AN EQUIVARIANCE TOOLBOX FOR LEARNING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many theoretical results in deep learning can be traced to symmetry or equivariance of neural networks under parameter transformations. However, existing analyses are typically problem-specific and focus on first-order consequences such as conservation laws, while the implications for second-order structure remain less understood. We develop a general equivariance toolbox that yields coupled first- and second-order constraints on learning dynamics. The framework extends classical Noether-type analyses in three directions: from gradient constraints to Hessian constraints, from symmetry to general equivariance, and from continuous to discrete transformations. At the first order, our framework unifies conservation laws and implicit-bias relations as special cases of a single identity. At the second order, it provides structural predictions about curvature: which directions are flat or sharp, how the gradient aligns with Hessian eigenspaces, and how the loss landscape geometry reflects the underlying transformation structure. We illustrate the framework through several applications, recovering known results while also deriving new characterizations that connect transformation structure to modern empirical observations about optimization geometry.

1 INTRODUCTION

A large body of recent theory in deep learning exploits, either explicitly or implicitly, *symmetry* or *equivariance* of neural networks with respect to parameter transformations (Kunin et al., 2020; Głuch & Urbanke, 2021; Lyu & Li, 2019; Min et al., 2023; Tarmoun et al., 2021). Such structures arise naturally in many settings: layer-wise rescaling in ReLU networks, reparameterizations in deep linear models, and various invariances induced by over-parameterization. While powerful, existing analyses are typically developed in a problem-specific manner: one identifies a particular transformation, derives a tailored identity for a specific architecture or loss, and builds a dedicated argument around that identity. As a result, the *common mechanism* underlying these diverse results remains obscured, and extensions beyond specific first-order identities are often unclear.

A related line of work applies Noether's theorem to systematize symmetry-based arguments (Głuch & Urbanke, 2021; Zhao et al., 2022; Kunin et al., 2020). This perspective elegantly derives conservation laws from continuous symmetries, but it is intrinsically a *first-order* tool: it constrains the gradient but says little about curvature. Yet many actively studied phenomena concern *second-order* structure such as the Hessian spectrum, the geometry of sharp and flat directions, and the low-dimensional subspaces that dominate optimization (Sagun et al., 2016a; Cohen et al., 2021; Gur-Ari et al., 2018; Foret et al., 2020). Although Kunin et al. (2020) derived some second-order identities for certain symmetries, these were not developed into a general characterization, and the scope was limited to continuous symmetry rather than general equivariance or discrete symmetry.

In this work, we develop a unified *equivariance toolbox* for learning dynamics that yields coupled first- and second-order constraints. Our framework applies whenever a model f satisfies an equivariance relation: transforming parameters by an operation H is equivalent to transforming outputs by an operation G. From this relation, we derive universal identities that constrain both the gradient and Hessian of the loss along transformation-induced directions. Importantly, this framework can handle both continuous and discrete equivariances (symmetries). When the equivariance is continuous (i.e., continuously parameterized), these identities hold throughout parameter space; while when

it is discrete (e.g., sign flips or permutations), weaker but still informative constraints hold on the fixed-point sets of the transformation.

At the first order, our framework recovers a broad family of known results as special cases. For symmetries (where G is the identity), it reduces to orthogonality between the gradient and symmetry-generated directions, recovering Noether-type conservation laws. For non-trivial equivariances such as homogeneity, it reproduces key scalar relations that underpin classical implicit-bias arguments (Lyu & Li, 2019). Both cases are handled by the same theoretical framework. At the second order, our framework provides characterizations that are largely missing in existing work. The derived identities constrain how the Hessian acts along transformation-induced directions, yielding concrete structural predictions: relationships between the gradient and the Hessian's column space, sharpness estimates along canonical directions, and constraints on eigenspace structure. These predictions connect directly to empirical observations about optimization geometry and late-stage training behaviors (Gur-Ari et al., 2018; Cohen et al., 2021; Papyan et al., 2020; Kothapalli, 2022).

Organization. Due to space limitations, we only present the major theorem as well as the necessary preliminaries in the main paper, leaving other discussions in the appendix. Section 2 introduces tensor notations used throughout. Section 3 presents the general framework and states the main identities for both continuous equivariance and discrete symmetry. In the appendix, Section A, Section B, and Section C discuss the consequences of the first-order identities, second-order identities, and discrete identities, recovering and developing important theoretical consequences such as conservation laws, homogeneity-based relations, and mirror-symmetry results as a special case. Section D concludes this paper and discuss some of the limitations and important future directions.

2 Preliminaries: Tensor Operations

Throughout this paper we will frequently manipulate higher-order derivatives and compositions of tensor-valued maps. To avoid ambiguity, we fix a notation system for tensors and tensor functions. Throughout, we view a tensor as a (curried) linear mapping: it takes one vector as input and outputs another tensor. This convention is consistent with the treatment in modern formal language systems such as Lean 4 (Moura & Ullrich, 2021).

Formally, a **tensor shape** is a finite sequence of positive integers. For shapes s and t, we write (s, t) for their concatenation. For an integer n, we write (n) for the shape consisting of the single element n (and we may omit parentheses when there is no ambiguity). For a shape s, we denote by $\mathbb{T}(s)$ the set of all tensors of shape s, defined recursively:

1. A linear mapping $\mathbb{R}^n \to \mathbb{R}$ is a **tensor** of shape (n).
2. Given a shape s, a linear mapping $\mathbb{R}^n \to \mathbb{T}(s)$ is a **tensor** of shape (n, s).

For a shape s, a positive integer n, and a (sufficiently smooth) function $f : \mathbb{R}^n \to \mathbb{T}(s)$, then the Fréchet derivative (gradient) of f is denoted as $\nabla f : \mathbb{R}^n \to \mathbb{T}(n, s)$.

For functions f and g, we use $g \circ f$ to denote their (usual) composition whenever it is well-defined. We also define a more general notion of composition, the **k-th component composition** $\circ_k$, which is convenient under the curried representation of tensors. We define it recursively as follows:

1. The first component composition coincides with standard composition: $\circ_1 := \circ$.
2. For $k \geq 2$, define $(g \circ_k f)(x) := g(x) \circ_{k-1} f$ whenever the right-hand side is well-defined.

Intuitively, $\circ_k$ composes f into the k-th input component of g when one interprets curried tensors as functions of multiple arguments. For example, if $g \in \mathbb{T}(a, b, c)$ and $f \in \mathbb{T}(d, b)$, then $g \circ_2 f \in \mathbb{T}(a, d, c)$ and $(g \circ_2 f)(x, y, z) = g(x, f(y), z)$.

Example. Readers familiar with matrix operations may find the following correspondence helpful. A tensor of shape (n, m) can be viewed with an $n \times m$ matrix; a tensor of shape (n) can be viewed as an n-dimensional vector. For example, if $A \in \mathbb{T}(n, m)$ and B is another tensor (can be a matrix or a vector, as long as the dimension matches), the composition $A \circ B$ corresponds to matrix multiplication AB, and the second-component composition $A \circ_2 B$ corresponds to $B^\top A$. Specifically, if $x \in \mathbb{T}(n)$ and $y \in \mathbb{T}(m)$ are vectors, we have $A \circ x \circ_2 y = y^\top A x = \langle y, Ax \rangle$.

3 Main Results: A General Equivariance Framework

In this section we present the main theoretical results: universal first- and second-order identities induced by equivariance. We consider two cases separately. When the equivariance is continuous, the

identities hold at every point in parameter space. When the transformation is not smoothly parameterized but admits fixed points, weaker constraints still hold on the fixed-point set. All statements use the tensor notations from Section 2; concrete instantiations appear in Sections A to C.

Throughout, we consider a model $f : \mathbb{R}^d \to \mathbb{R}^c$, representing a neural network as a function of its d trainable parameters with c-dimensional output (input data can be absorbed into the definition of f). We also consider a smooth loss function $\ell : \mathbb{R}^c \to \mathbb{R}$, so that the overall objective is $L := \ell \circ f : \mathbb{R}^d \to \mathbb{R}$.

We introduce two families of transformations. The first, $H : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$, acts on parameters and is parameterized by a p-dimensional vector $\boldsymbol{\lambda} \in \mathbb{R}^p$. We write $\nabla_\theta H$ and $\nabla_\lambda H$ for the Fréchet derivatives (gradients) of H with respect to its first and second arguments, respectively. The second, $G : \mathbb{R}^c \times \mathbb{R}^p \to \mathbb{R}^c$, acts on outputs, with gradients $\nabla_y G$ and $\nabla_\lambda G$.

Definition 1 (Equivariance and good position)**.** *Assume f, $H(\cdot, \boldsymbol{\lambda})$, and $G(\cdot, \boldsymbol{\lambda})$ are smooth for every $\boldsymbol{\lambda} \in \mathbb{R}^p$. We say that f admits an (H, G)**-equivariance** if*

$$f\big(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\big) = G\big(f(\boldsymbol{\theta}), \boldsymbol{\lambda}\big) \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^d \text{ and } \boldsymbol{\lambda} \in \mathbb{R}^p. \tag{1}$$

*If additionally H and G are both smooth in $\boldsymbol{\lambda}$, we say this equivariance is **continuous**. A pair $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is called a **good position** if both $\nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})$ are invertible.*

An important special case of equivariance is *symmetry* (or *invariance*), where $G(\cdot, \boldsymbol{\lambda}) = \mathrm{Id}$ is the identity mapping. In some physics-related contexts, the set of all symmetries is required to form a group (Zhao et al., 2022; 2025; Kunin et al., 2020). While this is often the case in practice, our framework does not require any algebraic structure on the symmetric or equivariant transformations. We only require smoothness near the good position, which is a purely analytic condition.

Below, we present two results addressing continuous and discrete equivariances separately. We emphasize that both theorems arise from the same equivariance condition and share a common derivation strategy (see Section E.1 for details).

3.1 Continuous Equivariance

We first present results for continuous equivariances, where the transformation parameter $\boldsymbol{\lambda}$ varies continuously.

Definition 2 (Characteristic direction and characteristic output)**.** *Suppose f admits a continuous (H, G)-equivariance and $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a good position. Define*

$$\begin{cases} X(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}), \\ Y(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}). \end{cases}$$

*We call X the **characteristic direction** in parameter space and Y the **characteristic output** in output space.*

Informally, $X(\boldsymbol{\theta}, \boldsymbol{\lambda})$ captures how an infinitesimal change in $\boldsymbol{\lambda}$ induces a direction in parameter space after "solving for" the corresponding change through H; $Y(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the analogous quantity on the output side through G. With these definitions in place, we state the main result for continuous equivariances.

Theorem 1 (First- and second-order identities)**.** *Assume f admits a continuous (H, G)-equivariance and $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a good position. Then the following hold:*

1. ***(First-order identity).*** *The gradient of L satisfies*
   (i). $\nabla L(\boldsymbol{\theta}) \circ X(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ Y(\boldsymbol{\theta}, \boldsymbol{\lambda})$.

2. ***(Second-order identities).*** *The Hessian of L satisfies the following two equations:*
   (ii).
$$\begin{aligned} &\nabla^2 L(\boldsymbol{\theta}) \circ X(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &= \nabla^2 \ell\big(f(\boldsymbol{\theta})\big) \circ Y(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_2 \nabla f(\boldsymbol{\theta}) \\ &\quad - \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &\quad + \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ X(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &\quad + \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ_2 \nabla f(\boldsymbol{\theta}) \\ &\quad - \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ Y(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_2 \nabla f(\boldsymbol{\theta}). \end{aligned}$$

(iii). $\nabla^2 L(\boldsymbol{\theta}) \circ X(\boldsymbol{\theta},\boldsymbol{\lambda}) \circ_2 X(\boldsymbol{\theta},\boldsymbol{\lambda})$

$$
\begin{aligned}
&= \nabla^2 \ell\big(f(\boldsymbol{\theta})\big) \circ Y(\boldsymbol{\theta},\boldsymbol{\lambda}) \circ_2 Y(\boldsymbol{\theta},\boldsymbol{\lambda}) \\
&\quad - \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta},\boldsymbol{\lambda})^{-1} \circ \nabla^2_\lambda H(\boldsymbol{\theta},\boldsymbol{\lambda}) \\
&\quad + \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta},\boldsymbol{\lambda})^{-1} \circ \nabla^2_\theta H(\boldsymbol{\theta},\boldsymbol{\lambda}) \circ X(\boldsymbol{\theta},\boldsymbol{\lambda}) \circ_2 X(\boldsymbol{\theta},\boldsymbol{\lambda}) \\
&\quad + \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla_y G(f(\boldsymbol{\theta}),\boldsymbol{\lambda})^{-1} \circ \nabla^2_\lambda G(f(\boldsymbol{\theta}),\boldsymbol{\lambda}) \\
&\quad - \nabla \ell\big(f(\boldsymbol{\theta})\big) \circ \nabla_y G(f(\boldsymbol{\theta}),\boldsymbol{\lambda})^{-1} \circ \nabla^2_y G(f(\boldsymbol{\theta}),\boldsymbol{\lambda}) \circ Y(\boldsymbol{\theta},\boldsymbol{\lambda}) \circ_2 Y(\boldsymbol{\theta},\boldsymbol{\lambda}).
\end{aligned}
$$

Although the formulas in Theorem 1 appear complicated, they admit a simple interpretation and often simplify substantially. Eq. (i) states that the directional derivative of L along the characteristic direction X is determined entirely by output-side quantities: the loss gradient $\nabla \ell(f(\boldsymbol{\theta}))$ and the characteristic output Y, which are both zeroth-order properties of f. The second-order identities eqs. (ii) and (iii) provide analogous constraints for the Hessian, expressing its action and quadratic form along X in terms of lower-order information.

Two common special cases lead to significant simplification. First, when the equivariance is a *symmetry*, we have $\nabla_\lambda G \equiv 0$ and hence $Y \equiv 0$; all terms shown in green vanish. Second, when $H(\boldsymbol{\theta},\boldsymbol{\lambda})$ is *linear* in $\boldsymbol{\theta}$, we have $\nabla^2_\theta H = 0$; all terms shown in blue vanish. Many equivariances of practical interest satisfy one or both conditions, yielding compact and interpretable identities.

3.2 Discrete Symmetry

When the transformation H is not continuously parameterized in $\boldsymbol{\lambda}$ (for instance, when $\boldsymbol{\lambda}$ indexes a discrete set of transformations such as sign flips or permutations), the continuous identities of Theorem 1 do not directly apply. Nevertheless, the equivariance condition still implies useful constraints when we restrict attention to the fixed-point set of a given transformation. This approach is common in the study of discrete symmetries (Brea et al., 2019; Ziyin, 2023). We first provide a formal definition of the fixed-point set.

Definition 3 (Fixed-point set)**.** *For $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$, we say that $\mathcal{S} \subseteq \mathbb{R}^d$ is a **fixed-point set** of H at $\boldsymbol{\lambda}_0$ if $H(\boldsymbol{\theta},\boldsymbol{\lambda}_0) = \boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathcal{S}$. We denote the union of all fixed-point sets of H at $\boldsymbol{\lambda}_0$ by $\mathrm{Fix}(H,\boldsymbol{\lambda}_0)$.*

In this subsection, we focus on *symmetry*, i.e., $G = \mathrm{Id}$. In most cases of discrete symmetry, $H(\cdot,\boldsymbol{\lambda}_0)$ is a linear transformation which only has eigenvalues ± 1 (e.g. permutations); the eigenspace corresponding to eigenvalue 1 is then a fixed-point set. Below we present our results for discrete symmetries.

Theorem 2 (Discrete symmetry identities)**.** *Suppose f admits an (H,Id)-equivariance, i.e., $f(H(\boldsymbol{\theta},\boldsymbol{\lambda})) = f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Fix $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$ and let $\boldsymbol{\theta} \in \mathrm{Fix}(H,\boldsymbol{\lambda}_0)$. Define $\widehat{X}(\boldsymbol{\theta},\boldsymbol{\lambda}_0) := \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta},\boldsymbol{\lambda}_0)$. Then:*

1. ***(First-order identity).*** *The gradient of L satisfies*

   (i'). $\nabla L(\boldsymbol{\theta}) \circ \widehat{X}(\boldsymbol{\theta},\boldsymbol{\lambda}_0) = \nabla L(\boldsymbol{\theta})$.

2. ***(Second-order identity).*** *The Hessian of L satisfies:*

   (ii'). $\nabla^2 L(\boldsymbol{\theta}) \circ \widehat{X}(\boldsymbol{\theta},\boldsymbol{\lambda}_0) \circ_2 \widehat{X}(\boldsymbol{\theta},\boldsymbol{\lambda}_0) = \nabla^2 L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}) \circ \nabla^2_\theta H(\boldsymbol{\theta},\boldsymbol{\lambda}_0)$.

The key difference from the continuous case is that here $\widehat{X} = \nabla_{\boldsymbol{\theta}} H$ is the Jacobian of H with respect to $\boldsymbol{\theta}$ (not an inverse times a $\boldsymbol{\lambda}$-derivative), and the identities hold only on the fixed-point set where $H(\boldsymbol{\theta},\boldsymbol{\lambda}_0) = \boldsymbol{\theta}$.

Eq. (i') states that the gradient is an eigenvector of $\widehat{X}^\top$ with eigenvalue 1: the transformation H leaves the gradient invariant (when viewed as a linear functional). Eq. (ii') constrains the structure of the Hessian: conjugating by $\widehat{X}$ reproduces the original Hessian up to a correction term involving ∇L and the second derivative $\nabla^2_{\boldsymbol{\theta}} H$. When H is linear in $\boldsymbol{\theta}$ (so that $\nabla^2_{\boldsymbol{\theta}} H = 0$), the blue term vanishes, and the identity simplifies to $\widehat{X}^\top \nabla^2 L\, \widehat{X} = \nabla^2 L$, i.e., the Hessian is invariant under conjugation by the transformation Jacobian.

These discrete identities are weaker than their continuous counterparts (they hold only on the fixed-point set rather than everywhere), but they still yield nontrivial structural information. In Section C, we show that they recover recent results on mirror symmetry (Ziyin, 2023) as a special case.

REFERENCES

Yossi Arjevani and Michael Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. *Advances in Neural Information Processing Systems*, 33:5441–5452, 2020.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Robert Bamler and Stephan Mandt. Improving optimization for models with continuous symmetry breaking. In *International Conference on Machine Learning*, pp. 423–432. PMLR, 2018.

Johanni Brea, Berfin Simsek, Bernd Illing, and Wulfram Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *arXiv preprint arXiv:1907.02911*, 2019.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Clémentine CJ Dominé, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro AM Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv preprint arXiv:2409.14623*, 2024.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. *Advances in Neural Information Processing Systems*, 36:51621–51662, 2023.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. *Advances in Neural Information Processing Systems*, 37:76848–76900, 2024.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer relu networks with small initialization. *arXiv preprint arXiv:2307.12851*, 2023.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *International Conference on Automated Deduction*, pp. 625–635. Springer, 2021.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016a.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016b.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

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

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. *Advances in Neural Information Processing Systems*, 34: 23914–23927, 2021.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pp. 10153–10161. PMLR, 2021.

Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10473–10486, 2022.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Geonhui Yoo, Minhak Song, and Chulhee Yun. Understanding sharpness dynamics in nn training with a minimalist example: The effects of dataset difficulty, depth, stochasticity, and more. *arXiv preprint arXiv:2506.06940*, 2025.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. *arXiv preprint arXiv:2210.17216*, 2022.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *arXiv preprint arXiv:2506.13018*, 2025.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Liu Ziyin. Symmetry induces structure and constraint of learning. *arXiv preprint arXiv:2309.16932*, 2023.

Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. Parameter symmetry and noise equilibrium of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 37:93874–93906, 2024.

A First-Order Consequences: Conservation Laws and Implicit Bias

We now instantiate the first-order identity eq. (i) in concrete settings. This section shows that many known results in learning dynamics, including conservation laws (Zhao et al., 2022; Głuch & Urbanke, 2021), gradient flow invariants (Arora et al., 2019; Min et al., 2023; Maennel et al., 2018), and scalar identities for implicit bias (Lyu & Li, 2019), arise as special cases of eq. (i).

A.1 Symmetry

When $G(\boldsymbol{y}, \boldsymbol{\lambda}) = \boldsymbol{y}$ (symmetry), we have $Y \equiv \mathbf{0}$, and eq. (i) reduces to

$$\langle \nabla L(\boldsymbol{\theta}),\, X(\boldsymbol{\theta}, \boldsymbol{\lambda}) \rangle = 0. \tag{2}$$

If the model is trained through gradient flow (GF), we have $\dot{\boldsymbol{\theta}} = -\nabla L(\boldsymbol{\theta})$, and the identity becomes $\langle \dot{\boldsymbol{\theta}}, X(\boldsymbol{\theta}, \boldsymbol{\lambda}) \rangle = 0$: the parameter motion is orthogonal to the symmetry-generated direction. Integrating this orthogonality along the flow yields conservation laws, recovering Noether-type results (Głuch & Urbanke, 2021; Kunin et al., 2020). Two representative examples are:[1]

1. **Rescaling symmetry.** If two consecutive layers with trainable parameters $\boldsymbol{W}_1, \boldsymbol{W}_2$ satisfy $f(\boldsymbol{W}_1, \boldsymbol{W}_2) = f(\lambda \boldsymbol{W}_1, \lambda^{-1} \boldsymbol{W}_2)$, then eq. (2) implies $\|\boldsymbol{W}_1\|_F^2 - \|\boldsymbol{W}_2\|_F^2$ is constant during training (Saxe et al., 2013). This conservation law is often used in analyses of two-layer ReLU networks to tie the absolute value of the second-layer coefficients directly with the norm of the first-layer parameters and eliminates its freedom (Min et al., 2023; Maennel et al., 2018).
2. **Reparameterization symmetry.** If a network contains a component $f_0(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{W}_1 \boldsymbol{W}_2$, then for any invertible matrix $\boldsymbol{P}$, we have $f(\boldsymbol{P}\boldsymbol{W}_1, \boldsymbol{P}^{-1}\boldsymbol{W}_2) = f(\boldsymbol{W}_1, \boldsymbol{W}_2)$. This reparameterization symmetry appears in deep linear networks and self-attention mechanisms. Eq. (2) then implies $\boldsymbol{W}_1 \boldsymbol{W}_1^\top - \boldsymbol{W}_2 \boldsymbol{W}_2^\top = \mathrm{const}$, which has been used to analyze solution stability (Marion & Chizat, 2024; Arora et al., 2019; Tarmoun et al., 2021; Du et al., 2018; Dominé et al., 2024).

In discrete-time gradient descent, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t)$, the identity eq. (2) still implies that the update is orthogonal to the characteristic direction at each iterate, capturing the same geometric constraint on the update direction.

A.2 Homogeneity

We say f is m**-homogeneous** if $f(\lambda \boldsymbol{\theta}) = \lambda^m f(\boldsymbol{\theta})$ for all $\lambda \in \mathbb{R}$. For example, an m-layer ReLU network without bias is m-homogeneous. Homogeneity fits our framework with $H(\boldsymbol{\theta}, \lambda) = \lambda \boldsymbol{\theta}$ and $G(\boldsymbol{y}, \lambda) = \lambda^m \boldsymbol{y}$. At any good position with $\lambda \neq 0$, we have $X(\boldsymbol{\theta}, \lambda) = \lambda^{-1} \boldsymbol{\theta}$ and $Y(\boldsymbol{\theta}, \lambda) = m \lambda^{-1} f(\boldsymbol{\theta})$. Eq. (i) then yields the well-known Euler-type relation

$$\langle \nabla L(\boldsymbol{\theta}),\, \boldsymbol{\theta} \rangle = m \left\langle \nabla \ell\big(f(\boldsymbol{\theta})\big),\, f(\boldsymbol{\theta}) \right\rangle.$$

Under gradient flow and specific loss functions, this identity translates into a monotonicity statement. To see this, let $c = 1$ and $\dot{\boldsymbol{\theta}} = -\nabla L(\boldsymbol{\theta})$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{2} \|\boldsymbol{\theta}\|^2 = -\langle \boldsymbol{\theta}, \nabla L(\boldsymbol{\theta}) \rangle = -m\, \ell'\big(f(\boldsymbol{\theta})\big)\, f(\boldsymbol{\theta}).$$

For classification losses where $\ell'(y) \cdot y < 0$ when correctly classified (e.g., exponential or logistic loss), once the model enters the correct-classification regime, the right-hand side is positive, so $\|\boldsymbol{\theta}(t)\|$ diverges. This norm divergence is the key step in max-margin implicit bias proofs (Lyu & Li, 2019).

Beyond these two examples. The two cases above already capture the general workflow of the toolbox. Once a continuous equivariance is specified, eq. (i) immediately yields a scalar identity that can be turned into either 1) an invariant along gradient flow (in the symmetry case) or 2) a differential

[1] For notational simplicity, when focusing on a subset of parameters, we view f as a function of those parameters.

inequality controlling the evolution of a meaningful quantity such as the parameter norm or margin (in the non-symmetric case). Such identities are exactly the starting point of many implicit-bias and stability arguments for gradient methods. We emphasize again that the framework is not restricted to $G = \mathrm{Id}$: allowing general equivariances produces non-vanishing right-hand sides in eq. (i).

B Second-Order Consequences: Hessian Structure and Loss Landscape

While the first-order results in Section A unify many known phenomena, most of those consequences have been established (albeit separately) in the literature. The second-order implications of equivariance, however, have not been explored nearly as systematically, which can be essential for studying the geometry of optimization in modern networks, as they translate symmetry/equivariance directly into testable structural predictions about curvature.

In this section, we focus on the second-order identities eqs. (ii) and (iii) that constrain the Hessian $\nabla^2 L(\boldsymbol{\theta})$ through its action and quadratic forms along the same transformation-induced directions. This is precisely the level at which many empirical observations are formulated: sharp versus flat directions, low-rank or low-dimensional structure of curvature, and the alignment between optimization dynamics and the leading eigenspaces of the Hessian (Gur-Ari et al., 2018; Sagun et al., 2016a; Yoo et al., 2025; Kalra & Barkeshli, 2023). Below we illustrate the powerfulness of the second-order identities through several examples.

B.1 Homogeneity

For m-homogeneous models (discussed in Section A.2), recall that $X(\boldsymbol{\theta}, \lambda) = \lambda^{-1}\boldsymbol{\theta}$ and $Y(\boldsymbol{\theta}, \lambda) = m\lambda^{-1}y$ where $y = f(\boldsymbol{\theta})$. As in Section A.2, here we consider the case where $c = 1$ for simplicity of the illustration. Substituting the equivariance into eqs. (ii) and (iii) yields the following identities:

$$\nabla^2 L(\boldsymbol{\theta})\,\boldsymbol{\theta} = \left(my\frac{\ell''(y)}{\ell'(y)} + m - 1\right)\nabla L(\boldsymbol{\theta}), \tag{3}$$

$$\left\langle \boldsymbol{\theta},\, \nabla^2 L(\boldsymbol{\theta})\,\boldsymbol{\theta}\right\rangle = \ell''(y)\, m^2 y^2 + \ell'(y)\, m(m-1)\, y, \tag{4}$$

whenever $\ell'(y) \neq 0$. These identities themselves already provide strong information about the structure of the Hessian and the Hessian-gradient relationship, as they relate the action (and quadratic form) of the Hessian along $\boldsymbol{\theta}$ to low-order quantities.

Gradient descent happens in a tiny subspace. By performing eigen-decomposition of the Hessian, the first identity eq. (3) immediately gives the following corollary.

Corollary 1. *For a m-homogeneous f and a good position $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, there exists $\alpha(\boldsymbol{\theta}) \in \mathbb{R}$ such that $\langle \boldsymbol{g}, \boldsymbol{u}_k\rangle = \lambda_k \alpha(\boldsymbol{\theta})\langle \boldsymbol{\theta}, \boldsymbol{u}_k\rangle$, where $\boldsymbol{g} = \nabla L(\boldsymbol{\theta})$ and $(\lambda_k, \boldsymbol{u}_k)$ are eigenvalue-eigenvector pairs of $\nabla^2 L(\boldsymbol{\theta})$.*

Corollary 1 reveals a deep connection of the angles between the gradient, eigenvector of the Hessian, and the parameter. A first consequence is that $\nabla L(\boldsymbol{\theta})$ must lie in the column space of $\nabla^2 L(\boldsymbol{\theta})$, which has been proven to be low-dimensional around the stable points (Arjevani & Field, 2020; Wei et al., 2022). Moreover, in regimes where $\boldsymbol{\theta}$ does not concentrate on a particular subset of eigenvectors (e.g., as a rough heuristic, $\langle \boldsymbol{\theta}, \boldsymbol{u}_k\rangle$ being of comparable magnitude across many k), the gradient preferentially aligns with directions of large $|\lambda_k|$, concentrating in leading eigenspaces. This provides a concrete structural explanation for the empirical observation that gradient descent effectively evolves within a small subspace associated with dominant curvature directions (Gur-Ari et al., 2018).

Sharpness lower bound. The second identity eq. (4) yields an immediate curvature lower bound along the direction of $\boldsymbol{\theta}$. Let K denote the sharpness of the loss (i.e. largest eigenvalue of $\nabla^2 L(\boldsymbol{\theta})$), eq. (4) gives a computable lower bound on K:

$$K = \sup_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{\left\langle \boldsymbol{x}, \nabla^2 L(\boldsymbol{\theta})\boldsymbol{x}\right\rangle}{\|\boldsymbol{x}\|^2} \;\geq\; \frac{\left\langle \boldsymbol{\theta}, \nabla^2 L(\boldsymbol{\theta})\boldsymbol{\theta}\right\rangle}{\|\boldsymbol{\theta}\|^2} = \frac{m}{\|\boldsymbol{\theta}\|^2}\Big(\ell''(y)\, my^2 + \ell'(y)\,(m-1)\,y\Big).$$

Although this bound probes only a single direction, it is a particularly canonical one for homogeneous models: scaling $\boldsymbol{\theta}$ changes the output in a controlled manner, and thus curvature along $\boldsymbol{\theta}$ captures a fundamental component of the loss geometry induced by homogeneity.

As an example of a concrete setting, consider $c = 1, m = 1$ and the exponential loss $\ell(y) = \exp(-y\hat{y})$ for binary classification, this becomes $K \geq \frac{\ell(y)y^2}{\|\boldsymbol{\theta}\|^2}$. In early training when $|y|$ is small, the lower bound is dominated by the y^2 term, suggesting a monotonic growth of the sharpness lower bound as training goes, partially verifying the previously observed progressive sharpening phenomenon (Cohen et al., 2021).

B.2 Noether Flow under Stochastic Dynamics

Now, we consider the setting where the equivariance is a symmetry ($G = \mathrm{Id}$) and the training is stochastic gradient flow (SGF). Recall from Section A.1 that under gradient flow, symmetries induce conserved quantities. Specifically, assume $p = 1$ and $H(\boldsymbol{\theta}, 0) = \boldsymbol{\theta}$ (so $\lambda = 0$ corresponds to the identity transformation). Then $X(\boldsymbol{\theta}, 0) = \nabla_\lambda H(\boldsymbol{\theta}, 0)$. If there exists $C : \mathbb{R}^d \to \mathbb{R}$ with

$$\nabla C(\boldsymbol{\theta}) = \nabla_\lambda H(\boldsymbol{\theta}, 0), \tag{5}$$

then $\dot{C}(\boldsymbol{\theta}) = \langle \nabla C(\boldsymbol{\theta}), \dot{\boldsymbol{\theta}} \rangle = -\langle \nabla_\lambda H(\boldsymbol{\theta}, 0), \nabla L(\boldsymbol{\theta}) \rangle = 0$, which means C is conserved under GF (called a Noether charge).

Under SGF, however, C is no longer conserved but can exhibit a systematic drift called *Noether flow* (Ziyin et al., 2024). Our framework provides a precise characterization of this drift. Specifically, in this subsection we follow the setting of (Ziyin et al., 2024), which adopts a multi-sample setting. Let Ω be the set of all possible input data and μ be a distribution over Ω. For a data sample $x \in \Omega$, let $L_x(\boldsymbol{\theta})$ represent the loss function with data x and trainable parameter $\boldsymbol{\theta}$. Let $L(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mu} L_x(\boldsymbol{\theta})$ be the expected loss function. Let $\Sigma(\boldsymbol{\theta}) = \mathbb{E}\left(\nabla L_x(\boldsymbol{\theta}) \nabla L_x(\boldsymbol{\theta})^\top\right) - \nabla L(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta})^\top$ be the noise covariance matrix. Specifically, we can obtain the following characterization of the Noether flow using Theorem 1:

Corollary 2. *If L is sufficiently smooth, and the parameter $\boldsymbol{\theta}$ is trained through SGF (defined in (Ziyin et al., 2024)), the conserved quantity C from eq. (5) evolves as*

$$\dot{C}(\boldsymbol{\theta}) = -\frac{\sigma^2}{2} \left\langle \nabla C(\boldsymbol{\theta}), \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{Tr}(\Sigma(\boldsymbol{\theta})) \right\rangle,$$

where $\Sigma(\boldsymbol{\theta})$ is the noise covariance matrix and σ is the strength of the noise in the SGF.

Corollary 2 shows that the Noether charge drifts in a direction determined by how the total gradient noise $\mathrm{Tr}(\Sigma(\boldsymbol{\theta}))$ changes along the symmetry direction $\nabla C(\boldsymbol{\theta})$. If noise increases along ∇C, the charge decreases, and vice versa. This provides a precise mechanism for how stochasticity breaks conservation laws: the drift is not arbitrary but is governed by the interaction between the symmetry structure and the noise geometry.

B.3 Last-Layer Equivariance

We now consider an equivariance that is neither a symmetry nor homogeneity. Consider a model of the form

$$f(\boldsymbol{W}, \boldsymbol{\theta}') = \boldsymbol{W} h(\boldsymbol{\theta}'), \tag{6}$$

where $\boldsymbol{W} \in \mathbb{R}^{c\times s}$ is the last-layer weight matrix, $\boldsymbol{\theta}' \in \mathbb{R}^{d'}$ contains the remaining parameters, and $h : \mathbb{R}^{d'} \to \mathbb{R}^s$ represents the feature extractor, $d' = d - cs$ is the number of parameters of the model excluding the last layer. For any $\boldsymbol{P} \in \mathbb{R}^{c\times c}$, the model admits the equivariance $f(\boldsymbol{P}\boldsymbol{W}, \boldsymbol{\theta}') = \boldsymbol{P} f(\boldsymbol{W}, \boldsymbol{\theta}')$, fitting our framework with $H((\boldsymbol{W}, \boldsymbol{\theta}'), \boldsymbol{P}) = (\boldsymbol{P}\boldsymbol{W}, \boldsymbol{\theta}')$ and $G(\boldsymbol{y}\boldsymbol{P}) = \boldsymbol{P}\boldsymbol{y}$.

Applying this equivariance to the the second-order identity eq. (iii) yields the following result:

Corollary 3. *For f satisfying eq. (6) and any $\boldsymbol{V} \in \mathbb{R}^{c\times s}$ whose row space is contained in that of $\boldsymbol{W}$,*

$$\langle \mathrm{vec}\, \boldsymbol{V}, \nabla^2_{\mathrm{vec}\, \boldsymbol{W}} L(\boldsymbol{\theta}) \,\mathrm{vec}\, \boldsymbol{V} \rangle = \langle \boldsymbol{V}\boldsymbol{h}, \nabla^2 \ell(\boldsymbol{y}) \boldsymbol{V}\boldsymbol{h} \rangle, \tag{7}$$

where $\boldsymbol{h} = h(\boldsymbol{\theta}')$, $\boldsymbol{y} = f(\boldsymbol{\theta})$, and $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{\theta}')$.

This identity directly connect the parameter space sharpness along direction $\boldsymbol{V}$ with the output space sharpness along direction $\boldsymbol{V}\boldsymbol{h}$. This connection could be potentially related to the self-duality structure of the last layer weight matrix observed in (Papyan et al., 2020).

As a specific example, consider the softmax cross-entropy loss. The Hessian of ℓ satisfies $\nabla^2 \ell(\boldsymbol{y}) = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ where $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{y})$, so eq. (7) becomes $\langle \mathrm{vec}\, \boldsymbol{V}, \nabla^2_{\mathrm{vec}\, \boldsymbol{W}} L \ \mathrm{vec}\, \boldsymbol{V} \rangle = \mathrm{Var}_{\boldsymbol{p}}(\boldsymbol{V}\boldsymbol{h})$, where $\mathrm{Var}_{\boldsymbol{p}}$ is the variance under the distribution $\boldsymbol{p}$. As training progresses and $\boldsymbol{p}$ concentrates on the true class k, perturbations that redistribute mass among low-probability classes have vanishing curvature (flat directions), while perturbations affecting the top-vs-rest margin have large curvature (sharp directions). This explains why late-stage last-layer geometry becomes increasingly rigid, with directions stabilized up to rescaling (Papyan et al., 2020; Zhu et al., 2021).

B.4 Symmetry: Hessian Degeneracy at Stationarity

In Section B.2 we examined how symmetries affect dynamics under stochastic training. Here we consider a complementary aspect: what symmetries imply about the Hessian at stationary points.

For symmetry ($G = \mathrm{Id}$), the characteristic output $Y \equiv \mathbf{0}$. At any first-order stable point $\boldsymbol{\theta}_\star$ (where $\nabla L(\boldsymbol{\theta}_\star) = \mathbf{0}$) we have

$$\nabla^2 L(\boldsymbol{\theta}_\star) \circ X(\boldsymbol{\theta}_\star, \boldsymbol{\lambda}) = \mathbf{0}.$$

Thus, every symmetry-generated direction $X(\boldsymbol{\theta}_\star, \boldsymbol{\lambda})$ lies in the null space of the Hessian at stationarity. This is consistent with the principle that continuous symmetries create degenerate manifolds of equivalent solutions (Bamler & Mandt, 2018). Consequently, if a model has q independent continuous symmetries whose characteristic directions span a q-dimensional subspace at $\boldsymbol{\theta}_\star$, then $\nabla^2 L(\boldsymbol{\theta}_\star)$ has at least p zero eigenvalues. This provides a structural explanation for why Hessians of overparameterized networks often exhibit large near-zero spectral bulk (Arjevani & Field, 2020; Wu et al., 2020; Sagun et al., 2016b; Singh et al., 2021; Sagun et al., 2017).

We note that previous work (Kunin et al., 2020) has derived equivalent second-order identities for continuous symmetries but without much discussion; our contribution here is then to place these results within the broader equivariance framework that also handles non-symmetric and non-continuous cases.

C Discrete Symmetry Consequences

The preceding sections focused on smooth equivariances, where the transformation parameter $\boldsymbol{\lambda}$ varies continuously. However, many symmetries in neural networks are inherently discrete: sign flips, permutations of hidden neurons, or reflections across subspaces. For such transformations, the continuous identities of Theorem 1 do not directly apply since we cannot differentiate with respect to $\boldsymbol{\lambda}$. Nevertheless, Theorem 2 shows that useful constraints still hold on the fixed-point set of the transformation.

Specifically, consider a discrete symmetry H and a point $\boldsymbol{\lambda}_0$, assume $H(\cdot, \boldsymbol{\lambda}_0)$ is a linear transformation of the parameters, i.e. $H(\boldsymbol{\theta}, \boldsymbol{\lambda}_0) = \boldsymbol{P}\boldsymbol{\theta}$ for some fixed matrix $\boldsymbol{P}$ (depending on $\boldsymbol{\lambda}_0$). Since $G = \mathrm{Id}$ for symmetry and H is linear in $\boldsymbol{\theta}$, Theorem 2 simplifies to

$$\boldsymbol{P}^\top \nabla L(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta}), \tag{8}$$

$$\boldsymbol{P}^\top \nabla^2 L(\boldsymbol{\theta})\, \boldsymbol{P} = \nabla^2 L(\boldsymbol{\theta}), \tag{9}$$

for any $\boldsymbol{\theta} \in \mathrm{Fix}(H, \boldsymbol{\lambda}_0)$.

These identities have clear interpretations. Eq. (8) states that the gradient is an eigenvector of $\boldsymbol{P}^\top$ with eigenvalue 1: the symmetry leaves the gradient invariant on the fixed-point set. Eq. (9) states that the Hessian commutes with $\boldsymbol{P}$ (in the sense of conjugation): the curvature respects the same symmetry structure as the transformation.

An important consequence is that eigenvectors of $\boldsymbol{P}$ organize the Hessian's eigenspaces. If $\boldsymbol{P}$ has distinct eigenvalues, then eigenvectors of $\nabla^2 L(\boldsymbol{\theta})$ must align with eigenvectors of $\boldsymbol{P}$. More generally, eq. (9) implies that the Hessian block-diagonalizes according to the eigenspaces of $\boldsymbol{P}$: directions corresponding to different eigenvalues of $\boldsymbol{P}$ are decoupled in the Hessian.

C.1 Mirror Symmetry

As a concrete application, we recover the mirror symmetry results of Ziyin (2023) as a special case. Let $\boldsymbol{O} \in \mathbb{R}^{d \times d'}$ be a matrix with orthonormal columns ($d' \leq d$), defining a d'-dimensional subspace. The mirror transformation defined by $\boldsymbol{O}$ is $\boldsymbol{P} = \boldsymbol{I} - 2\boldsymbol{O}\boldsymbol{O}^\top$, which reflects vectors across the orthogonal complement of $\mathrm{col}(\boldsymbol{O})$. The fixed-point set consists of all $\boldsymbol{\theta} \perp \mathrm{col}(\boldsymbol{O})$. Applying eqs. (8) and (9) yields the following result:

Corollary 4. *If f has (H, Id)-equivariance with $H(\boldsymbol{\theta}, \boldsymbol{\lambda}_0) = (\boldsymbol{I} - 2\boldsymbol{O}\boldsymbol{O}^\top)\boldsymbol{\theta}$, then for any $\boldsymbol{\theta} \perp \mathrm{col}(\boldsymbol{O})$:*

1. *$\boldsymbol{O}^\top \nabla L(\boldsymbol{\theta}) = \boldsymbol{0}$, i.e., the gradient has no component in $\mathrm{col}(\boldsymbol{O})$;*
2. *The Hessian preserves the subspace decomposition: $\nabla^2 L(\boldsymbol{\theta})\, \mathrm{col}(\boldsymbol{O}) \subseteq \mathrm{col}(\boldsymbol{O})$ and $\nabla^2 L(\boldsymbol{\theta})\, \mathrm{col}(\boldsymbol{O})^\perp \subseteq \mathrm{col}(\boldsymbol{O})^\perp$.*

Corollary 4 recovers Theorems 1.1 and 1.2 of Ziyin (2023), which is the key theoretical result of it.

D Discussion

In this work, we have developed a general equivariance toolbox that provides coupled first- and second-order constraints on learning dynamics. The framework substantially extends classical Noether-type analyses in three directions: 1) from first-order to second-order, capturing Hessian structure beyond gradient orthogonality; 2) from symmetry to general equivariance, handling cases like homogeneity where outputs transform non-trivially; and 3) from continuous to discrete transformations, where fixed-point constraints replace differential identities. Together, these extensions yield a unified calculus for translating transformation structure into predictions about optimization geometry.

Beyond Noether: from symmetry to equivariance. Classical applications of Noether's theorem to learning dynamics focus on symmetries ($G = \mathrm{Id}$), deriving conservation laws from the orthogonality $\langle \nabla L, X \rangle = 0$. Our framework generalizes this by allowing non-trivial output transformations G, which produces the richer identity $\langle \nabla L, X \rangle = \langle \nabla \ell, Y \rangle$. This extension is essential for capturing equivariances like homogeneity, where the right-hand side is non-zero and drives the implicit bias. In this sense, the toolbox unifies two previously separate classes of results: Noether-type conservation laws and homogeneity-based scalar identities now emerge as special cases of the same first-order identity.

Second-order structure and curvature phenomena. A central contribution is that the same equivariance calculus yields second-order constraints (Theorem 1, eqs. (ii) and (iii)). These provide structural information about $\nabla^2 L(\boldsymbol{\theta})$: how the Hessian acts along characteristic directions and what its quadratic forms evaluate to. When H and G are linear in their main arguments, the general formulas simplify to compact expressions directly comparable with empirical observations. This connects transformation structure to modern curvature phenomena such as low-rank Hessians, progressive sharpening, and gradient concentration in leading eigenspaces.

Discrete symmetries and fixed-point constraints. The extension to discrete symmetries (Theorem 2) shows that useful constraints persist even without continuous parameterization. While continuous symmetries yield differential identities holding everywhere, discrete symmetries yield algebraic constraints on the fixed-point set: the gradient must be invariant under the transformation, and the Hessian must commute with it. These constraints predict block-diagonal Hessian structure aligned with symmetry eigenspaces, complementing the zero-eigenvalue predictions from continuous symmetries. The mirror symmetry results of Ziyin (2023) emerge as a special case, illustrating how the framework recovers and contextualizes recent findings.

Limitations and future directions. While the equivariance identities in Theorems 1 and 2 are universal, turning them into sharp predictions in concrete learning problems still requires additional modeling choices. We highlight two directions that we view as particularly important.

- **Training-dependent refinements.** Many of our current Hessian characterizations are obtained by directly "reading off" structure from the identities, which is broadly applicable but largely agnostic to the training setting. In practice, data distributions, optimizer details (e.g., learning rate, momentum, weight decay, sharpness-aware updates), and architectural choices may determine which equivariances are effectively expressed and how strongly the induced constraints

shape the trajectory. An important next step is to combine our identities with concrete training models to derive more quantitative predictions, such as evolution laws or scaling behaviors for curvature/sharpness along characteristic directions.

- **Beyond global equivariances: local and emergent structure.** Current analysis mainly emphasizes global equivariances/symmetries, whereas realistic training may exhibit additional local or approximate equivariances only in certain regimes (e.g., near particular solutions due to saturation, feature collapse, or local linearization). Since our framework is fundamentally local (as it only relies on derivatives around a good position), it can in principle accommodate such cases, but we do not yet provide a systematic characterization of when these emergent equivariances arise or how to validate them empirically. Developing practical criteria and second-order signatures for local equivariances may help further clarify end-stage phenomena such as progressive flattening and structured degeneracies.

Overall, the equivariance toolbox provides a unified interface between transformation structure and learning dynamics, extending classical Noether-type analyses to second-order constraints and discrete symmetries, and offering a systematic route to interpret modern empirical observations about optimization geometry.

E Proofs of the theoretical results

We first state some basic tensor-function operations to simplify the subsequent derivations.

Lemma 1. *Let a, b be positive integers, and let s, t be tensor shapes. The following statements hold.*

1. ***(Function composition).*** *If $f : \mathbb{R}^a \to \mathbb{T}(b)$ and $g : \mathbb{R}^b \to \mathbb{T}(s)$, then $g \circ f : \mathbb{R}^a \to \mathbb{T}(s)$.*
2. ***(Tensor composition).*** *If $f \in \mathbb{T}(t, b)$ and $g \in \mathbb{T}(b, s)$, then $g \circ f \in \mathbb{T}(t, s)$.*
3. ***(Gradient).*** *If $f : \mathbb{R}^a \to \mathbb{T}(s)$, then $\nabla f : \mathbb{R}^a \to \mathbb{T}(a, s)$.*
4. ***(Chain rule).*** *If $f : \mathbb{R}^a \to \mathbb{T}(b)$ and $g : \mathbb{R}^b \to \mathbb{T}(s)$, then $h = g \circ f : \mathbb{R}^a \to \mathbb{T}(s)$ and*
$$\forall \boldsymbol{x} \in \mathbb{R}^a, \qquad \nabla h(\boldsymbol{x}) = (\nabla g \circ f)(\boldsymbol{x}) \circ \nabla f(\boldsymbol{x}).$$
Moreover, if f, g are both tensors (i.e., $f \in \mathbb{T}(a, b)$ and $g \in \mathbb{T}(b, s)$), then h is also a tensor.
5. ***(Product rule).*** *If $f : \mathbb{R}^a \to \mathbb{T}(t, b)$ and $g : \mathbb{R}^a \to \mathbb{T}(b, s)$, define $h : \mathbb{R}^a \to \mathbb{T}(t, s)$ by $\forall \boldsymbol{x} \in \mathbb{R}^a,\ h(\boldsymbol{x}) = g(\boldsymbol{x}) \circ f(\boldsymbol{x})$. Then*
$$\forall \boldsymbol{x} \in \mathbb{R}^a, \qquad \nabla h(\boldsymbol{x}) = \nabla g(\boldsymbol{x}) \circ_2 f(\boldsymbol{x}) + g(\boldsymbol{x}) \circ \nabla f(\boldsymbol{x}).$$

E.1 Proof of the Main Theorems

In this subsection, we prove Theorem 1 and Theorem 2 together. Throughout, we assume that $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is at a good position.

Taking derivatives with respect to $\boldsymbol{\theta}$ on both sides of eq. (1), we obtain
$$\nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\theta H(\boldsymbol{\theta}, \lambda) = \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ \nabla f(\boldsymbol{\theta}). \tag{10}$$
Taking derivatives with respect to $\boldsymbol{\lambda}$ on both sides of eq. (1), we obtain
$$\nabla f\,(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\lambda G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}). \tag{11}$$

Since $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a good position, $\nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})$ are invertible. Note that $\nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \in \mathbb{T}(d, d)$ and $\nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \in \mathbb{T}(c, c)$. Therefore, using eq. (10) and eq. (11), we have
$$\begin{aligned}\nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &= \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &= \nabla_\lambda G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}).\end{aligned} \tag{12}$$
Left-multiplying by $\nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1}$, we obtain
$$\nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}). \tag{13}$$
Using $\nabla L(\boldsymbol{\theta}) = \nabla \ell(f(\boldsymbol{\theta})) \circ \nabla f(\boldsymbol{\theta})$, we obtain eq. (i).

Second-order identities. Now, right-composing eq. (10) with $\nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1}$, we obtain

$$\nabla f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) = \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1}. \tag{14}$$

Taking derivatives with respect to $\boldsymbol{\theta}$ on both sides of eq. (10), we obtain

$$\begin{aligned}&\nabla^2 f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_\pi \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &= \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ \nabla f(\boldsymbol{\theta}) \circ_2 \nabla f(\boldsymbol{\theta}) + \nabla_y G(f(\boldsymbol{\theta}), \lambda) \circ \nabla^2 f(\boldsymbol{\theta}).\end{aligned} \tag{15}$$

Taking derivatives with respect to $\boldsymbol{\lambda}$ on both sides of eq. (10), we obtain

$$\begin{aligned}&\nabla^2 f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_\pi \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\lambda \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &= \nabla_\lambda \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ_2 \nabla f(\boldsymbol{\theta}).\end{aligned} \tag{16}$$

Taking derivatives with respect to $\boldsymbol{\lambda}$ on both sides of eq. (11), we obtain

$$\nabla^2 f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_\pi \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\lambda^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \nabla_\lambda^2 G(f(\boldsymbol{\theta}), \lambda). \tag{17}$$

For simplicity, we denote $A = X(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $B = Y(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Right-composing eq. (15) with A and using eq. (i), we get

$$\begin{aligned}&\nabla^2 f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_\pi \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ A + \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ A \\ &= \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ (\nabla f(\boldsymbol{\theta}) \circ A) \circ_2 \nabla f(\boldsymbol{\theta}) + \nabla_y G(f(\boldsymbol{\theta}), \lambda) \circ \nabla^2 f(\boldsymbol{\theta}) \circ A \\ &= \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ B \circ_2 \nabla f(\boldsymbol{\theta}) + \nabla_y G(f(\boldsymbol{\theta}), \lambda) \circ \nabla^2 f(\boldsymbol{\theta}) \circ A.\end{aligned} \tag{18}$$

Note that the first term on the left-hand side of eq. (18) is the same as the first term on the left-hand side of eq. (16). Therefore, we can use eq. (16) to eliminate this term. Moreover, we can use eq. (14) to eliminate $\nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda}))$. We obtain

$$\begin{aligned}\nabla^2 f(\boldsymbol{\theta}) \circ A = &- \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &+ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ A \\ &+ \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ_2 \nabla f(\boldsymbol{\theta}) \\ &- \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ B \circ_2 \nabla f(\boldsymbol{\theta}).\end{aligned} \tag{19}$$

Similarly, if we right-compose eq. (15) with A in its first two arguments (i.e., compose with $A \circ_2 A$), we get

$$\begin{aligned}&\nabla^2 f\left(H(\boldsymbol{\theta}, \boldsymbol{\lambda})\right) \circ \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ_\pi \nabla_\lambda H(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \nabla f(H(\boldsymbol{\theta}, \boldsymbol{\lambda})) \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ A \circ_2 A \\ &= \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ B \circ_2 B + \nabla_y G(f(\boldsymbol{\theta}), \lambda) \circ \nabla^2 f(\boldsymbol{\theta}) \circ A \circ_2 A.\end{aligned} \tag{20}$$

Using eq. (17) and eq. (14) in eq. (20), we obtain

$$\begin{aligned}\nabla^2 f(\boldsymbol{\theta}) \circ A \circ_2 A = &- \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ &+ \nabla f(\boldsymbol{\theta}) \circ \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} \circ \nabla_\theta^2 H(\boldsymbol{\theta}, \boldsymbol{\lambda}) \circ A \circ_2 A \\ &+ \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_\lambda^2 G(f(\boldsymbol{\theta}), \lambda) \\ &- \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda})^{-1} \circ \nabla_y^2 G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}) \circ B \circ_2 B.\end{aligned} \tag{21}$$

Using $\nabla L(\theta) = \nabla \ell(f(\boldsymbol{\theta})) \circ \nabla f(\boldsymbol{\theta})$ and

$$\nabla^2 L(\theta) = \nabla^2 \ell(f(\boldsymbol{\theta})) \circ \nabla f(\boldsymbol{\theta}) \circ_2 \nabla f(\boldsymbol{\theta}) + \nabla \ell(f(\boldsymbol{\theta})) \circ \nabla^2 f(\boldsymbol{\theta}), \tag{22}$$

in eq. (19) and eq. (21), we obtain eq. (ii) and eq. (iii), respectively.

Discrete identities. In the discrete case, taking derivatives with respect to $\boldsymbol{\lambda}$ is prohibited, since we do not have structural information about H with respect to $\boldsymbol{\lambda}$. Therefore, only eq. (10) and eq. (15) are valid. For a specific $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$, if $\boldsymbol{\theta} \in \mathrm{Fix}(H, \boldsymbol{\lambda}_0)$, then $H(\boldsymbol{\theta}, \boldsymbol{\lambda}_0) = \boldsymbol{\theta}$. Substituting this identity into eq. (10) and eq. (15), we obtain

$$\nabla f(\boldsymbol{\theta}) \circ \widehat{A} = \widehat{B} \circ \nabla f(\boldsymbol{\theta}), \tag{23}$$

and

$$\nabla^2 f(\theta) \circ \widehat{A} \circ_\pi \widehat{A} + \nabla f(\boldsymbol{\theta}) \circ \nabla^2_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}_0) = \nabla^2_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}_0) \circ \nabla f(\boldsymbol{\theta}) \circ_2 \nabla f(\boldsymbol{\theta}) + \widehat{B} \circ \nabla^2 f(\boldsymbol{\theta}), \tag{24}$$

where $\widehat{A} = \nabla_\theta H(\boldsymbol{\theta}, \boldsymbol{\lambda}_0)$ and $\widehat{B} = \nabla_y G(f(\boldsymbol{\theta}), \boldsymbol{\lambda}_0)$. Since we only consider symmetry here, i.e., $G = \mathrm{Id}$, we have $\widehat{B} = \mathrm{Id}$. Using $\nabla L(\boldsymbol{\theta}) = \nabla \ell(f(\boldsymbol{\theta})) \circ \nabla f(\boldsymbol{\theta})$ and eq. (22) in eq. (23) and eq. (24), we obtain eq. (i') and eq. (9), respectively.

E.2 Proof of Corollary 1

If $\ell'(y) = 0$ or $my\ell''(y) + (m-1)\ell'(y) = 0$, then $\nabla^2 L(\boldsymbol{\theta})\,\boldsymbol{\theta} = \mathbf{0}$, in which case the proposition clearly holds. In the following, we assume $\ell'(y) \neq 0$ and $my\ell''(y) + (m-1)\ell'(y) \neq 0$.

Let $\alpha(\boldsymbol{\theta}) = \left(my\frac{\ell''(y)}{\ell'(y)} + m - 1\right)^{-1}$. Since $\ell'(y) \neq 0$ and $my\ell''(y) + (m-1)\ell'(y) \neq 0$, α is well-defined. Using eq. (3), we have

$$\begin{aligned}
\langle \boldsymbol{g}, \boldsymbol{u}_k \rangle &= \alpha(\boldsymbol{\theta}) \left\langle \boldsymbol{u}_k, \nabla^2 L(\boldsymbol{\theta})\,\boldsymbol{\theta} \right\rangle \\
&= \alpha(\boldsymbol{\theta}) \left\langle \nabla^2 L(\boldsymbol{\theta})\,\boldsymbol{u}_k, \boldsymbol{\theta} \right\rangle \\
&= \alpha(\boldsymbol{\theta}) \lambda_k \left\langle \boldsymbol{u}_k, \boldsymbol{\theta} \right\rangle .
\end{aligned}$$

E.3 Proof of Corollary 2

Using the characterization in (Ziyin et al., 2024), Section 4.1, we have

$$\dot{C}(\boldsymbol{\theta}) = \sigma^2 \mathrm{Tr}\big(\Sigma(\boldsymbol{\theta})\,\nabla^2 C(\boldsymbol{\theta})\big).$$

Note that $\nabla^2 C(\boldsymbol{\theta}) = \nabla_\theta \nabla_\lambda H(\boldsymbol{\theta}, 0)$. Since the symmetry holds for every input datum x, by Theorem 1 and eq. (ii), we have

$$\nabla^2 C(\boldsymbol{\theta})\,\nabla L_x(\boldsymbol{\theta}) = -\nabla^2 L_x(\boldsymbol{\theta})\,\nabla C(\boldsymbol{\theta}).$$

Moreover, the symmetry also holds for the expected loss (since it does not change the model output); therefore, we also have

$$\nabla^2 C(\boldsymbol{\theta})\,\nabla L(\boldsymbol{\theta}) = -\nabla^2 L(\boldsymbol{\theta})\,\nabla C(\boldsymbol{\theta}).$$

Therefore,

$$\begin{aligned}
\frac{1}{\sigma^2}\dot{C}(\boldsymbol{\theta}) &= \mathrm{Tr}\big(\Sigma(\boldsymbol{\theta})\,\nabla^2 C(\boldsymbol{\theta})\big) \\
&= \mathbb{E}\,\mathrm{Tr}\big(\nabla L_x(\boldsymbol{\theta})\,\nabla L_x(\boldsymbol{\theta})^\top \nabla^2 C(\boldsymbol{\theta})\big) - \mathrm{Tr}\big(\nabla L(\boldsymbol{\theta})\,\nabla L(\boldsymbol{\theta})^\top \nabla^2 C(\boldsymbol{\theta})\big) \\
&= \mathbb{E}\left\langle \nabla L_x(\boldsymbol{\theta}), \nabla^2 C(\boldsymbol{\theta})\,\nabla L_x(\boldsymbol{\theta}) \right\rangle - \left\langle \nabla L(\boldsymbol{\theta}), \nabla^2 C(\boldsymbol{\theta})\,\nabla L(\boldsymbol{\theta}) \right\rangle \\
&= -\mathbb{E}\left\langle \nabla L_x(\boldsymbol{\theta}), \nabla^2 L_x(\boldsymbol{\theta})\,\nabla C(\boldsymbol{\theta}) \right\rangle + \left\langle \nabla L(\boldsymbol{\theta}), \nabla^2 L(\boldsymbol{\theta})\,\nabla C(\boldsymbol{\theta}) \right\rangle \\
&= -\mathbb{E}\left\langle \nabla^2 L_x(\boldsymbol{\theta}) \nabla L_x(\boldsymbol{\theta}), \nabla C(\boldsymbol{\theta}) \right\rangle + \left\langle \nabla^2 L(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta}), \nabla C(\boldsymbol{\theta}) \right\rangle \\
&= -\left\langle \nabla C(\boldsymbol{\theta}), \mathbb{E}\,\nabla^2 L_x(\boldsymbol{\theta}) \nabla L_x(\boldsymbol{\theta}) - \nabla^2 L(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta}) \right\rangle \\
&= \frac{1}{2}\left\langle \nabla C(\boldsymbol{\theta}), \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{Tr}(\Sigma(\boldsymbol{\theta})) \right\rangle .
\end{aligned}$$

E.4 Proof of Corollary 3

Substituting the setting into Theorem 1 eq. (iii), we obtain that for any matrix $\boldsymbol{U} \in \mathbb{R}^{c \times c}$,

$$\left\langle \mathrm{vec}(\boldsymbol{U}\boldsymbol{W}), \nabla^2_{\mathrm{vec}(\boldsymbol{W})} L(\boldsymbol{\theta})\ \mathrm{vec}(\boldsymbol{U}\boldsymbol{W}) \right\rangle = \left\langle \boldsymbol{U}\boldsymbol{y},\ \nabla^2 \ell(\boldsymbol{y})\, \boldsymbol{U}\boldsymbol{y} \right\rangle.$$

Let $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{W}$. Since $\boldsymbol{U}$ ranges over all $c \times c$ matrices, $\boldsymbol{V}$ ranges over all $c \times s$ matrices whose row space is contained in the row space of $\boldsymbol{W}$. Note also that $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{h}$. Therefore,

$$\left\langle \mathrm{vec}(\boldsymbol{V}), \nabla^2_{\mathrm{vec}(\boldsymbol{W})} L(\boldsymbol{\theta})\ \mathrm{vec}(\boldsymbol{V}) \right\rangle = \left\langle \boldsymbol{U}\boldsymbol{W}\boldsymbol{h},\ \nabla^2 \ell(\boldsymbol{y})\, \boldsymbol{U}\boldsymbol{W}\boldsymbol{h} \right\rangle = \left\langle \boldsymbol{V}\boldsymbol{h},\ \nabla^2 \ell(\boldsymbol{y})\, \boldsymbol{V}\boldsymbol{h} \right\rangle.$$

E.5 Proof of Corollary 4

First, note that eq. (8) implies $\boldsymbol{O}\boldsymbol{O}^\top \nabla L(\boldsymbol{\theta}) = \boldsymbol{0}$, which is equivalent to $\boldsymbol{O}^\top \nabla L(\boldsymbol{\theta}) = \boldsymbol{0}$.

Moreover, eq. (9) implies

$$\boldsymbol{P} \nabla^2 L(\boldsymbol{\theta}) = \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{P}.$$

Therefore, for any $\boldsymbol{x} \in \mathrm{col}(\boldsymbol{O})$,

$$\boldsymbol{P} \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{x} = \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{P}\boldsymbol{x} = -\nabla^2 L(\boldsymbol{\theta}) \boldsymbol{x},$$

which means $\nabla^2 L(\boldsymbol{\theta}) \boldsymbol{x}$ is an eigenvector of $\boldsymbol{P}$ with eigenvalue -1, and hence $\nabla^2 L(\boldsymbol{\theta}) \boldsymbol{x} \in \mathrm{col}(\boldsymbol{O})$. Similarly, for $\boldsymbol{y} \perp \mathrm{col}(\boldsymbol{O})$, we have

$$\boldsymbol{P} \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{y} = \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{P}\boldsymbol{y} = \nabla^2 L(\boldsymbol{\theta}) \boldsymbol{y},$$

which means $\nabla^2 L(\boldsymbol{\theta}) \boldsymbol{y}$ lies in the eigenspace of $\boldsymbol{P}$ corresponding to eigenvalue 1. This is equivalent to $\nabla^2 L(\boldsymbol{\theta}) \boldsymbol{y} \perp \mathrm{col}(\boldsymbol{O})$.