix} -S_\Delta & 0 \\ 0 & \beta\mathrm{Id}_{(n+m)} \end{pmatrix}. \tag{36}$$

The matrix M is symmetric and belongs to $\mathcal{S}$ even when $n = m = 1$. Thus $\mathcal{W}_\phi^{\mathtt{mom}} \subseteq \mathcal{W}$ and $\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}) \subseteq \mathcal{D}(\mathcal{W})$.

We now prove that $\mathcal{D}(\mathcal{W})$ is a Lie algebra. Since for all a, b smooth real-valued functions and for all $X, Y \in \mathcal{W}$ one has:

$$[aX, bY] = ab[X,Y] + \underbrace{b(\partial aY)X}_{\in\mathcal{D}(\mathcal{W})} - \underbrace{a(\partial bX)Y}_{\in\mathcal{D}(\mathcal{W})},$$

where, due to dimensions, both $(\partial a)Y$ and $(\partial b)X$ are smooth scalar-valued functions, it is enough to show that $[\mathcal{W}, \mathcal{W}] \subseteq \mathcal{D}(\mathcal{W})$. Moreover, since $\eta_{\beta,M} = \eta_{0,M} + \beta\eta_{1,0}$, it is enough to check that $[\eta_{0,M}, \eta_{0,M'}]$ and $[\eta_{1,0}, \eta_{0,M}]$ are elements of $\mathcal{D}(\mathcal{W})$ whenever $M, M' \in \mathcal{S}$.

- We first obtain (see Appendix H.3 for more details) that

$$[\eta_{0,M}, \eta_{0,M'}] = \eta_{0,M''}, \text{ where } M'' := MHM' - M'HM. \tag{37}$$

  Since $H^\top = -H$ it is straightforward to check that $M'' \in \mathcal{S}_{2(n+m)}$, hence $M'' \in \mathcal{S}$ when $(n,m) \neq (1,1)$, and we let the reader check that when $(n,m) = (1,1)$ we also have $M'' \in \mathcal{S}$ as soon as $M, M' \in \mathcal{S}$ (note that $\mathcal{S}_2'$ is stable by matrix multiplication and is commutative). Therefore $[\eta_{0,M}, \eta_{0,M'}] = \eta_{0,M''} \in \mathcal{W} \subseteq \mathcal{D}(\mathcal{W})$.

- We now show that $[\eta_{0,M}, \eta_{1,0}] \in \mathcal{D}(\mathcal{W})$ for every $M \in \mathcal{S}$. Since $M \in \mathcal{S}$ we can write it as $M = \begin{pmatrix} S_1 & S \\ S^\top & S_2 \end{pmatrix}$ with $S \in \mathbb{R}^{(n+m)\times(n+m)}$, $S_1, S_2 \in \mathcal{S}_{n+m}$ (when $n = m = 1$ we further have $S, S_1, S_2 \in \mathcal{S}_2'$). We then obtain (see Appendix H.3 for more details):

$$[\eta_{0,M}, \eta_{1,0}](\alpha) = \tau(t)\eta_{0,M'}(\alpha) \text{ with } M' := \begin{pmatrix} S_1 & 0 \\ 0 & -S_2 \end{pmatrix}. \tag{38}$$

  Since $t \mapsto \tau(t)$ is $\mathcal{C}^\infty$ and $M' \in \mathcal{S}$ (even when $n = m = 1$) this implies $[\eta_{0,M}, \eta_{1,0}] \in \mathcal{D}(\mathcal{W})$.

This establishes as claimed that $\mathcal{D}(\mathcal{W})$ is indeed a Lie algebra.

Finally we prove that $\mathcal{W} \subseteq \mathrm{Lie}(\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}))$. This is where the necessity to impose the more restricted definition of $\mathcal{S}$ for $n = m = 1$ will become evident. Before proving this inclusion, observe that by Lemma H.2 it will imply $\mathcal{D}(\mathcal{W}) \subseteq \mathcal{D}\left(\mathrm{Lie}(\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}))\right) = \mathrm{Lie}(\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}))$, hence combined with what we already proved it implies $\mathcal{D}(\mathcal{W}) = \mathrm{Lie}(\mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}}))$.

As a shorthand denote $\mathcal{V} := \mathcal{D}(\mathcal{W}_\phi^{\mathtt{mom}})$. First, we will building matrix sets $\mathcal{M}_k, \mathcal{M}'_k$ such that $\eta_{0,M} \in \mathcal{V}_k$ for every $M \in \mathcal{M}_k \cup \mathcal{M}'_k$. Then we will show that $\mathcal{W} \subseteq \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V}$ where $\mathcal{M}$ is the linear span of all built matrix sets. Since $\mathcal{V}_k \subseteq \mathrm{Lie}(\mathcal{V})$ for every k, this will yield the desired conclusion.

- *We first prove that $\eta_{0,M} \in \mathcal{V}_k$ for every $M \in \mathcal{M}_k$.*

The set $\mathcal{M}_0$ is defined as the collection of all matrices that write as

$$M = \begin{pmatrix} -S_\Delta & 0 \\ 0 & 0 \end{pmatrix}$$

for some Δ. By (36), for any Δ we have $\eta_{0,M} = \chi_{0,\Delta} \in \mathcal{V}_0$.

The set $\mathcal{M}_1$ is defined as the linear span of $\mathcal{M}_0$ and of the set of all matrices that write as

$$M = \begin{pmatrix} 0 & S_\Delta \\ S_\Delta & 0 \end{pmatrix}.$$

Consider $M_1 := \begin{pmatrix} -S_\Delta & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_0$ and $M_2 := \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_{n+m} \end{pmatrix}$. As $\eta_{1,M_2} = \eta_{0,M_2} + \eta_{1,0}$, by bilinearity of Lie brackets:

$$[\chi_{0,\Delta}, \chi_{1,0}](\alpha) \overset{(36)}{=} [\eta_{0,M_1}, \eta_{1,M_2}](\alpha) = [\eta_{0,M_1}, \eta_{0,M_2}](\alpha) + [\eta_{0,M_1}, \eta_{1,0}](\alpha) \overset{(37),(38)}{=} \eta_{0,M_3}(\alpha) + \tau(t)\eta_{0,M_4}(\alpha),$$

where $M_3 \overset{(37)}{:=} M_1 H M_2 - M_2 H M_1 = \begin{pmatrix} 0 & -S_\Delta \\ -S_\Delta & 0 \end{pmatrix}$ and $M_4 \overset{(38)}{=} M_1 \in \mathcal{M}_0$. Thus: $\eta_{0,M_3}(\alpha) = [\chi_{0,\Delta}, \chi_{1,0}](\alpha) - \tau(t)\eta_{0,M_4}(\alpha) \in \mathcal{V}_1$ since $t \mapsto \tau(t)$ is $\mathcal{C}^\infty$.

The set $\mathcal{M}_2$ is defined as the linear span of $\mathcal{M}_1$ and of the set of matrices

$$M = \begin{pmatrix} 0 & 0 \\ 0 & S_\Delta \end{pmatrix}.$$

Consider $M_1 := \begin{pmatrix} 0 & S_\Delta \\ S_\Delta & 0 \end{pmatrix} \in \mathcal{M}_1$ and $M_2 := \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_{n+m} \end{pmatrix}$. Since $\eta_{0,M_1} \in \mathcal{V}_1$ and $\chi_{1,0} \in \mathcal{V}$, we have $[\eta_{0,M_1}, \eta_{1,M_2}] \overset{(36)}{=} [\eta_{0,M_1}, \chi_{1,0}] \in \mathcal{V}_2$ and as: $[\eta_{0,M_1}, \eta_{1,M_2}](\alpha) = [\eta_{0,M_1}, \eta_{0,M_2}](\alpha) + [\eta_{0,M_1}, \eta_{1,0}](\alpha) \overset{(37),(38)}{=} \eta_{0,M_3}(\alpha) + \tau(t)\eta_{0,M_4}(\alpha)$, where $M_3 \overset{(37)}{:=} M_1 H M_2 - M_2 H M_1 = -2\begin{pmatrix} 0 & 0 \\ 0 & S_\Delta \end{pmatrix}$ and $M_4 \overset{(38)}{=} 0$. Thus $\eta_{0,M_4} = 0$ and we obtain $\eta_{0,M_3} = [\eta_{0,M_1}, \chi_{1,0}] \in \mathcal{V}_2$.

The set $\mathcal{M}_3$ is defined as the linear span of $\mathcal{M}_2$ and of the set of matrices

$$M = \begin{pmatrix} 0 & S_\Delta S_{\Delta'} \\ S_{\Delta'} S_\Delta & 0 \end{pmatrix},$$

for any Δ, Δ'. Such a matrix satisfies $M = M_2 H M_1 - M_1 H M_2$ where M_1, M_2 are specified as $M_1 := \begin{pmatrix} 0 & 0 \\ 0 & S_\Delta \end{pmatrix} \in \mathcal{M}_2$ and $M_2 := \begin{pmatrix} S_{\Delta'} & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_0$. By (37) it follows that $\eta_{0,M} = [\eta_{0,M_2}, \eta_{0,M_1}] \in \mathcal{V}_3$.

Before defining $\mathcal{M}_5$ we further explicit matrices contained in $\mathcal{M}_3$. Given any $\Delta, \Delta' \in \mathbb{R}^{n\times m}$, denote $B_1 := \Delta\Delta'^\top$ and $B_2 := \Delta^\top \Delta'$. Since $S_\Delta S_{\Delta'} = \begin{pmatrix} \Delta\Delta'^\top & 0 \\ 0 & \Delta^\top\Delta' \end{pmatrix}$ we have

$$\begin{pmatrix} 0 & 0 & B_1 & 0 \\ 0 & 0 & 0 & B_2 \\ B_1^\top & 0 & 0 & 0 \\ 0 & B_2^\top & 0 & 0 \end{pmatrix} \in \mathcal{M}_3, \tag{39}$$

for any pair of matrices B_1, B_2 that can be written as above. We explicit a few such matrices.

In the case $n = m = 1$, $B_1 = B_2 = b \in \mathbb{R}$, and as $\begin{pmatrix} 0 & S_\Delta \\ S_\Delta & 0 \end{pmatrix} \in \mathcal{M}_1$ and $S_\Delta = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}$ with $a \in \mathbb{R}$, one has

$$\begin{pmatrix} 0 & B \\ B & 0 \end{pmatrix} \in \mathcal{M}_3, \text{ for any } B \in \mathcal{S}_2'. \tag{40}$$

When $(n, m) \neq (1, 1)$, consider any $1 \leq i, k \leq n$, $1 \leq j, l \leq m$, and $\Delta := E_{i,j} \in \mathbb{R}^{n\times m}$ and $\Delta' := E_{k,l} \in \mathbb{R}^{n\times m}$. Since $B_1 := \Delta\Delta'^\top = E_{i,j}E_{l,k} = \delta_{j,l}E_{i,k} \in \mathbb{R}^{n\times n}$ and $B_2 := \Delta^\top\Delta' = E_{j,i}E_{k,l} = \delta_{i,k}E_{j,l} \in \mathbb{R}^{m\times m}$, we can reach:

- $(B_1, B_2) = (E_{i,i}, E_{j,j})$, for any $1 \leq i \leq n$ and $1 \leq j \leq m$, by choosing $k := i$ and $l := j$;
- (in the case $m > 1$) $(B_1, B_2) = (0, E_{j,l})$ for any $1 \leq j \neq l \leq m$, by choosing e.g. $i = k = 1$;
- (in the case $n > 1$): $(B_1, B_2) = (E_{i,k}, 0)$ for any $1 \leq i \neq k \leq n$, by choosing $j = l = 1$.

The two following steps are specific to the case $(n, m) \neq (1, 1)$ only.

The set $\mathcal{M}_5$ (defined only when $(n, m) \neq (1, 1)$ – and we skip the definition of $\mathcal{M}_4$) is defined as the linear span of $\mathcal{M}_3$ and of the set of matrices

$$M = \begin{pmatrix} 0 & 0 & 0 & E_{i,j} \\ 0 & 0 & -E_{i,j}^\top & 0 \\ 0 & -E_{i,j} & 0 & 0 \\ E_{i,j}^\top & 0 & 0 & 0 \end{pmatrix} \tag{41}$$

with $1 \leq i \leq n$ and $1 \leq j \leq m$. Since $(n, m) \neq (1, 1)$, without loss of generality, assume that $n > 1$ (a similar construction can be done in the case $m > 1$). Observe that $M = M_2HM_1 - M_1HM_2$ where M_1, M_2 are specified as $M_1 := \begin{pmatrix} 0 & S_\Delta \\ S_\Delta & 0 \end{pmatrix} \in \mathcal{M}_1$ (NB: not $\mathcal{M}_0$) with $\Delta := E_{i,j}$, and $M_2 := \begin{pmatrix} 0 & 0 & E_{k,k} & 0 \\ 0 & 0 & 0 & E_{j,j} \\ E_{k,k} & 0 & 0 & 0 \\ 0 & E_{j,j} & 0 & 0 \end{pmatrix} \in \mathcal{M}_3$ (by (39)) for some $1 \leq k \neq i \leq n$ (such a choice of k is possible since $n > 1$). By (37) it follows (using Jacobi identity and the fact that $M_1 \in \mathcal{M}_1$) that $\eta_{0,M} = [\eta_{0,M_2}, \eta_{0,M_1}] \in \mathcal{V}_5$ (but not $\mathcal{V}_4$).

Again, before defining the set $\mathcal{M}_7$ we show that for any $\Delta_1 \in \mathbb{R}^{n\times m}$, $\Delta_2 \in \mathbb{R}^{m\times n}$ we have

$$M' := \begin{pmatrix} 0 & 0 & 0 & \Delta_1 \\ 0 & 0 & \Delta_2 & 0 \\ 0 & \Delta_2^\top & 0 & 0 \\ \Delta_1^\top & 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}_5. \tag{42}$$

Indeed $2M' = \begin{pmatrix} 0 & 0 & 0 & \Delta_1 \\ 0 & 0 & -\Delta_1^\top & 0 \\ 0 & -\Delta_1 & 0 & 0 \\ \Delta_1^\top & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & S_{\Delta_1} \\ S_{\Delta_1} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & -\Delta_2^\top \\ 0 & 0 & \Delta_2 & 0 \\ 0 & \Delta_2^\top & 0 & 0 \\ -\Delta_2 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & S_{\Delta_2^\top} \\ S_{\Delta_2^\top} & 0 \end{pmatrix}$ where the first and third terms are combinations of matrices shaped as (41), while the second and last belong to $\mathcal{M}_1$.

The set $\mathcal{M}_7$ (defined only when $(n, m) \neq (1, 1)$ – and again we skip the definition of $\mathcal{M}_6$) is defined as the linear span of $\mathcal{M}_5$ and of the set of matrices

$$M = \begin{pmatrix} 0 & 0 & E_{i,i} & 0 \\ 0 & 0 & 0 & -E_{j,j} \\ E_{i,i} & 0 & 0 & 0 \\ 0 & -E_{j,j} & 0 & 0 \end{pmatrix} \tag{43}$$

with $1 \leq i \leq n$ and $1 \leq j \leq m$. Observe that $M = M_2HM_1 - M_1HM_2$ where $M_1 := \begin{pmatrix} 0 & S_\Delta \\ S_\Delta & 0 \end{pmatrix} \in \mathcal{M}_1$ (NB: not $\mathcal{M}_0$) with $\Delta = E_{i,j}$, and $M_2 := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & E_{j,i} & 0 \\ 0 & E_{j,i}^\top & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}_5$ (by (42)). By (37) it follows (using again Jacobi identity

and the fact that $M_1 \in \mathcal{M}_1$) that $\eta_{0,M} = [\eta_{0,M_2}, \eta_{0,M_1}] \in \mathcal{V}_7$ (and not $\mathcal{V}_6$).

As $\mathcal{M}_7$ is a vector space and since we have already (39) with $(B_1, B_2) = (E_{ii}, E_{jj})$, by linear combination with matrices shaped as in (43) we obtain that any matrix shaped as in (39) with arbitrary diagonal B_1, B_2 also belongs to $\mathcal{M}_7$. Arbitrary off-diagonal terms can be obtained by combining matrices shaped as in (39) (if $m > 1$, $B_1 = 0$ and $B_2 = E_{j,l}$ and if $n > 1$, $B_1 = E_{i,k}$ and $B_2 = 0$), and we obtain

$$M = \begin{pmatrix} 0 & 0 & B_1 & 0 \\ 0 & 0 & 0 & B_2 \\ B_1^\top & 0 & 0 & 0 \\ 0 & B_2^\top & 0 & 0 \end{pmatrix} \in \mathcal{M}_7, \text{ for each } B_1 \in \mathbb{R}^{n\times n}, B_2 \in \mathbb{R}^{m\times m}. \tag{44}$$

Finally, combining (44) and (42) one has

$$\begin{pmatrix} 0 & B \\ B^\top & 0 \end{pmatrix} \in \mathcal{M}_7 \text{ for any } B \in \mathbb{R}^{(n+m)\times(n+m)}. \tag{45}$$

The set $\mathcal{M}_8$ (resp. $\mathcal{M}'_4$) is defined as the linear span of $\mathcal{M}_7$ (resp. of $\mathcal{M}_3$) and of the set of matrices

$$M = \begin{pmatrix} \Delta_1\Delta^\top + \Delta\Delta_1^\top & 0 & 0 & 0 \\ 0 & \Delta_2\Delta + \Delta^\top\Delta_2^\top & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \text{ (resp. with } \Delta_2 = \Delta_1^\top\text{)},$$

and

$$M' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \Delta\Delta_2 + \Delta_2^\top\Delta^\top & 0 \\ 0 & 0 & 0 & \Delta^\top\Delta_1 + \Delta_1^\top\Delta \end{pmatrix}, \text{ (resp. with } \Delta_2 = \Delta_1^\top\text{)},$$

which satisfy $M = M_2HM_1 - M_1HM_2$ and $M' = M'_2HM_1 - M_1HM'_2$ where $M_1 := \begin{pmatrix} 0 & 0 & 0 & \Delta_1 \\ 0 & 0 & \Delta_2 & 0 \\ 0 & \Delta_2^\top & 0 & 0 \\ \Delta_1^\top & 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}_5$ (in the case $(n,m) = (1,1)$, $M_1 = \begin{pmatrix} 0 & S_{\Delta_1} \\ S_{\Delta_1} & 0 \end{pmatrix} \in \mathcal{M}_1$), $M_2 := \begin{pmatrix} S_\Delta & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_0$ and $M'_2 := -\begin{pmatrix} 0 & 0 \\ 0 & S_\Delta \end{pmatrix} \in \mathcal{M}_2$.

By (37) it directly follows that $\eta_{0,M} = [\eta_{0,M_2}, \eta_{0,M_1}] \in \mathcal{V}_6 \subseteq \mathcal{V}_8$ (resp. $\eta_{0,M} \in \mathcal{V}_2 \subseteq \mathcal{V}_4$). Similarly, using Jacobi identity and the fact that $M'_2 \in \mathcal{M}_2$ (resp. that $M_1 \in \mathcal{M}_1$) we obtain that $\eta_{0,M'} = [\eta_{0,M'_2}, \eta_{0,M_1}] \in \mathcal{V}_8$ (resp. $\eta_{0,M'} \in \mathcal{V}_4$).

Again, we now explicit matrices belonging to $\mathcal{M}_8$ (resp. to $\mathcal{M}'_4$).

By considering $M_3 := \begin{pmatrix} S_{\Delta''} & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_0$, we obtain that $M + M_3 = \begin{pmatrix} S_1 & \Delta'' & 0 & 0 \\ \Delta''^\top & S_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}'_8$ (resp. $\mathcal{M}'_4$) , with $S_1 := \Delta_1\Delta^\top + \Delta\Delta_1^\top \in \mathcal{S}_n$ and $S_2 := \Delta_2\Delta + \Delta^\top\Delta_2^\top \in \mathcal{S}_m$ (resp. in the case $n = m = 1$, one has $S_1 = S_2$ as $\Delta_2 = \Delta_1^\top$), and since this holds for any choice of $\Delta, \Delta_1, \Delta_2, \Delta''$, one has

$$\begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}'_8 \text{ (resp. } \mathcal{M}'_4\text{) for any } S \in \mathcal{S}_{n+m} \text{ (resp. } \mathcal{S}'_{n+m}\text{)}. \tag{46}$$

Similarly, by considering $M'_3 := \begin{pmatrix} 0 & 0 \\ 0 & S_{\Delta''} \end{pmatrix} \in \mathcal{M}_2$, we obtain that $M' + M'_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & S_1 & \Delta'' \\ 0 & 0 & \Delta''^\top & S_2 \end{pmatrix} \in \mathcal{M}'_8$ (resp. $\mathcal{M}'_4$) , with $S_1 := \Delta_1\Delta^\top + \Delta\Delta_1^\top \in \mathcal{S}_n$ and $S_2 := \Delta_2\Delta + \Delta^\top\Delta_2^\top \in \mathcal{S}_m$ (resp. in the case $n = m = 1$, one

has $S_1 = S_2$ as $\Delta_2 = \Delta_1^\top$), and thus one has

$$\begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \in \mathcal{M}'_8 \text{ (resp. } \mathcal{M}'_4\text{) for any } S \in \mathcal{S}_{n+m} \text{ (resp. } \mathcal{S}'_2\text{).} \tag{47}$$

Thus by combining (46) and (47), one has:

$$\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \in \mathcal{M}'_8 \text{ (resp. } \mathcal{M}'_4\text{) for each } S_1, S_2 \in \mathcal{S}_{n+m} \text{ (resp. } \mathcal{S}'_2\text{).} \tag{48}$$

- *We now show that $\mathcal{W} \subseteq \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V}$ where $\mathcal{M}$ is the linear span of all built matrix sets.*

First by combining Equation (45) and Equation (48), one has $\{\eta_{0,M}, M \in \mathcal{S}\} \subseteq \{\eta_{0,M}, M \in \mathcal{M}\}$

For any $\beta \in \mathbb{R}$ one has: $\beta\eta_{1,0}(\alpha) = \beta \begin{pmatrix} 1 \\ 0 \\ -\tau(t)\dot{\theta} \end{pmatrix} = \chi_{\beta,0}(\alpha) - \beta \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} = \chi_{\beta,0}(\alpha) - \eta_{0,\beta M}(\alpha)$, where $M := \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_{n+m} \end{pmatrix} \in \mathcal{M}$ (by (47)), and $\chi_{\beta,0} \in \mathcal{V}$. Therefore $\beta\eta_{1,0} \in \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V}$. Then, for any β, $M \in \mathcal{S}$, $\eta_{\beta,M} = \beta\eta_{1,0} + \eta_{0,M} \in \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V}$ so that: $\mathcal{W} \subseteq \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V}$.

- *Conclusion.*

Since $\mathcal{V}_k \subseteq \mathrm{Lie}(\mathcal{V})$ for every k and since $\mathcal{V} \subseteq \mathrm{Lie}(\mathcal{V})$, we get $\mathcal{W} \subseteq \{\eta_{0,M}, M \in \mathcal{M}\} + \mathcal{V} \subseteq \cup_k \mathcal{V}_k \subseteq \mathrm{Lie}(\mathcal{V})$.

Conclusion. By using Lemma H.2, one has for all α, $\mathrm{Lie}\left(\mathcal{D}\left(\mathcal{W}_\phi^{\mathtt{mom}}\right)\right)(\alpha) = \mathrm{Lie}\left(\mathcal{W}_\phi^{\mathtt{mom}}\right)(\alpha)$, and thus $\mathcal{W}(\alpha) = \mathcal{D}(\mathcal{W})(\alpha) = \mathrm{Lie}\left(\mathcal{D}\left(\mathcal{W}_\phi^{\mathtt{mom}}\right)\right)(\alpha) = \mathrm{Lie}\left(\mathcal{W}_\phi^{\mathtt{mom}}\right)(\alpha)$, which concludes the proof of Proposition H.3. □

Eventually, what we need to compute is the dimension of the trace $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha)$ for any $\alpha = (t, U, V, \dot{U}, \dot{V})$.

Proposition H.4. *Consider $\alpha = (t, \theta, \dot{\theta})$ such that $\left(U; V; \dot{U}; \dot{V}\right) \in \mathbb{R}^{2(n+m)\times r}$ has full rank where $\theta = \mathtt{vec}(U; V)$ and $\dot{\theta} = \mathtt{vec}(\dot{U}; \dot{V})$. Then:*

1. *if $2(n+m) \leq r$ and if $(n,m) \neq (1,1)$, then $\mathrm{dimLie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = (n+m)(2(n+m)+1)+1$;*
2. *if $2(n+m) > r$ and if $(n,m) \neq (1,1)$, then $\mathrm{dimLie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = 2(n+m)r + 1 - r(r-1)/2$;*
3. *if $(n,m) = (1,1)$ and if $r \geq 4$, then $\mathrm{dimLie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = 6+1$.*

Proof. Let us consider the linear applications:

$$\Gamma : M \in \mathcal{S} \mapsto (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} = \mathtt{vec}\left(HM \begin{pmatrix} U \\ V \\ \dot{U} \\ \dot{V} \end{pmatrix} \right) \quad \text{and} \quad \bar{\Gamma} : M \in \mathcal{S} \mapsto \begin{pmatrix} 0 \\ \Gamma(M) \end{pmatrix}.$$

By Proposition H.3, we have $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = \mathrm{span}\{\eta_{\beta,M}(\alpha) : \beta \in \mathbb{R}, M \in \mathcal{S}\}$. By linearity of $(\beta, M) \mapsto \eta_{\beta,M}(\alpha)$ it follows that $\mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = \mathbb{R}\eta_{1,0}(\alpha) + \bar{\Gamma}(\mathcal{S})$. Since the first coordinate of $\eta_{1,0}(\alpha)$ is nonzero, it does not belong to $\bar{\Gamma}(\mathcal{S})$, hence $\dim\ \mathrm{Lie}(\mathcal{W}_\phi^{\mathtt{mom}})(\alpha) = \dim(\bar{\Gamma}(\mathcal{S})) + 1 = \dim(\Gamma(\mathcal{S})) + 1 = \mathrm{rank}(\Gamma) + 1$. By the rank–nullity theorem, we have: $\dim\ \ker\ (\Gamma) + \mathrm{rank}\ (\Gamma) = \dim\ \mathcal{S}$. We now distinguish two cases.

1st case: $2(n+m) \leq r$. Then as H is invertible ($H^{-1} = -H$) and $\left(U; V; \dot{U}; \dot{V}\right)$ has full rank $2(n+m)$, Γ is injective and we obtain $\mathrm{rank}(\Gamma) = \dim\mathcal{S}$. When $(n,m) \neq (1,1)$ we have $\mathcal{S} = \mathcal{S}_{2(n+m)}$ hence this yields $\mathrm{rank}(\Gamma) = \frac{2(n+m)[2(n+m)-1]}{2} = (n+m)[2(n+m)+1]$. In the case $n = m = 1$, the assumption $2(n+m) \leq r$ reads $r \geq 4$, and the associated rank is equal to 6.

2d case: $2(n+m) > r$. Since H is invertible, $\ker(\Gamma)$ is the set of matrices $M \in \mathcal{S}$ such that $M(U;V;\dot{U};\dot{V}) = 0$. Denote M_i, $1 \leq i \leq 2(n+m)$ the rows of such a matrix, so that $M^\top = (M_1; \cdots; M_{2(n+m)})$. Denoting C_j, $1 \leq j \leq r$ the columns of $\left(U;V;\dot{U};\dot{V}\right)$. and $\mathcal{C} \coloneqq \underset{j=1,\cdots,r}{\text{span}}\ C_j$, we observe that since $\left(U;V;\dot{U};\dot{V}\right)$ has full rank $r = \min(2(m+n), r)$ the columns C_j are linearly independent and $\dim \mathcal{C} = r$. Since $M \times \left(U;V;\dot{U};\dot{V}\right) = 0$, we have $\langle M_i, C_j \rangle = 0$ for all $1 \leq i \leq 2(n+m)$ and $1 \leq j \leq r$, i.e., each $M_i \in \mathbb{R}^{2(n+m)}$ belongs to $\mathcal{C}^\perp$, of dimension $\dim \mathcal{C}^\perp = 2(n+m) - r$.

To determine $\dim\ \ker(\Gamma)$ we now count the number of degrees of freedom to choose $M \in \mathcal{S}$ such that $M_i \in \mathcal{C}^\perp$ for every i. We only treat the case $(n,m) \neq (1,1)$, where $\mathcal{S}$ is simply the set of symmetric matrices characterized by $M^\top = M$.

We first show the following lemma.

Lemma H.5. *The matrix $C \in \mathbb{R}^{2(n+m)\times r}$ has full rank r if and only if there exists a subset T of $2(n+m) - r$ indices such that the horizontal concatenation (C, Id_T) is invertible, where $\mathrm{Id}_T \in \mathbb{R}^{2(n+m)\times(2(n+m)-r)}$ is the restriction of the identity matrix to its columns indexed by T.*

Proof. The converse implication is clear. Let us show the direct one. By denoting $e_1, \cdots, e_{2(n+m)}$ the canonical basis in $\mathbb{R}^{2(n+m)}$, there is i_1 such that e_{i_1} is linearly independent from all C_j: otherwise all e_i would be spanned by $C_1, \cdots, C_r$, i.e. we would have $\text{span}\{e_i : 1 \leq i \leq 2(n+m)\} \subseteq \mathcal{C}$ hence $2(n+m) \leq r$, which contradicts our assumption. Similarly, by recursion, after finding $i_1, \cdots, i_k$ for some $k < 2(n+m) - r$ such that $i_1, \cdots, i_k$ are linearly independent from $C_1, \ldots, C_r$ (so that $\tilde{\mathcal{C}} \coloneqq \text{span}\{\mathcal{C}, e_{i_l} : 1 \leq l \leq k\}$ has dimension $r + k < 2(n+m)$), there exists i_{k+1} such that $e_{i_{k+1}}$ is linearly independent from all C_j and all $e_{i_1}, \cdots, e_{i_k}$. Stopping this construction when $k = 2(n+m) - r$ yields $T \coloneqq \{i_1, \ldots, i_k\}$. □

Consider the index set $T = \{i_1, \cdots, i_{2(n+m)-r}\}$ given by Lemma H.5, so that $(C, \mathrm{Id}_T) \in \mathbb{R}^{2(n+m)\times 2(n+m)}$ is invertible.

We first build the column M_{i_1}, which can be chosen arbitrarily in $\mathcal{C}^\perp$, a space of dimension $2(n+m) - r$. Then, the i_1-th coordinate of M_{i_2} is determined by M_{i_1} (and equal to its i_2-th one) as M is a symmetric matrix, and its remaining $2(n+m) - 1$ coordinates can be freely chosen provided that M_{i_2} belongs to $\mathcal{C}^\perp$. Thus, M_{i_2} can be arbitrarily chosen in the affine space of dimension $2(n+m) - r - 1$ defined by

$$(C, e_{i_1})^\top M_{i_2} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ M_{i_1}[i_2] \end{pmatrix},$$

where the matrix $(C, e_{i_1})^\top \in \mathbb{R}^{(r+1)\times 2(n+m)}$ has full rank $r+1$ by construction. By recursion, after building k columns $M_{i_1}, \cdots, M_{i_k}$ with $k < 2(n+m) - r$, the coordinates indexed by $i_1, \cdots, i_k$ of the column $M_{i_{k+1}}$ are determined by $M_{i_1}, \cdots, M_{i_k}$ to ensure that M is a symmetric matrix, and the remaining $2(n+m) - k$ coordinates must ensure that $M_{i_{k+1}} \in \mathcal{C}^\perp$. Thus $M_{i_{k+1}}$ can be arbitrarily chosen in the affine space of dimension $2(n+m) - r - k$ defined by

$$(C, e_{i_1}, \cdots, e_{i_k})^\top M_{i_{k+1}} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ M_{i_1}[i_{k+1}] \\ \cdots \\ M_{i_k}[i_{k+1}] \end{pmatrix},$$

where the matrix $(C, e_{i_1}, \cdots, e_{i_k})^\top \in \mathbb{R}^{(r+k)\times 2(n+m)}$ has full rank $r+k$ by construction. Finally the dimension of $\ker(\Gamma)$ is equal to:

$$\sum_{i=0}^{2(n+m)-r} (2(n+m) - r - i) = (2(n+m) - r)(2(n+m) - r + 1)/2.$$

Eventually we obtain $\text{rank}(\Gamma) = 2(n+m)r - r(r-1)/2$. □

H.3. Some derivations.

Details on how we obtained (36). Given β, Δ the matrix $M \coloneqq \begin{pmatrix} -S_\Delta & 0 \\ 0 & \beta \mathrm{Id}_{(n+m)} \end{pmatrix}$ is symmetric and belongs to $\mathcal{S}$ even when $n = m = 1$, and $HM = \begin{pmatrix} 0 & \beta \mathrm{Id}_{n+m} \\ S_\Delta & 0 \end{pmatrix}$ so that for any $\alpha = (t, \theta, \dot\theta)$ we have

$$(\mathrm{Id}_r \otimes HM)(\theta; \dot\theta) = \mathtt{vec}(HM(U; V; \dot U; \dot V)) = \mathtt{vec}((\beta(\dot U; \dot V); S_\Delta(U; V)) = (\beta\dot\theta; (\mathrm{Id}_r \otimes S_\Delta)\theta)$$

and we deduce that $\chi_{\beta,\Delta} \in \mathcal{W}$ since

$$\chi_{\beta,\Delta}(\alpha) = \begin{pmatrix} \beta \\ \beta\dot\theta \\ (\mathrm{Id}_r \otimes S_\Delta)\theta - \tau(t)\beta\dot\theta \end{pmatrix} = \begin{pmatrix} 0 \\ (\mathrm{Id}_r \otimes (HM)) \cdot \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 0 \\ -\tau(t)\dot\theta \end{pmatrix} = \eta_{\beta,M}(\alpha).$$

Details on how we obtained (37). Using that $(\mathrm{Id} \otimes \mathbf{A})(\mathrm{Id} \otimes \mathbf{B}) = \mathrm{Id} \otimes (\mathbf{AB})$ we obtain that for any α

$$\partial\eta_{0,M}(\alpha)\eta_{0,M'}(\alpha) = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_r \otimes (HM) \end{pmatrix} \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HM')] \cdot \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HMHM')] \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix}$$

$$[\eta_{0,M}, \eta_{0,M'}](\alpha) = \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HMHM' - HM'HM)] \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HM'')] \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix} = \eta_{0,M''}(\alpha)$$

with $M'' \coloneqq MHM' - M'HM$.

Details on how we obtained (38). Similarly one has

$$\partial\eta_{1,0}(\alpha)\eta_{0,M}(\alpha) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\tau'(t)\dot\theta & 0 & -\tau(t)\mathrm{Id} \end{pmatrix} \begin{pmatrix} 0 \\ [\mathrm{Id}_r \otimes (HM)] \cdot \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix} = \tau(t) \begin{pmatrix} 0 \\ \begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} (\mathrm{Id}_r \otimes (HM)) \cdot \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix}$$

$$\partial\eta_{0,M}(\alpha)\eta_{1,0}(\alpha) = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_r \otimes (HM) \end{pmatrix} \begin{pmatrix} 1 \\ \begin{pmatrix} 0 \\ -\tau(t)\dot\theta \end{pmatrix} \end{pmatrix} = -\tau(t) \begin{pmatrix} 0 \\ (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 \\ \dot\theta \end{pmatrix} \end{pmatrix}, \text{ and thus,}$$

$$[\eta_{1,0}, \eta_{0,M}](\alpha) = \tau(t) \begin{pmatrix} 0 \\ \left[\begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} (\mathrm{Id}_r \otimes (HM)) + (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id} \end{pmatrix} \right] \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} \end{pmatrix}.$$

Since $M \in \mathcal{S}$ we can write it as $M = \begin{pmatrix} S_1 & S \\ S^\top & S_2 \end{pmatrix}$ with $S \in \mathbb{R}^{(n+m)\times(n+m)}, S_1, S_2 \in \mathcal{S}_{n+m}$ (when $n = m = 1$ we further have $S, S_1, S_2 \in \mathcal{S}_2'$). We now prove that

$$\left[\begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} (\mathrm{Id}_r \otimes (HM)) + (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id} \end{pmatrix} \right] = \mathrm{Id}_r \otimes (HM') \tag{49}$$

with $M' \coloneqq \begin{pmatrix} -S_1 & 0 \\ 0 & S_2 \end{pmatrix}$. To establish (49), denoting $\mathtt{mat}(\theta) \coloneqq (U;V)$ and $\mathtt{mat}(\dot\theta) \coloneqq (\dot U;\dot V)$ we compute

$$
\begin{aligned}
HM &= \begin{pmatrix} S^\top & S_2 \\ -S_1 & -S \end{pmatrix} \\
\begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} \mathtt{vec}\left(HM \begin{pmatrix} \mathtt{mat}(\theta) \\ \mathtt{mat}(\dot\theta) \end{pmatrix} \right) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} \begin{pmatrix} \mathtt{vec}\left(S^\top \mathtt{mat}(\theta) + S_2 \mathtt{mat}(\dot\theta) \right) \\ \mathtt{vec}\left(-S_1 \mathtt{mat}(\theta) - S \mathtt{mat}(\dot\theta) \right) \end{pmatrix} \quad = \begin{pmatrix} 0 \\ \mathtt{vec}\left(S_1 \mathtt{mat}(\theta) + S \mathtt{mat}(\dot\theta) \right) \end{pmatrix} \\
(\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id} \end{pmatrix} \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} &= (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 \\ \dot\theta \end{pmatrix} = \mathtt{vec}\left(HM \begin{pmatrix} 0 \\ \mathtt{mat}(\dot\theta) \end{pmatrix} \right) \qquad = \begin{pmatrix} \mathtt{vec}\left(S_2 \mathtt{mat}(\dot\theta) \right) \\ \mathtt{vec}\left(-S \mathtt{mat}(\dot\theta) \right) \end{pmatrix}
\end{aligned}
$$

so that

$$
\begin{aligned}
\left[\begin{pmatrix} 0 & 0 \\ 0 & -\mathrm{Id} \end{pmatrix} (\mathrm{Id}_r \otimes (HM)) + (\mathrm{Id}_r \otimes (HM)) \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id} \end{pmatrix} \right] \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix} &= \begin{pmatrix} \mathtt{vec}\left(S_2 \mathtt{mat}(\dot\theta) \right) \\ \mathtt{vec}\left(S_1 \mathtt{mat}(\theta) \right) \end{pmatrix} \\
&= \mathtt{vec}\left(\underbrace{\begin{pmatrix} 0 & S_2 \\ S_1 & 0 \end{pmatrix}}_{= HM'} \begin{pmatrix} \mathtt{mat}(\theta) \\ \mathtt{mat}(\dot\theta) \end{pmatrix} \right) = (\mathrm{Id}_r \otimes (HM')) \begin{pmatrix} \theta \\ \dot\theta \end{pmatrix}.
\end{aligned}
$$

I. Proof of Proposition 4.4

Proposition 4.4. *Consider $g(\theta, x) \coloneqq UV^\top x$ with $(U,V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. Assume both $(U;V) \in \mathbb{R}^{(n+m)\times r}$ and $(U;V;\dot U;\dot V) \in \mathbb{R}^{2(n+m)\times r}$ have full rank and $(n,m) \neq (1,1)$. Denote N_{GF} (resp. N_{MF}) the number of independent conservation laws of $\phi_{\mathtt{Lin}}$ for (2) (resp. for (3)) in a neighborhood of (U,V) (resp. of $(t,U,V,\dot U,\dot V)$, then if $r \leq n+m$: $N_{GF} - N_{MF} = r > 0$, else: $N_{GF} - N_{MF} \leq 0$.*

Proof. Denoting $\mathtt{rk}_1 = \min(r, n+m)$ the rank of $(U;V) \in \mathbb{R}^{(n+m)\times r}$ and $\mathtt{rk}_2 = \min(r, 2(n+m))$ the rank of $(U;V;\dot U;\dot V) \in \mathbb{R}^{2(n+m)\times r}$ and using Proposition 3.8 and Theorem 4.3, one has

$$
N_{GF} \coloneqq \mathtt{rk}_1/2(2r+1-\mathtt{rk}_1) \text{ and } N_{MF} \coloneqq \mathtt{rk}_2/2(2r-1-\mathtt{rk}_2).
$$

Thus:

$$
\begin{aligned}
N_{GF} - N_{MF} &= \frac{\mathtt{rk}_1}{2}(2r+1-\mathtt{rk}_1) - \frac{\mathtt{rk}_2}{2}(2r-1-\mathtt{rk}_2) \\
&= -\frac{\mathtt{rk}_1^2}{2} + \frac{\mathtt{rk}_2^2}{2} + r(\mathtt{rk}_1 - \mathtt{rk}_2) + \frac{\mathtt{rk}_1 + \mathtt{rk}_2}{2} \\
&= \tfrac{1}{2}[(\mathtt{rk}_2 - r)^2 - (\mathtt{rk}_1 - r)^2] + \frac{\mathtt{rk}_1 + \mathtt{rk}_2}{2}.
\end{aligned}
$$

We now distinguish 3 cases.

1st case: $r \leq n+m$. Then $\mathtt{rk}_1 = \mathtt{rk}_2 = r$, and thus $N_{GF} - N_{MF} = \frac{\mathtt{rk}_1 + \mathtt{rk}_2}{2} = r > 0$.

2d case: $n+m < r \leq 2(n+m)$. Then $\mathtt{rk}_1 = n+m$ and $\mathtt{rk}_2 = r$, and thus: $N_{GF} - N_{MF} = \frac{-(n+m-r)^2}{2} + \frac{r+n+m}{2}$. Let us show that in that case, we always have: $N_{GF} - N_{MF} \leq 0$. Denoting $x \coloneqq n+m \geq 1$ and $y \coloneqq r \geq 2$, we have: $N_{GF} - N_{MF} = -(x-y)^2/2 + (x+y)/2 = -\frac{1}{2}[x^2 - (1+2y)x + y(y-1)] = -\frac{1}{2}(x-x_1)(x-x_2)$, where: $x_1x_2 = y(y-1)$ (and thus x_1 and x_2 have the same sign) and $x_1 + x_2 = 1+2y$ (and thus they are positive). By using the equality $x_1x_2 = y(y-1)$, we necessarily have $x_1, x_2 \in [y-1, y]$. Thus as $x \leq y-1$ ($n,m,r \in \mathbb{N}$), we then have $N_{GF} - N_{MF} \leq 0$.

3rd case: $2(n+m) < r$. Then $\mathtt{rk}_1 = n+m$ and $\mathtt{rk}_2 = 2(n+m)$, and thus:

$$\begin{aligned} N_{GF} - N_{MF} &= -\frac{\mathtt{rk}_1^2}{2} + \frac{\mathtt{rk}_2^2}{2} + r(\mathtt{rk}_1 - \mathtt{rk}_2) + \frac{\mathtt{rk}_1 + \mathtt{rk}_2}{2} \\ &= (n+m)\left[\tfrac{3}{2}(n+m) + \tfrac{3}{2} - r\right] \\ &\leq 0 \text{ as } 2(n+m)+1 \leq r. \end{aligned}$$

□

J. Proof of Theorem 4.5

Theorem 4.5. *Consider $g(\theta, x) := UV^\top x$ with $\theta = (U, V) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$. Then*

$$1_n^\top U - 1_m^\top V \tag{20}$$

defines r independent conservation laws for (2) with $M(\theta) := \mathrm{diag}(\theta)$.

Proof. One has: $\frac{\mathrm{d}}{\mathrm{d}t}\left(1_n^\top U - 1_m^\top V\right) = 1_n^\top \dot{U} - 1_m^\top \dot{V} \overset{(2)}{=} -1_n^\top\Big(U \odot \underbrace{\nabla_U \mathcal{E}_Z}_{\in\mathbb{R}^{n\times r}}\Big) + 1_m^\top\Big(V \odot \underbrace{\nabla_V \mathcal{E}_Z}_{\in\mathbb{R}^{n\times r}}\Big)$. Then by (27) one has:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(1_n^\top U - 1_m^\top V\right) = -1_n^\top\Big(U \odot (\nabla F(\phi(\theta))V)\Big) + 1_m^\top\Big(V \odot \big(\nabla F(\phi(\theta))^\top U\big)\Big).$$

Then by denoting $\nabla F(\phi(\theta)) = (A_{i,j})_{i,j} \in \mathbb{R}^{n\times m}$, one has for all $j = 1, \cdots, r$:

$$\begin{aligned} \Big[1_n^\top\Big(U \odot (\nabla F(\phi(\theta))V)\Big) - 1_m^\top\Big(V \odot \big(\nabla F(\phi(\theta))^\top U\big)\Big)\Big][j] &= \sum_{i=1}^n U_{i,j} \sum_{k=1}^m A_{i,k} V_{k,j} - \sum_{i=1}^m V_{i,j} \sum_{k=1}^n A_{k,i} U_{k,j} \\ &= 0. \end{aligned}$$

This proves as claimed that for $1 \leq j \leq r$ the j-th column of $1_n^\top U - 1_m^\top V$ defines a conserved function $h_j(\theta)$ for (2) with $M(\theta)$. Since h_j only depends on the corresponding columns u_j, v_j of U, V, the gradients $\nabla h_j(\theta)$, $\nabla h_\ell)(\theta)$ are orthogonal, $j \neq \ell$, hence these conserved functions are also independent. □

K. Proof of Theorem 4.6

Theorem 4.6. *Consider $g(\theta, x) = U\sigma(V^\top x + b^\top)$. Denote U_j (resp. V_j / b_j) the j-th column of U (resp column of V / entry of b). For all $j = 1, \cdots, r$, the function*

$$1_n^\top U_j - \frac{1}{2}\left(\|V_j\|^2 + b_j^2\right) \tag{21}$$

is a conservation law for (2) with $M(U, V, b) := \mathrm{diag}[(U, 1_{m\times r}, 1_{1\times r})]$.

Proof. Consider the following extension of the linear transformation T^A from (15) to cover the presence of biases, where by convention $U \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{m\times r}$, and $b \in \mathbb{R}^{1\times r}$:

$$T^A(\epsilon, U, V, b) := \Big(U\exp(\epsilon A), V\exp(-\epsilon A^\top), b\exp(-\epsilon A^\top)\Big).$$

Observe that as soon as $A \in \mathbb{R}^{r\times r}$ is diagonal, T^A is a linear transformation that leaves $g(\cdot, x)$ invariant for each x, hence it also leaves $\mathcal{E}_Z$ invariant. Moreover, $\Delta_{T^A}(U, V, b) = (UA, -VA^\top, -bA^\top) = (UA, -VA, -bA)$, as diagonal matrices are symmetric. Thus by (14):

$$\left\langle \nabla\mathcal{E}_Z(\theta), \left(\begin{smallmatrix} UA \\ -VA \\ -bA \end{smallmatrix}\right)\right\rangle = 0.$$

Specializing to $A = E_{i,i} \in \mathbb{R}^{r\times r}$, the one-hot matrix with the (i,i)-th entry being 1, we obtain

$$\langle \nabla_{U_i}\mathcal{E}_Z(\theta), U_i\rangle - \langle \nabla_{V_i}\mathcal{E}_Z(\theta), V_i\rangle - \nabla_{b_i}\mathcal{E}_Z(\theta) b_i = 0. \tag{50}$$

with U_i, V_i the columns of U, V. Finally, given any $j \in \{1, \dots r\}$ we compute using that $M(\theta) = \mathrm{diag}[(U, 1_{n\times r}, 1_{1\times r})]$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(1_n^\top U_j - \frac{1}{2}\left(\|V_j\|^2 + b_j^2\right)\right) = 1_n^\top \dot{U}_j - \left(\langle \dot{V}_j, V_j\rangle + b_j \dot{b}_j\right)$$
$$\overset{(2)}{=} -\underbrace{1_n^\top\left(U_j \odot \nabla_{U_j}\mathcal{E}_Z(\theta)\right)}_{=\langle \nabla_{U_j}\mathcal{E}_Z(\theta), U_j\rangle} + \langle \nabla_{V_j}\mathcal{E}_Z(\theta), V_j\rangle + b_j \nabla_{b_j}\mathcal{E}_Z(\theta) \overset{(50)}{=} 0. \qquad \square$$

L. About the Natural gradient flow case (cf Section 4.5).

Let $\theta \in \Theta$. We consider $Z = (x_i, y_i)_i$ such that $g(\cdot, x_i)$ is $\mathcal{C}^2$ in a neighborhood of θ, and the ODE:

$$\dot{\theta} = \underbrace{\left(\frac{1}{n}\sum_i [\partial_1 g(\theta, x_i)]^\top \partial_1 g(\theta, x_i)\right)^\dagger}_{=: M_Z(\theta)} \nabla \mathcal{E}_Z(\theta), \tag{51}$$

where $A^\dagger$ denotes the pseudo-inverse of A. We consider ϕ as in Assumption 2.13. Then we deduce from (6) that $\partial_1 g(\theta, x_i) = \partial f_1(\phi(\theta), x_i)\partial\phi(\theta)$ for each i, so that:

$$M_z(\theta) = \left(\frac{1}{n}\sum_i [\partial_1 g(\theta, x_i)]^\top \partial_1 g(\theta, x_i)\right)^\dagger = \left(\partial\phi(\theta)^\top \frac{1}{n}\sum_i \partial_1 f(\theta, x_i)^\top \partial_1 f(\theta, x_i)\partial\phi(\theta)\right)^\dagger.$$

As $A^\dagger = \lim_{\delta\to 0} A^\top(AA^\top + \delta I)^{-1}$, using the definitions of $W_\theta^{\mathtt{grad}}$ and $\mathcal{W}_\phi(\theta)$ (cf Proposition 2.12 and (7)) we get $W_\theta^{\mathtt{grad}} := \mathrm{span}_{Z\in\mathcal{Z}'_\theta}\{M_Z(\theta)\nabla\mathcal{E}_Z(\theta)\} \subseteq \mathrm{span}\{\nabla\phi_1(\theta), \cdots, \nabla\phi_d(\theta)\} =: \mathcal{W}_\phi(\theta)$. By Definition 2.19 and (9), any GF conservation law h of ϕ for the Euclidean gradient flow (2) (i.e., with $M = \mathrm{Id}$) satisfies $\nabla h(\theta) \perp \mathcal{W}_\phi(\theta)$ for all $\theta \in \Theta$, hence $\nabla h(\theta) \perp W_\theta^{\mathtt{grad}}$ for every $\theta \in \Theta$. By Proposition 2.12, this shows that h is indeed locally conserved on Θ for any data set *through the natural gradient flow* (2) with $M_Z(\theta)$.

M. About Experiments with formal calculus

Our code is open-sourced and is available at `https://github.com/sibyllema/Conservation_laws_ICML`. We used the software SageMath (The Sage Developers, 2022), which relies on a Python interface. We compare the number of independent conservation laws given Theorem 3.7, with the number of independent polynomial conservation laws found (with or without a change of variable) as explained in Section 3.1.

M.1. Euclidean MF

We first considered the case of Euclidean MF (the Euclidean GF has been studied in (Marcotte et al., 2024)) from Sections 4.1 and 4.2. We tested both linear and ReLU neural networks (with and without biases) of various depths and widths, and observed that the two numbers matched in all our examples, with a constant τ either equal to 0 or 1. For this, we drew 20 random linear (resp. ReLU) neural networks, with depth drawn uniformly at random between 2 to 4 and i.i.d. layer widths drawn uniformly at random between 1 to 4, with a τ randomly chosen between 0 or 1. For ReLU architectures, the probability of including biases was $1/2$. Then we checked that the two numbers match. In particular, for ReLU neural networks, the number of conservation laws is always equal to zero.

M.2. NMF

For the GF scenario (resp. the MF scenario) with NMF from Section 4.3, we drew 20 random 2-layer linear neural networks, with i.i.d. layer widths drawn uniformly at random between 1 to 8 (resp. 1 and 6, with a τ randomly chosen between 0 or 1). We observed that the two numbers (number of independent conservation laws/number of independent "polynomial" conservation laws) matched in all our examples in the GF scenario and that there is no conservation law for the MF scenario.

M.3. ICNN

For the GF scenario (resp. the MF scenario), we draw 20 random 2-layer ReLU neural networks (with and without biases), with i.i.d. layer widths drawn uniformly at random between 1 to 6 (resp. and with a τ randomly chosen between 0 or 1),

with a probability of including biases of $1/2$. We observed that the two numbers (number of independent conservation laws/number of independent "polynomial" conservation laws) matched in all our examples in the GF scenario, and that there is no conservation law for the MF scenario.

N. Numerical Simulation

In this section, we show numerical simulations on gradient flows and momentum flows to explore: (a) the influence of the time discretization on the preserved quantities, (b) the impact of momentum on the preservation of conservation laws for the gradient flows. Our code is open-sourced and is available at https://github.com/sibyllema/Conservation_laws_ICML.

N.1. Discretization of the flows

To ease the description, we consider the following parameterization of the flows

$$\mu\ddot{\theta} + \nu\dot{\theta} = -M(\mu\dot{\theta} + \nu\theta)\nabla\mathcal{E}_Z(\theta).$$

Gradient flows correspond to $\mu = 0$, while the momentum parameter is $\tau = \nu/\mu$. We consider the following time discretization of the flows, where time at step k is $t = k\delta$ and $\delta > 0$ is the time step

$$\mu\frac{\theta_{k+1} + \theta_{k-1} - 2\theta_k}{\delta^2} + \nu\frac{\theta_{k+1} - \theta_k}{\delta} = -M_k\nabla\mathcal{E}_Z(\theta_k) \quad \text{where} \quad M_k := M\Big(\mu\frac{\theta_k - \theta_{k-1}}{\delta} + \nu\theta_k\Big)$$

This can be re-written in the usual form of a gradient descent with momentum

$$\theta_{k+1} = \theta_k - \alpha M_k\nabla\mathcal{E}_Z(\theta_k) + \beta(\theta_k - \theta_{k-1})$$

where

$$\alpha := \frac{\delta}{\nu + \mu/\delta} \quad \text{and} \quad \beta := \frac{\mu}{\delta\nu + \mu} < 1.$$

Here $\beta \in [0, 1)$ is the momentum (extrapolation) parameter, so that $\beta = 0$ corresponds to usual gradient descent, and setting $\beta = 1$ is maximum momentum (which is not in general ensured to converge).

N.2. MLP example

We consider here a 3-layer MLP trained for classification on the MNIST dataset (LeCun et al., 2010) with the cross entropy loss function and a ReLU non-linearity. The input dimension is $28 \times 28 = 784$ (number of pixels), the inner layer dimensions are $(512, 256)$ and the output dimension is 10 (number of classes). We focus on the conservation laws associated to the $r = 512$ neurons of the first two layers, and we denote $(U \in \mathbb{R}^{784\times512}, V \in \mathbb{R}^{256\times512})$ the associated matrices, with associated columns neurons $(u_i \in \mathbb{R}^{784})_i$ and $(v_i \in \mathbb{R}^{256})_i$. The r conserved quantities for the gradient flow, $\tau = +\infty$ are $\|u_i\|^2 - \|v_i\|^2$ and there is no *exactly* preserved quantity for the momentum flow $\tau < +\infty$.

Figure 1, left, shows the evolution of the loss for a range of step size δ up to almost no convergence, all with the same initialization of the weights. Note that despite the non-convexity of the loss function, the evolution converges to approximately the same loss value. Figure 1, right, shows one of the conservation laws (associated with the neurons of the first layer). One can see that even for relatively large step sizes, these quantities are almost perfectly conserved. It is only for step size on the edge of instabilities that these quantities are not well preserved. This validates the relevance of these conservation laws for the regime of the step size used for stable training of neural networks. Figure 2 shows how the evolution of the loss and the preserved quantities for GF is impacted by the momentum parameter $\mu = 1/\tau$. As expected, increasing μ deteriorates the preservation of the conservation law.

N.3. NMF example

We consider here a non-negative matrix factorization so that the loss function is $\mathcal{E}_Z(U, V) = \|UV^\top - Y\|^2$ under positivity constraints. We thus use the metric of the mirror flow associated to the Shannon potential function $M_Z(t, \theta, \dot{\theta}) = \mathrm{diag}(\dot{\theta}+\tau\theta)$ The column of $Y \in \mathbb{R}^{n\times p}$ are $p = 6903$ images of $n = 28 \times 28$ pixels from both the test and training sets of MNIST dataset (LeCun et al., 2010) associated to the digit 0, see Fig. 3, left. We compute a factorization of rank $r = 10$ so that

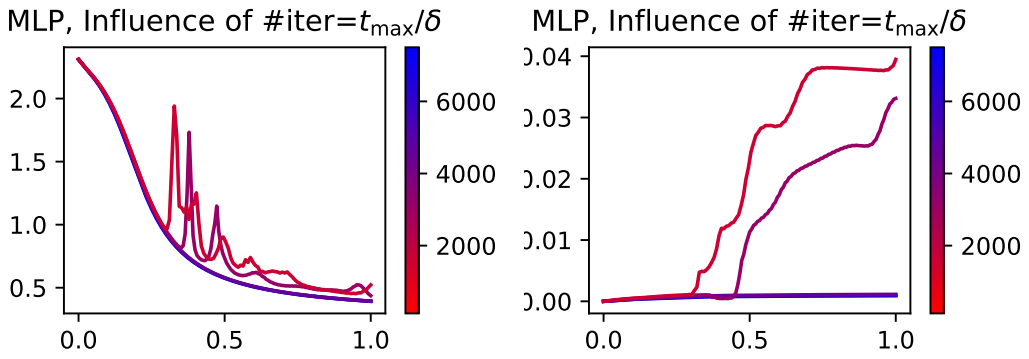

Figure 1. Impact on the step size δ on the evolution of the loss (left) and on the preservation of one of the conservation laws (right). The colors are associated with the number of iterations $t_{\max}/\delta$ used to train the networks.

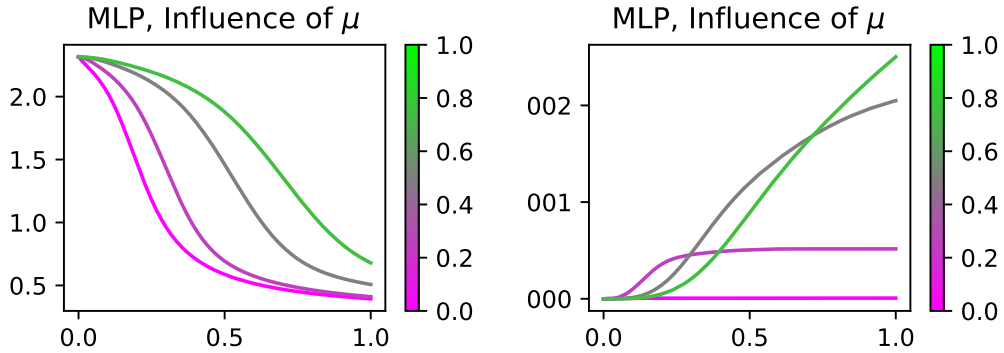

Figure 2. Impact on the momentum parameter $\mu = 1/\tau$ on the evolution of the loss (left) and on the preservation of one of the conservation laws (right).

$U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{p\times r}$. Figure 3 shows examples of the r factors (columns of U displayed as positive images). Note that while the function is non-convex, in practice, gradient descent and momentum descent converge to global minimizers (and loss curves converge to the same values), as shown on the left of Figure 4. We use a small step size to avoid discretization error (which impact is similar to the one reported in the previous section). For $\tau = +\infty$ the r conservation laws are $1^\top U - 1^\top V$ and Figure 4 displays the evolution in time of the first of the quantities (associated with the first factor). As it is expected when $\mu = 0$ (gradient flow) this law is perfectly conserved and is only approximately preserved for larger value of the momentum parameter μ. Note however that an interesting phenomenon arises, that similarly to the MLP case, these quantities stay bounded for all time within a range depending on the momentum parameter μ (so if μ is small, approximate conservation holds for all time). Analyzing theoretically this non-trivial phenomenon is an interesting avenue for future work.

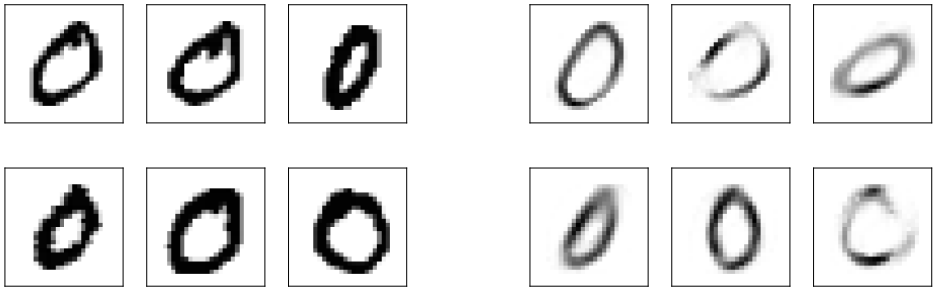

Figure 3. Left: example of input images (columns of Y). Right: example of NMF factors (columns of U) at optimality.

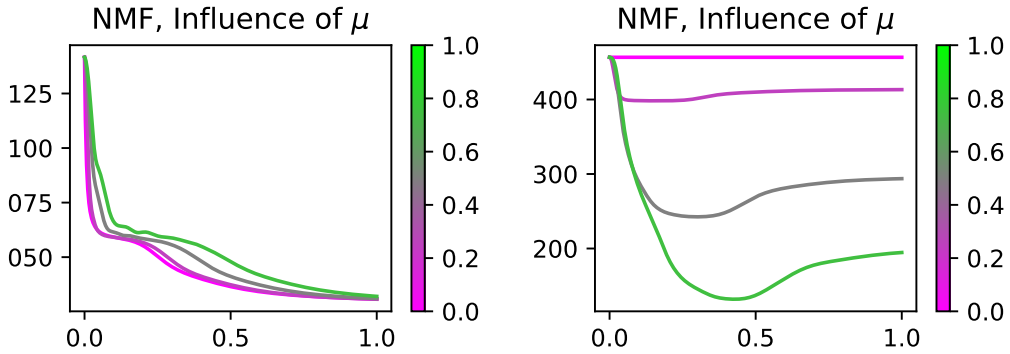

Figure 4. Impact on the momentum parameter $\mu = 1/\tau$ on the evolution of the loss (left) and on the preservation of one of the conservation laws (right). Here $\mu = 0$ corresponds to the gradient flow (no momentum).