Regularity in Canonicalized Models: A Theoretical Perspective

Behrooz Tahmasebi
MIT

Stefanie Jegelka
TUM and MIT

Abstract

In learning with invariances (or symmetries), canonicalization is a widely used technique that projects data onto a smaller subset of the input space to reduce associated redundancies. The transformed dataset is then processed through a function from a designated function class to obtain the final invariant representation. Although canonicalization is often simple and flexible, both theoretical and empirical evidence suggests that the projection map can be discontinuous and unstable, which poses challenges for machine learning applications. However, the overall end-to-end representation can still remain continuous. Focusing on the importance of end-to-end regularity rather than the projection mapping itself, this paper explores the continuity and regularity of canonicalized models from a theoretical perspective. In a broad setting of input spaces and group actions, we establish necessary and sufficient conditions for the continuity or regularity of canonicalized models of any order, thereby characterizing the minimal conditions required for stability. To our knowledge, this represents the first comprehensive investigation into the end-to-end regularity of canonicalized models, offering critical insights into their design and application, as well as guidance for enhancing stability in practical settings.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Learning with invariances, or symmetries, has long been a key focus within the machine learning community (Hinton, 1987; Anselmi et al., 2016; Bouvrie et al., 2009; Poggio and Anselmi, 2016). The ability to incorporate known symmetries into learning algorithms can lead to models that are both more efficient and more interpretable. This approach has gained further prominence with applications in scientific domains such as physics (Batzner et al., 2022, 2023; Unke et al., 2021; Grisafi et al., 2018). These advancements have spurred interest in developing neural networks and other machine learning architectures that can directly utilize inherent symmetries in data (Smidt, 2021; Bronstein et al., 2017).

Despite the growing number of applications, there is currently no universally accepted framework for learning with general invariances. Existing techniques include widely-used methods such as data augmentation, as well as group averaging approaches (Murphy et al., 2019), frame averaging (Puny et al., 2022), and canonicalization (Kaba et al., 2023), each of which attempts to integrate invariance into models in different ways. In this paper, we focus on the concept of canonicalization.

Canonicalization is a method for constructing invariant function classes that first projects data onto a substantially smaller subset of the input space, aiming to eliminate redundancies induced by symmetries. An arbitrary function is then applied to form the final representation. Figure Figure 1 illustrates this method. In this paper, the class of functions used to build the canonicalized models is referred to as the class of base functions.

The simplest instance of canonicalization is sorting. A method for constructing permutation-invariant functions over arrays of real numbers is to first sort the array (i.e., canonicalize it) and then apply an arbitrary learning algorithm (i.e., find an appropriate base function) to learn from the canonicalized

data. Although finding permutation-invariant functions by averaging over all invariances is prohibitive, canonicalization offers a straightforward approach to constructing invariant function classes by simply applying the sorting function, which can be executed in nearly linear time. This example illustrates the power of canonicalization in achieving invariant representations with low complexity.

However, it has recently been observed that canonicalized models suffer from the problem of continuity: the projection used to remove invariances from the data is often not continuous. Indeed, recent findings indicate that, for many problems, it is impossible to find canonicalized models via continuous projections (Dym et al., 2024). Given that continuity is the first design principle in machine learning architecture, this issue undermines the computational benefits of canonicalization and poses challenges for utilizing these models across various types of invariances. This challenge extends beyond continuity; it also applies to differentiability and, more generally, higher-order regularities.

Let us further explore the computational advantages of canonicalization, which often provides a simpler approach to obtaining invariant function classes. Consider the case of learning on point clouds under permutation invariance. Suppose each datapoint is an element from $\mathbb{R}^{n\times d}$, representing n points in $\mathbb{R}^d$. The group of permutations of length d (with $d!$ elements) acts on this space by permuting the elements. To utilize group averaging for obtaining an invariant function class, one would need to compute functions of the form:

$$\frac{1}{d!}\sum_{\sigma\in S_d} f(\sigma x),$$

where f is an arbitrary base function, $x \in \mathbb{R}^{d\times n}$ is a point cloud, S_d is the group of permutations of length d, and σx denotes the point cloud obtained by permuting the points within x according to σ.

It is apparent that to use group averaging, one must compute an average over $d!$ permutations, which is computationally intractable. However, there is a natural canonicalization map in this space. Specifically, lexicographic sorting of the points in the point cloud (according to their coordinates) provides a simple and efficient way to obtain invariant function classes, expressed as:

$$f(\pi(x)),$$

where f is an arbitrary base function and $\pi(.)$ denotes the lexicographic sorting. This demonstrates that canonicalization offers significant computational benefits compared to group averaging (as a baseline) for constructing invariant function classes.

However, the lexicographic sorting function $\pi(.)$ is not continuous when $d > 1$. Recent theoretical results [1] indicate that this holds for any canonicalization map $\pi(.)$ in this particular setting. Consequently, while canonicalization is computationally feasible, it is numerically sensitive and unstable near the discontinuity points. In contrast, group averaging preserves continuity (i.e., if f is continuous, then the averaged version is also continuous), though it remains computationally intractable.

It is important to note that the representations obtained through canonicalization can still exhibit end-to-end continuity or differentiability, even if the projection step itself is discontinuous. This leads to a crucial question: what are the minimal conditions necessary to ensure that the end-to-end representation remains continuous (or differentiable)? Is achieving this difficult, or can it be accomplished through simple regularizations applied to the training of base functions? This question is particularly relevant to evaluating the effectiveness of canonicalization, as it pertains to the continuity of the entire architecture and the complexity involved in identifying base functions that ensure overall model regularity.

In this paper, we seek to address the aforementioned question through a rigorous examination of the regularity of end-to-end representations derived from canonicalization. Our primary contribution is a comprehensive characterization of the necessary and sufficient conditions for base functions to ensure continuity in end-to-end canonicalization, while imposing minimal assumptions on the problem setup. Additionally, we apply these findings to various examples, ranging from permutation invariance to sign invariance. We find that, in all cases, the conditions needed to achieve continuity (or differentiability) of the end-to-end representation are prohibitively complex, often requiring exponentially many constraints to be satisfied or necessitating the resolution of a similar learning under invariance problem to achieve regularity. Thus, this work complements existing research on the difficulties and impossibilities associated with achieving regularity in canonicalization, focusing specifically on the regularity of end-to-end representations.

Let us further elaborate on the challenges associated with the derived conditions.

- **Permutation Invariance** (Section 5.4). In this example, we considered lexicographic sorting for point clouds, specifically n points in $\mathbb{R}^d$, modeled as $x \in \mathbb{R}^{n\times d}$. The condition derived in Section 5.4 indicates that continuity can only be achieved if $f(x) = f(\tilde{x})$ for any $x, \tilde{x} \in \mathbb{R}^{n\times d}$ whose columns are permutations of one another and whose first rows are ordered (see Section 5.4). Focusing on the case where the first rows of both x and $\tilde{x}$ are all zeros, we conclude that f must be invariant with respect to all permutations applied to the columns of its inputs. This implies that f must be invariant under permutations of point clouds of size $n \times (d-1)$. Therefore, to achieve continuity, we must construct function classes that are invariant under permutations of n points, which contradicts the goal of canonicalization. Canonicalization seeks to use a general function f, or impose minimal constraints on f, to achieve invariance when composed with a projection.

- **Tori** (Section 5.3). In this example, we consider translation invariance under integer shifts, i.e., invariance to the action of $\mathbb{Z}^d$ on $\mathbb{R}^d$ via shifting. The required conditions for continuity were derived in Section 5.3. Specifically, among other conditions, we require $f(x) = f(\tilde{x})$ for any $x, \tilde{x} \in \{0,1\}^d$ such that $\tilde{x} \prec x$, where $\prec$ denotes lexicographic sorting. The number of such pairs $x, \tilde{x}$ grows exponentially with d, which is what is meant by needing to satisfy exponentially many conditions. It is important to note that the actual continuity condition requires functional equality for infinitely many pairs of points, as illustrated in Section 5.3, which represents an even larger space.

- **Sign Invariance** (Section 5.5). In this example, we consider the action of flipping the sign of all elements in a d-dimensional vector. The required conditions for ensuring continuity in sign canonicalization were derived in Section 5.5. Specifically, this demands that the functional equation $f(v) = f(-v)$ hold for all vectors $v \in \mathbb{R}^d$ such that $v_1 = 0$. In other words, ensuring continuity under sign canonicalization restricts f to be sign-invariant on $\mathbb{R}^{d-1}$, thus transforming the problem of constructing invariant function classes to one for $\mathbb{R}^{d-1}$. However, in this case, group averaging offers more efficient solutions, ensuring both continuity and computational efficiency. We note that the continuity issue in canonicalization also extends to the case where n vectors are invariant to 2^n possible sign flips.

In summary, this paper presents the following contributions:

- We provide a comprehensive study of the end-to-end continuity and differentiability of canonicalized models, with minimal assumptions. Specifically, we establish necessary and sufficient conditions on the class of base functions to ensure that the canonicalized representation is always continuous or differentiable.

- Due to the generality of our results, we can apply them to various settings, ranging from sorting and permutation invariance to sign invariances. We present several tight results demonstrating how continuity and differentiability issues can significantly restrict canonicalized function classes.

2 RELATED WORK

The problem of learning under invariances has been an important topic of research in machine learning, spanning from the early stages of the field (Hinton, 1987) to more recent contributions (Bouvrie et al., 2009; Poggio and Anselmi, 2016; Anselmi et al., 2016). The applications encompass both Euclidean geometry (Smidt, 2021) and broader studies in geometric deep learning (Bronstein et al., 2017).

Many learning architectures have been proposed for specific types of symmetries, such as PointNet for point clouds (Qi et al., 2017a,b). Additionally, methods for constructing general invariant function classes include group averaging (Murphy et al., 2019), frame averaging (Puny et al., 2022; Lin et al., 2024; Baker et al., 2024), and canonicalization (Ma et al., 2024a; Kaba et al., 2023; Panigrahi and Mondal, 2024), among others (Dym and Gortler, 2024; Kim et al., 2023; Pozdnyakov and Ceriotti, 2024).

The continuity problem of canonicalization and frame averaging is recently studied in Dym et al. (2024); Ma et al. (2024b); Zhang et al. (2019); for more, see Huang et al. (2024). The hardness of learning under invariance has also recently been explored (Kiani et al., 2024). Dym et al. (2024) propose alternative methods for achieving continuity using weighted frames. However, their results for canonicalization itself are largely negative. Moreover, their proposed solution via weighted frames requires exponentially large cardinalities (see Table 1 in (Dym et al., 2024)) for permutation or rotational symmetries acting on $\mathbb{R}^{n\times d}$. Specifically, in these cases, the input is $O(nd)$-dimensional, whereas the cardinality size scales as

$n^{\Omega(d)}$, which grows exponentially with respect to the input size.

3 PROBLEM STATEMENT

This section is divided into two subsections. The first covers preliminaries and definitions, and the second explains the problem of continuity, smoothness, and regularization of canonicalized models.

3.1 Preliminaries

Consider a dataset of n labeled samples $\mathcal{S} = \{(x_i, y_i) : i \in [n]\} \subseteq (\mathcal{X} \times \mathbb{R})^n$, where $\mathcal{X}$ is a complete locally compact metric space representing the input space. Here, $[n] := \{1, 2, \ldots, n\}$. The primary objective here is to find a target function $f : \mathcal{X} \to \mathbb{R}$, from a class of functions $\mathcal{F}$, that minimizes, for example, the empirical risk $\sum_{i\in[n]} \ell(f(x_i), y_i)$, where $\ell(.,.)$ denotes a differentiable loss function. Moreover, the data is generated according to the model $y_i = f^\star(x_i) + \epsilon_i$, for all $i \in [n]$. Here, $f^\star \in \mathcal{F}$ represents the (unknown) *optimal* target function, and ϵ_i, for $i \in [n]$, are independent noise terms.

In this paper, we study the problem of learning under invariances (or symmetries). Specifically, consider an L-Lipschitz action of a topological group G on $\mathcal{X}$, meaning that each group element corresponds to an L-Lipschitz bijection on $\mathcal{X}$. Moreover, assume that the optimal target function is invariant under this group action, meaning $f^\star(x) = f^\star(gx)$ for all $g \in G$ and $x \in \mathcal{X}$. Here, gx represents the action of the group element g on x, and for technical reasons, we assume that the map $\theta : G \times \mathcal{X} \to \mathcal{X}$ defined by $\theta(g, x) = gx$ is continuous, where $G \times \mathcal{X}$ is equipped with the product topology.

However, the class of target functions $\mathcal{F}$ is not always G-invariant in practice and may need to be adjusted to ensure G-invariance. For instance, $\mathcal{F}$ might be represented using deep neural networks or kernel methods, which can include functions that are not invariant. To address this, various approaches have been developed to create invariant function spaces from a more general function space $\mathcal{F}$ (i.e., the class of base functions) in the context of learning with invariances. This includes techniques such as feature averaging (Puny et al., 2022) and canonicalization (Ma et al., 2024a), with the latter being the primary focus of this paper.

The idea of canonicalization is first to project data on the *quotient space* of the group action and then apply an arbitrary base function $f \in \mathcal{F}$. In particular, define the quotient space $\mathcal{X}/G := \{[x] : x \in \mathcal{X}\}$, where $[x] := \{gx : g \in G\}$ denotes the orbit of a point $x \in \mathcal{X}$. Then, consider the projection map $\pi : \mathcal{X} \to \mathcal{X}/G$ and assume that $\mathcal{X}/G$ is identified as a subset[1] of $\mathcal{X}$ so we might assume $\mathcal{X}/G \subseteq \mathcal{X}$. For technical reasons, throughout this paper, we assume that at least one of the following conditions holds:

- The group G is compact.
- The set $\mathcal{X}/G$ (the quotient of $\mathcal{X}$ by the action of G) is precompact, meaning its closure is compact in $\mathcal{X}$. Additionally, for every element x in $\mathcal{X}$, the orbit of x under G, denoted by $[x]$, is a closed set in $\mathcal{X}$.

The conditions ensure that the mathematical structures we work with have well-behaved topological properties, which are essential for achieving the main results of the paper. The *canonicalization scheme*, denoted as $(\pi, \mathcal{X}/G)$, proposes using target functions of the form $f \circ \pi(x)$.

Definition 1. The class of *canonicalized* target functions $\mathcal{F}_{\text{can}}$ is defined as:

$$\mathcal{F}_{\text{can}} := \big\{ f \circ \pi : \mathcal{X} \to \mathbb{R} \mid f \in \mathcal{F} \big\}, \tag{1}$$

where $f \circ \pi$ denotes the composition of the two functions.

Figure 1 illustrates the canonicalization method used for constructing invariant function classes. Note that any $f_{\text{can}} \in \mathcal{F}_{\text{can}}$ is G-invariant: $f_{\text{can}}(gx) = f(\pi(gx)) = f(\pi(x)) = f_{\text{can}}(x)$ for all x.

3.2 Continuity, Smoothness, and Regularization

Given the projection map $\pi : \mathcal{X} \to \mathcal{X}/G$ and noting that $\mathcal{X}$ is a topological space, note that π is continuous if we endow $\mathcal{X}/G$ with the quotient topology, which is arguably the most natural topology to choose for $\mathcal{X}/G$. However, when $\mathcal{X}/G$ is embedded as a subset of $\mathcal{X}$, the function π can be discontinuous! More precisely, when $\mathcal{X}/G$ is identified as a subset of $\mathcal{X}$, as we always do in this paper, the mapping $\pi : \mathcal{X} \to \mathcal{X}/G$ is not necessarily continuous (with respect to the topology of $\mathcal{X}$). Here is an example (see Dym et al. (2024) for reference):

Example 1 (Discontinuous canonicalization)**.** Consider the action of the set of integers $G = \mathbb{Z}$ on

[1] Although the quotient space can be studied as an abstract topological space, this paper focuses on a more concrete setting where it is embedded into a subset of the input space $\mathcal{X}$ through a surjection.

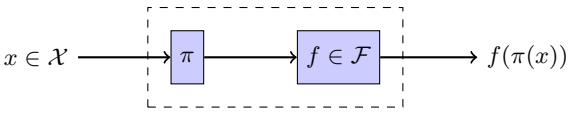

Figure 1: Canonicalization.

$\mathcal{X} = \mathbb{R}$ via shifting. The quotient space can be simply embedded into the input space, e.g., $\mathcal{X}/G = [0, 1)$. However, the projection map $\pi : \mathbb{R} \to [0, 1)$, given by $\pi(x) = x - [x]$, is clearly not continuous (with respect to $\mathbb{R}$ topology), as one can see $\lim_{x\to 1^-} \pi(x) = 1 \neq \pi(1) = 0$. Consequently, even for a continuous base function such as $f(x) = x$, the final end-to-end function is $f(\pi(x)) = x - [x]$, which is discontinuous on $\mathcal{X} = \mathbb{R}$.

We note that in designing machine learning architectures, it is always desirable to have continuous functions and, in general, be able to regularize the loss function up to a specified order of derivatives. To do that, we first need to make sure that the class of canonicalized functions $\mathcal{F}_{\text{can}}$ only includes continuous functions, or in general, it only includes functions having derivatives up to a specified order.

Regularity plays a crucial role in obtaining sample complexity bounds for learning under invariances. Consider the Sobolev space of base functions defined on a manifold input space of dimension d. The Sobolev space, with parameter s denoting the order of square-integrable derivatives, is a Reproducing Kernel Hilbert Space (RKHS) if and only if $s > d/2$. Since kernel methods can be applied in an RKHS, Sobolev spaces with $s > d/2$ exhibit favorable convergence rates for the generalization error in (kernel) regression. Moreover, group averaging further accelerates convergence, resulting in improved sample complexity, as demonstrated, for instance, in (Tahmasebi and Jegelka, 2023).

However, when canonicalization is used, the kernel generalization theory breaks down because the space of canonicalized functions $f(\pi(x))$ is no longer continuous, let alone an RKHS. This implies that the convergence theory of kernel methods is not applicable to canonicalization. Nevertheless, if one ensures continuous derivatives up to order k for the end-to-end canonicalized models, this guarantees that the resulting functions lie within an RKHS of a Sobolev space with the appropriate s (using, for example, Sobolev inequalities). Thus, regularity facilitates the construction of an RKHS of invariant functions, which, in turn, enables provable generalization bounds via kernel methods. This explanation highlights how regularity enables the application of recent convergence proofs for group averaging (see, for example, (Tahmasebi and Jegelka, 2023)) to canonicalization as well.

Thus, we ask the following question:

> What are the minimum requirements for the class of base functions $\mathcal{F}$ to ensure that the end-to-end representation of canonicalized functions $f \circ \pi \in \mathcal{F}_{\text{can}}$ is *always* continuous (or continuously differentiable up to a given order k)?

Note that the conditions imposed on $\mathcal{F}$ should be minimal. For instance, overly broad conditions could be trivially satisfied by degenerate function classes, such as the set of constant functions on $\mathcal{X}$. Additionally, we are seeking to impose constraints on the class of base functions while assuming that the canonicalization scheme $(\pi, \mathcal{X}/G)$ is fixed.

The following examples motivate the need for specific constraints to obtain continuous/smooth canonicalized functions and show having continuous/smooth base functions $f \in \mathcal{F}$ is not enough.

Example 2 (Continuity conditions)**.** Consider the same setup as in Example 1 and observe that for $f(\pi(x)) = f(x - [x])$ to be continuous on $\mathbb{R}$, it is necessary to assume that f is continuous on $[0, 1]$ and further satisfies $f(0) = f(1)$. The continuity of f is required by considering *generic* points in the quotient space, while the condition $f(0) = f(1)$ arises from examining the behavior at integer points.

Example 3 (Differentiability conditions)**.** Consider reflecting across the origin with $\mathcal{X} = \mathbb{R}$. Note that the projection $\pi(x) = |x|$ maps onto the quotient space $\mathcal{X}/G = [0, \infty)$. Any continuous function $f : \mathbb{R} \to \mathbb{R}$ results in a continuous canonicalized function $f(\pi(x)) = f(|x|)$. However, even if f is differentiable on $\mathbb{R}$, the canonicalized function $f(\pi(x)) = f(|x|)$ may not be differentiable. For instance, with $f(x) = x$, the function $f(\pi(x)) = |x|$ is not differentiable at the origin. To ensure differentiability at the origin,

one must check:

$$\lim_{x\to 0^+} \frac{f(|x|)}{x} = f'(0), \tag{2}$$

$$\lim_{x\to 0^-} \frac{f(|x|)}{x} = \lim_{x\to 0^-} \frac{f(-x)}{x} = -f'(0), \tag{3}$$

which implies that $f'(0)$ must be zero.

Example 4 (Smoothness conditions)**.** Consider the same setup as in Example 3. How can we ensure second-order differentiability? Let us assume that f is twice differentiable and note that the first-order derivative of the canonicalized function $f(|x|)$ is[2] $\text{sign}(x)f'(|x|)$ which is continuous if $f'(0) = 0$. Now observe that

$$\lim_{x\to 0^+} \frac{\text{sign}(x)f'(|x|)}{x} = \lim_{x\to 0^-} \frac{\text{sign}(x)f'(|x|)}{x} = f''(0).$$

In other words, the canonicalized function $f(|x|)$ is twice differentiable *for free*, and we don't need any additional conditions to achieve second-order differentiability. More generally, one can simply show that to obtain k-th order differentiability, we need the following conditions:

$$\forall \ell = 1, 2, \ldots, \left\lceil \frac{k}{2} \right\rceil : f^{(2\ell-1)}(0) = 0, \tag{4}$$

in addition to f being differentiable up to order k.

Driven by the examples discussed, this paper provides a comprehensive characterization of the necessary and sufficient conditions for obtaining continuous and smooth canonicalized function classes with arbitrary differentiability orders.

4 MAIN RESULTS

To present the main results, we need to introduce a few definitions in the following subsection.

4.1 Preliminaries

Let $\mathcal{B}(x;\epsilon) \subseteq \mathcal{X}$ denote the (open) ball around $x \in \mathcal{X}$ with radius ϵ. Let $\Pi(x;\epsilon)$ denote the image of $\mathcal{B}(x;\epsilon)$ under the mapping π, defined as

$$\Pi(x;\epsilon) := \Big\{\pi(\xi) : \xi \in \mathcal{B}(x;\epsilon)\Big\} \subseteq \mathcal{X}/G. \tag{5}$$

Let us emphasize that, unless specified otherwise, in this paper, the notation $\mathcal{X}/G$ will denote the embedded version of the quotient space with respect to the topology of $\mathcal{X}$.

[2]Define $\text{sign}(x) = x/|x|$ for $x \neq 0$ and $\text{sign}(0) = 0$.

Let $\overline{\Pi(x;\epsilon)}$ denote the closure of $\Pi(x;\epsilon)$ in $\mathcal{X}$. Note that $\Pi(x;\epsilon)$ may not be a closed set. For example, in Example 1, we have $\Pi(0;1) = [0,1)$ which is not closed in $\mathcal{X} = \mathbb{R}$.

The (topological) boundary of the quotient space in $\mathcal{X}$ is defined as $\partial(\mathcal{X}/G) := \overline{\mathcal{X}/G} \setminus (\mathcal{X}/G)^\circ$, where $\overline{\mathcal{X}/G}$ and $(\mathcal{X}/G)^\circ$ denote the closure and the interior of the quotient space with respect to the topology of $\mathcal{X}$, respectively.

Next, we define the equivalence class of any point $x \in \mathcal{X}$.

Definition 2 (Equivalence classes)**.** For any $x \in \mathcal{X}$, define

$$\Pi(x) := \bigcap_{\epsilon>0} \overline{\Pi(x;\epsilon)}. \tag{6}$$

Note that according to the above definition, one has $\pi(x) \in \Pi(x)$. Therefore, the equivalence classes of points are always non-empty. Moreover, it is an invariant set under the group action, and oftentimes, $\pi(x)$ is the only element of this set:

Proposition 1. *The following statements hold:*

- *For any $x \in \mathcal{X}$ and $g \in G$, one has $\Pi(gx) = \Pi(x)$.*
- *If $x \in (\mathcal{X}/G)^\circ$, then $\Pi(x) = \{x\}$.*
- *For any $x \in \mathcal{X}$, one has $\Pi(x) \subseteq [x]$.*
- *If $\mathcal{X}/G$ is closed in $\mathcal{X}$, then $\Pi(x) = \{x\}$ for any $x \in \mathcal{X}/G$.*
- *$\Pi(x) \subseteq \mathcal{X}$ is compact, for any $x \in \mathcal{X}$.*

We present the proof of Proposition 1 in Appendix A. To examine the continuity of canonicalized functions, we need to analyze the set of points $x \in \mathcal{X}$ that have a non-trivial equivalence class, referred to as critical points.

Definition 3 (Critical points)**.** The set of critical point $\Pi^\circ \subseteq \mathcal{X}$ is defined as follows:

$$\Pi^\circ := \Big\{x \in \mathcal{X}/G \subseteq \mathcal{X} : \Pi(x) \neq \{x\}\Big\}. \tag{7}$$

Observe that $\Pi^\circ \subseteq \partial(\mathcal{X}/G)$, according its definition and Proposition 1. Intuitively, for any critical point $x \in \mathcal{X}/G$, there are other points such as $\hat{x} \in \mathcal{X}$ such that the image of sequences converging to x (under the quotient map) can converge to them.

4.2 Continuity Conditions

We observed in Example 2 that achieving continuity may demand non-trivial assumptions on base functions. Next, we use the concept of critical points to establish the necessary and sufficient conditions for the continuity of canonicalized functions. The main result of this subsection is the following theorem.

Theorem 1 (Continuous canonicalized functions)**.** *Consider a canonicalized function class $\mathcal{F}_{\text{can}}$ derived from a set of continuous base functions $\mathcal{F}$. The class $\mathcal{F}_{\text{can}}$ contains only continuous functions if and only if for any base function $f \in \mathcal{F}$ the following condition holds:*

$$\forall \hat{x} \in \Pi(x): \quad f(\hat{x}) = f(x), \tag{8}$$

for every critical point $x \in \Pi^{\circ}$.

We present the proof of Theorem 1 in Appendix B.

Remark 2. According to the above theorem, having continuous base functions $\mathcal{F}$ alone is insufficient to ensure continuity. However, if each base function $f \in \mathcal{F}$ produces the same values at the equivalence classes of critical points, then the resulting function space consists only of continuous functions. Notably, this represents the minimal condition required for continuity, as stated by the theorem. We examine the results of Theorem 1 in several applications in Section 5.

4.3 Smoothness Conditions

Example 3 shows that achieving higher-order differentiability requires more conditions than continuity. Here, we establish necessary and sufficient conditions for the continuous differentiability of canonicalized functions up to a given order k. First, let us introduce some notation.

Notation. Throughout this subsection, G denotes a finite matrix group[3] acting faithfully[4] on $\mathcal{X}$, and for any $g \in G$, let $D(g) \in \mathbb{R}^{d\times d}$ denote its matrix representation, so $gx = D(g)x$ for any $x \in \mathcal{X}$, and assume that $D(g)^{\mathsf{T}} = D(g)$. Let $\nabla^k f : \mathcal{X} \to \mathbb{R}^{d^k}$ denote the k-th order tensor representing the k-th order derivatives of a function $f : \mathcal{X} \to \mathbb{R}$. We treat this object as a vector in this paper. Finally, remember that for any $\hat{x} \in \Pi(x)$, we have $\hat{x} \in [x]$. Let us define $\Xi(x) := \{(g, \hat{x}) \in G \times \Pi(x) : \hat{x} = gx\}$.

[3]The results in this subsection can also be extended to infinite groups, but for clarity and simplicity, we focus on smoothness for finite groups.

[4]A group action is considered faithful if the only group element that acts trivially on the space is the identity element.

Theorem 3 (Smooth canonicalized functions)**.** *Consider a canonicalized function class $\mathcal{F}_{\text{can}}$ derived from a set of k-th order continuously differentiable base functions $\mathcal{F}$. The class $\mathcal{F}_{\text{can}}$ contains only k-th order continuously differentiable functions if and only if for any base function $f \in \mathcal{F}$, the following condition holds:*

$$\forall (g, \hat{x}) \in \Xi(x): \quad \nabla^{\ell} f(x) = D(g)^{\otimes \ell} \nabla^{\ell} f(\hat{x}), \tag{9}$$

for all $0 \le \ell \le k$ and any point $x \in \partial(\mathcal{X}/G)$, where $D(g)^{\otimes \ell}$ denotes the ℓ-times tensor product of the matrix $D(g)$ with itself.

We present the proof of Theorem 3 in Appendix C.

Remark 4. The theorem above describes the minimal conditions required, formulated as several linear constraints on the boundary points of the quotient space, to guarantee that the canonicalized function class is continuously differentiable up to the k-th order. We examine the results of Theorem 3 in several applications in Section 5.

Remark 5. To ensure continuity, we previously conditioned the values of the base function on the critical points. However, to achieve k-th order differentiability, restricting only the critical points is insufficient. Specifically, we need to impose conditions on *all* boundary points $x \in \partial(\mathcal{X}/G)$ to achieve smoothness. We will further explain this distinction in the next section.

Remark 6. Note that our results regarding the differentiability of canonicalized functions are obtained in a general setting of metric spaces and Lipschitz group actions. However, to define and achieve smoothness, we require more structure than just metric spaces. Specifically, we consider embedded input spaces $\mathcal{X} \subseteq \mathbb{R}^d$ and finite matrix group actions, both for the clarity of the paper's presentation and due to their numerous applications. Nonetheless, these results can also be extended to infinite non-linear group actions using similar techniques, provided the appropriate conditions are met.

5 EXAMPLES AND APPLICATIONS

In this section, we evaluate the main results through several examples and applications.

5.1 Sort

Consider the action of the permutation group $G = \mathcal{S}_d$ on vectors $x \in \mathcal{X} = \mathbb{R}^d$ defined by $\sigma x :=$

$\big(x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_d}\big)^\mathsf{T}$ for any $\sigma \in \mathcal{S}_d$. The most natural canonicalization scheme on this space is the *sort* function: $\pi(x) := \big(x_{\min}, \ldots, x_{\max}\big)^\mathsf{T}$. The quotient space is then given by:

$$\mathcal{X}/G = \big\{x \in \mathbb{R}^d \mid \forall i \in [d-1] : x_{i+1} \geq x_i\big\}.$$

Moreover, the boundary of the quotient space can be expressed as:

$$\partial(\mathcal{X}/G) = \big\{x \in \mathcal{X}/G \mid \exists i \in [d-1] : x_{i+1} = x_i\big\}.$$

Since $\mathcal{X}/G$ is closed in $\mathcal{X}$, by Proposition 1, there are no critical points in this case. Indeed, the sort function is continuous, and $\Pi(x) = \{x\}$ for $x \in \mathcal{X}/G$.

Now, consider continuous (first-order) differentiability. According to Theorem 3, we need the following condition to hold for any base function $f \in \mathcal{F}$ and any $x \in \partial(\mathcal{X}/G)$:

$$\nabla f(x) = D(g)\nabla f(x) \quad \text{for any } g \in G_x, \tag{10}$$

where $G_x := \{g \in G : gx = x\}$ is the *stabilizer* of the group action at the point x. Rearranging this condition, it follows that:

$$\forall i, j \in [d] : x_i = x_j \implies \partial_i f(x) = \partial_j f(x), \tag{11}$$

for any $x \in \partial(\mathcal{X}/G)$. Since the elements of $\partial(\mathcal{X}/G)$ are sorted, this condition is equivalent to:

$$\forall i \in [d-1] : x_{i+1} = x_i \implies \partial_{i+1} f(x) = \partial_i f(x), \tag{12}$$

for any $x \in \partial(\mathcal{X}/G)$. This provides the necessary and sufficient conditions for the continuous (first-order) differentiability of the canonicalized model.

Let us further evaluate this result to a few cases.

Linear functions. A linear function $f : \mathbb{R}^d \to \mathbb{R}$ can be written as $f(x) = \sum_{i=1}^d a_i x_i$. Note that we have $\partial_i f(x) = a_i$ for each $i \in [d]$. Thus, in order to use linear functions to achieve differentiable end-to-end canonicalization, one needs to have $a_{i+1} = a_i$ for any $i \in [d-1]$. In other words, the function must be of the form $f(x) = a\sum_{i=1}^d x_i$. This shows that, differentiability condition can reduce the dimension of the space of functions drastically; here from d to one. Moreover, note that in this case such $f(x)$ are already invariant, thus for linear functions there is no difference in canonicalization or fully invariant learning on whole space if we require differentiability.

Quadratic functions. Now let us consider the space of quadratic polynomials $f(x) = \frac{1}{2}x^\mathsf{T}Ax + b^\mathsf{T}x$ for $x \in \mathbb{R}^d$, and parameters $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$. For simplicity, that A is a symmetric matrix.

Then, we have

$$\partial_i f(x) = b_i + a_{ii}x_i + \frac{1}{2}\sum_{j\neq i} a_{ij}x_j.$$

Therefore, to satisfy the differentiability conditions, we need to have $b_i = b$ for all $i \in [d]$ and

$$a_{ij} = \begin{cases} a & \text{if } i = j, \\ 2a & \text{if } |i-j| = 1, \\ c & \text{otherwise.} \end{cases}$$

Here, $a, b, c \in \mathbb{R}$ are three parameters. In this case, the function $f(x)$ is *not* permutation invariance, as opposed to linear functions. However, we still lose a lot of flexibility by forcing to choose among just three parameters instead of the general case with $O(d^2)$ parameters.

Dot-product kernels. A dot-product kernel is of the form $f\big(\langle\tilde{x}, x\rangle\big)$, $\tilde{x}, x \in \mathbb{R}^d$, for an appropriate function f. For example, the Gaussian kernel is a dot-product kernel. To see if kernel feature functions $f\big(\langle\tilde{x}, .\rangle\big)$ satisfy the condition, note

$$\partial_i f(x) = \tilde{x}_i f\big(\langle\tilde{x}, x\rangle\big),$$

thus demanding $\tilde{x} \in \mathbb{R}^d$ being a scalar multiple of the all-ones vectors. In other words, dot-product kernels produce only differentiable canonicalized models, if and only if they are evaluated over scalar multiples of the all-ones vector, an assumption that can barely happen.

5.2 Rotation

Consider the unit circle

$$\mathbb{S}^1 = \big\{\big(\cos(\theta), \sin(\theta)\big) \in \mathbb{R}^2 : \theta \in \mathbb{R}\big\}$$

as the input space. The group of rotations by integer multiples of $\frac{2\pi}{|G|}$ radians about the origin acts on the input space. The critical points of the canonicalization scheme $\pi(\theta) \in \Big[0, 2\pi/|G|\Big)$ are $\theta = 0$ and $\theta = 2\pi/|G|$. According to Theorem 1, the canonicalized functions are continuous if $f(0) = f(2\pi/|G|)$ for any base function $f \in \mathcal{F}$.

5.3 Tori

The flat torus $\mathbb{T}^d$ is defined as the quotient of $\mathbb{R}^d$ modulo the action of integer translations $\mathbb{Z}^d$, i.e., $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. Consider the canonicalization scheme

$$\pi(x) = \Big(x_1 - [x_1], x_2 - [x_2], \ldots, x_d - [x_d]\Big) \in [0, 1)^d,$$

where $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$. The critical points of this scheme are any $x \in [0,1]^d$ such that $x_i = 1$ for some $i \in [d]$. The necessary conditions for minima to ensure continuity are given by $f(\tilde{x}) = f(x)$ for any critical point x, where $\tilde{x}$ differs from x only by having a number of its coordinates flipped from one to zero. Even the discretized version of this requires exponentially many functional equations to hold, which is prohibitive.

5.4 Permutation Invariance

Consider sets of size n consisting of points in $\mathbb{R}^d$, also known as point clouds. Any point cloud can be represented as an element $x \in \mathbb{R}^{d\times n}$, which is invariant under permutations of its columns. A natural canonicalization scheme for this space is given by lexicographic sorting: one first sorts the columns of x based on the elements in the first row, as in Section 5.1. For columns that are identical in the first row, the sorting continues based on the second row, and so on.

To achieve continuity, by Theorem 1, we must have $f(\tilde{x}) = f(x)$ for any $\tilde{x}$ and x whose columns are permutations of each other and which satisfy the following condition:

$$x_{11} \leq x_{12} \leq \ldots \leq x_{1n} \quad \text{and} \quad \tilde{x}_{11} \leq \tilde{x}_{12} \leq \ldots \tilde{x}_{1n}.$$

A simple count shows that this requires exponentially many conditions on n, which is prohibitive.

5.5 Sign Invariance

Consider the action of the two-element group G on $\mathbb{R}^d$, which flips the sign of all coordinates: $v \mapsto -v$ for any $v \in \mathbb{R}^d$. Such symmetries commonly arise when learning with spectral data (i.e., eigenvectors). A canonicalization scheme for sign invariance is $\pi(v) = \pm v$, where the sign is chosen such that the first nonzero coordinate of $\pi(v)$ is positive.

The critical points of this scheme are vectors $v \in \mathbb{R}^d$ such that $0 = v_1 = v_2 = \cdots = v_i < v_{i+1}$ for some $i \in [d-1]$, or $v = 0$. For such v, we must have $f(v) = f(-v)$. In other words, this shows that continuously learning canonicalized eigenvectors $v \in \mathbb{R}^d$ is equivalent to learning vectors $v \in \mathbb{R}^{d-1}$ with sign invariance. This example demonstrates that sign canonicalization may be of limited use when continuity of the end-to-end representation is required, as the continuity condition necessitates solving an almost identical sign-invariance problem.

6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we investigate the continuity and smoothness of canonicalized models. We present a thorough analysis of the problem and identify the minimal conditions required for base functions to satisfy these properties. Our theoretical results offer insight into the feasibility of enforcing these conditions through regularization. We provide examples demonstrating the feasibility of smooth sorting for arrays and point clouds with permutation invariance. Additionally, we show that for canonicalized models of sign invariance, the conditions are nearly equivalent to achieving full sign invariance, indicating that regularization is only marginally applicable in this case.

For future work, the possibility of extending the results to spaces of equivariant functions can be explored. Moreover, our results indicate the following: classical theory, including that for kernel methods, falls short in explaining the success of canonicalization for generalizing specific tasks. This highlights the need for new approaches to understand generalization in these contexts, a conclusion that also points to future directions for investigating generalization in such settings using kernel methods.

Another direction for future work could involve extending the results to infinite groups. In cases with infinite groups, the conclusion preceding Lemma 2 no longer holds, meaning the functional identity $f_{\text{CAN}}(\xi) = f(D(g)\xi)$ does not necessarily hold locally around each ξ (provided ξ is not on the boundary of $\mathcal{X}/G$). Specifically, in such cases, we have $f_{\text{CAN}}(\xi) = f(D(g_\xi)\xi)$, where g_ξ (under certain conditions) depends smoothly on ξ. Differentiating with respect to ξ introduces non-trivial terms that depend on the derivative(s) of $D(g_\xi)$ with respect to ξ, which do not appear in the cases involving finite groups. To address this, these additional terms (arising from the chain rule) must be incorporated into Equation (9) to ensure that a modified version of Theorem 3 holds in the context of infinite groups.

ACKNOWLEDGEMENTS

This work was supported by the NSF AI Institute TILOS, ONR grant N00014-20-1-2023 (MURI ML-SCOPE), and an Alexander von Humboldt fellowship.

References

Anselmi, F., Rosasco, L., and Poggio, T. (2016). On invariance and selectivity in representation learning. *Information and Inference: A Journal of the IMA*, 5(2):134–158. 1, 3

Baker, J., Wang, S.-H., de Fernex, T., and Wang, B. (2024). An explicit frame construction for normalizing 3d point clouds. In *Int. Conference on Machine Learning (ICML)*. 3

Batzner, S., Musaelian, A., and Kozinsky, B. (2023). Advancing molecular simulation with equivariant interatomic potentials. *Nature Reviews Physics*, 5(8):437–438. 1

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453. 1

Bouvrie, J., Rosasco, L., and Poggio, T. (2009). On invariance in hierarchical models. *Advances in Neural Information Processing Systems*, 22. 1, 3

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42. 1, 3

Dym, N. and Gortler, S. J. (2024). Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, pages 1–41. 3

Dym, N., Lawrence, H., and Siegel, J. W. (2024). Equivariant frames and the impossibility of continuous canonicalization. In *Int. Conference on Machine Learning (ICML)*. 2, 3, 4

Grisafi, A., Wilkins, D. M., Csányi, G., and Ceriotti, M. (2018). Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Physical review letters*, 120(3):036002. 1

Hinton, G. E. (1987). Learning translation invariant recognition in a massively parallel networks. In *International conference on parallel architectures and languages Europe*, pages 1–13. Springer. 1, 3

Huang, Y., Lu, W., Robinson, J., Yang, Y., Zhang, M., Jegelka, S., and Li, P. (2024). On the stability of expressive positional encodings for graphs. In *Int. Conference on Learning Representations (ICLR)*. 3

Kaba, S.-O., Mondal, A. K., Zhang, Y., Bengio, Y., and Ravanbakhsh, S. (2023). Equivariance with learned canonicalization functions. In *Int. Conference on Machine Learning (ICML)*. 1, 3

Kiani, B., Le, T., Lawrence, H., Jegelka, S., and Weber, M. (2024). On the hardness of learning under symmetries. In *Int. Conference on Learning Representations (ICLR)*. 3

Kim, J., Nguyen, D., Suleymanzade, A., An, H., and Hong, S. (2023). Learning probabilistic symmetrization for architecture agnostic equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3

Lin, Y., Helwig, J., Gui, S., and Ji, S. (2024). Equivariance via minimal frame averaging for more symmetries and efficiency. In *Int. Conference on Machine Learning (ICML)*. 3

Ma, G., Wang, Y., Lim, D., Jegelka, S., and Wang, Y. (2024a). A canonization perspective on invariant and equivariant learning. *arXiv preprint arXiv:2405.18378*. 3, 4

Ma, G., Wang, Y., and Wang, Y. (2024b). Laplacian canonization: A minimalist approach to sign and basis invariant spectral embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3

Murphy, R., Srinivasan, B., Rao, V., and Ribeiro, B. (2019). Relational pooling for graph representations. In *Int. Conference on Machine Learning (ICML)*. 1, 3

Panigrahi, S. S. and Mondal, A. K. (2024). Improved canonicalization for model agnostic equivariance. *arXiv preprint arXiv:2405.14089*. 3

Poggio, T. A. and Anselmi, F. (2016). *Visual cortex and deep networks: learning invariant representations*. MIT press. 1, 3

Pozdnyakov, S. and Ceriotti, M. (2024). Smooth, exact rotational symmetrization for deep learning on point clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3

Puny, O., Atzmon, M., Smith, E. J., Misra, I., Grover, A., Ben-Hamu, H., and Lipman, Y. (2022). Frame averaging for invariant and equivariant network design. In *Int. Conference on Learning Representations (ICLR)*. 1, 3, 4

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3

Smidt, T. E. (2021). Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry*, 3(2):82–85. 1, 3

Tahmasebi, B. and Jegelka, S. (2023). The exact sample complexity gain from invariances for kernel regression. *Advances in Neural Information Processing Systems*, 36. 5

Unke, O., Bogojeski, M., Gastegger, M., Geiger, M., Smidt, T., and Müller, K.-R. (2021). Se(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1

Zhang, Y., Hare, J., and Prügel-Bennett, A. (2019). Fspool: Learning set representations with featurewise sort pooling. *arXiv preprint arXiv:1906.02795*. 3

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Proof of Proposition 1

Proof. We prove the four parts of the proposition below.

Part 1. Fix any $x \in \mathcal{X}$ and $g \in G$, and assume that $\hat{x} \in \Pi(x)$. We need to show that $\hat{x} \in \Pi(gx)$, and this completes the proof. Note that, according to the definition, $\hat{x} \in \overline{\Pi(x;\epsilon)}$ for all $\epsilon > 0$. Thus, for a fixed ϵ, there exists a sequence $\xi_i \in \mathcal{X}$, $i \in \mathbb{N}$, such that $\operatorname{dist}(\xi_i, x) < \epsilon$ and $\pi(\xi_i) \to \hat{x}$. Therefore, for the sequence $g\xi_i$, $i \in \mathbb{N}$, one has $\operatorname{dist}(g\xi_i, gx) \leq L \operatorname{dist}(\xi_i, x) < L\epsilon$, since the mapping $\xi \mapsto g\xi$ is L-Lipschitz by assumption.

Furthermore, observe that $\pi(g\xi_i) = \pi(\xi_i) \to \hat{x}$. Therefore, we conclude that $\hat{x} \in \overline{\Pi(gx; L\epsilon)}$. This holds for any ϵ, and thus, by definition, $\hat{x} \in \Pi(gx)$, which completes the proof.

Part 2. Now, we prove the second part of the proposition. If $x \in (\mathcal{X}/G)^\circ$, then there exists positive ϵ_0 such that $\mathcal{B}(x;\epsilon_0) \subset \mathcal{X}/G$. However, this means that $\Pi(x;\epsilon_0) = \mathcal{B}(x;\epsilon_0)$ since the mapping π acts trivially on the quotient space. Therefore, according to the definition, we have

$$\Pi(x) = \bigcap_{\epsilon>0} \overline{\Pi(x;\epsilon)} = \bigcap_{\epsilon_0>\epsilon>0} \overline{\mathcal{B}(x;\epsilon)} = \{x\}, \tag{13}$$

and this completes the proof.

Part 3. Fix an arbitrary $x \in \mathcal{X}$ and let $\hat{x} \in \Pi(x)$. Note that, according to the definition, for any ϵ, there exists a sequence $\xi_{i,\epsilon} \in \mathcal{X}$, $i \in \mathbb{N}$, such that $\operatorname{dist}(\xi_{i,\epsilon}, x) < \epsilon$ and $\pi(\xi_{i,\epsilon}) \to \hat{x}$. This means that for any $j \in \mathbb{N}$, there exists $N(j) \in \mathbb{N}$, such that $|\pi(\xi_{i,1/j}) - \hat{x}| \leq 1/j$ for any $i \geq N(j)$. Now consider the sequence $\xi_i = \xi_{N(i),1/i}$. Note that $\xi_i \to x$ and $\pi(\xi_i) \to \hat{x}$.

But, there exists a sequence $g_i \in G$, such that $\pi(\xi_i) = g_i\xi_i$, so we have $g_i\xi_i \to \hat{x}$. Let us first consider the case that G is a compact group. This means that, by passing to a subsequence, there exists $g \in G$ such that $g_i \to g \in G$, in the topology of G. Thus, for any i,

$$\begin{aligned} \operatorname{dist}\big(g_i\xi_i, gx\big) &\leq \operatorname{dist}\big(g_i\xi_i, g_ix\big) + \operatorname{dist}\big(g_ix, gx\big) && (14)\\ &\leq L \operatorname{dist}\big(\xi_i, x\big) + \operatorname{dist}\big(g_ix, gx\big), && (15)\end{aligned}$$

where in the last step, we used the Lipschitzness of the group action. Note that $\operatorname{dist}(\xi_i, x) \to 0$ by assumption, and also, from $g_i \to g$, we have $\operatorname{dist}(g_ix, gx) \to 0$. Thus, we conclude that $g_i\xi_i \to gx$, and this means that $gx = \hat{x}$, or $\hat{x} \in [x]$, which completes the proof.

Now consider the case where the group G is not compact. We can use the same argument as in the case of compact groups, and with the same notation, we only need to show that $\operatorname{dist}(g_ix, gx) \to 0$ for some $g \in G$.

Note that $\xi_i \to x$ and $g_i\xi_i \to \hat{x}$. Thus, $g_ix \to \hat{x}$, using the Lipschitzness of the group action:

$$\operatorname{dist}\big(g_ix, \hat{x}\big) \leq \operatorname{dist}\big(g_ix, g_i\xi_i\big) + \operatorname{dist}\big(g_i\xi_i, \hat{x}\big) \leq L \operatorname{dist}\big(x, \xi_i\big) + \operatorname{dist}\big(g_i\xi_i, \hat{x}\big) \to 0. \tag{16}$$

This means that for sufficiently large $i \in \mathbb{N}$, g_ix belongs to a compact set A since $\mathcal{X}$ is locally compact. Note that, according to the assumption, $[x]$ is a closed set, and thus[5] $[x] \cap A$ is compact. Since $g_ix \in [x] \cap A$ for

[5]The intersection of a closed set and a compact set is compact.

all sufficiently large i, by passing to a subsequence, we conclude that $g_i x \to gx \in [x]$ for some $g \in G$, which completes the proof.

Part 4. Note that $\mathcal{X}/G$ is the image of π. Therefore, by the closedness of $\mathcal{X}/G$, we have $\Pi(x) = \bigcap_{\epsilon>0} \overline{\Pi(x;\epsilon)} \subseteq \mathcal{X}/G$ for any $x \in \mathcal{X}/G$. However, from the previous part of the proposition, we know that $\Pi(x) \subseteq [x] \ni x$ for any $x \in \mathcal{X}/G$. Combining these results, we conclude that $\Pi(x) = \{x\}$ for all $x \in \mathcal{X}/G$, thus completing the proof.

Part 5. Fix an arbitrary $x \in \mathcal{X}$. Note that since the function $\theta(g, x) = gx$ is continuous, it maps compact sets to compact sets. Also, since $\mathcal{X}$ is locally compact, for sufficiently small ϵ, the ball $\mathcal{B}(x;\epsilon)$ is precompact. If G is a compact group, then the image of $G \times \mathcal{B}(x;\epsilon)$ under θ, denoted by B, is also precompact. Since $\Pi(x;\epsilon) \subseteq B$, it follows that $\Pi(x;\epsilon)$ is precompact. Given that $\Pi(x)$ is the intersection of infinitely many closed sets and is a subset of a precompact set, it is compact. This completes the proof.

Now assume that G is not compact. According to the assumption, $\mathcal{X}/G$ is precompact, and note that $\Pi(x;\epsilon) \subseteq \mathcal{X}/G$ for all ϵ. Therefore, $\Pi(x;\epsilon)$ is also precompact. Similar to the previous case, the proof is complete.

□

B Proof of Theorem 1

Proof. We divide the proof into two parts.

Part 1 ($\Longrightarrow$). Fix an arbitrary point $x \in \mathcal{X}$ and a function $f \in \mathcal{F}$. Consider an arbitrary sequence $\xi_i \in \mathcal{X}$, $i \in \mathbb{N}$, such that $\xi_i \to x$. To demonstrate the continuity of the canonicalized function at x, we need to show that $f(\pi(\xi_i)) \to f(\pi(x))$. Without loss of generality, assume that $\pi(x) = x$. This is due to the invariance of the rest of the proof under any group transformation.

By definition, we can assume that there exists a sequence of positive reals ϵ_i, $i \in \mathbb{N}$, such that $\epsilon_i \to 0$ and $\xi_i \in \mathcal{B}(x, \epsilon_i)$ for all $i \in \mathbb{N}$. Consequently, $\pi(\xi_i) \in \overline{\Pi(x;\epsilon_i)}$ for all i. We will use the following lemma.

Lemma 1. *Let $A_1 \supseteq A_2 \supseteq \ldots$ be a nested sequence of closed sets in a topological space $\mathcal{X}$, and let $A := \bigcap_{i=1}^{\infty} A_i$. Consider an arbitrary sequence $a_i \in \mathcal{X}$, $i \in \mathbb{N}$, and assume that $a_i \in A_i$ for all i. If $a_i \to a \in \mathcal{X}$, then $a \in A$.*

Proof. Note that $a_j \in A_j \subseteq A_i$ for each $j \geq i$, and since A_i is closed, by taking the limit as $j \to \infty$, we conclude that $a \in A_i$ for each i. This implies that $a \in A = \bigcap_{i=1}^{\infty} A_i$. □

We apply the above lemma to the closed sets $\overline{\Pi(x;\epsilon_i)}$ and the sequence $\pi(\xi_i)$ as $i \to \infty$. Suppose $\pi(\xi_i) \to \hat{x} \in \mathcal{X}$. By the lemma, $\hat{x} \in \Pi(x) = \bigcap_{\epsilon>0} \overline{\Pi(x;\epsilon)}$. According to our assumption, $f(\hat{x}) = f(x)$. Since $f \in \mathcal{F}$ is continuous on $\mathcal{X}$ and $\pi(\xi_i) \to \hat{x}$, it follows that $f(\pi(\xi_i)) \to f(\hat{x}) = f(x)$, thus completing the proof. Note that if $x \notin \Pi^{\circ}$, we obtain a stronger result where $\hat{x} = x$, still allowing us to conclude $f(\hat{x}) = f(x)$. The condition $f(\hat{x}) = f(x)$ is non-trivial only for critical points.

What if the sequence $\pi(\xi_i)$ does not converge to any limit in $\mathcal{X}$? We define

$$\delta_i := \min_{\hat{x} \in \Pi(x)} \operatorname{dist}\left(\pi(\xi_i), \hat{x}\right), \tag{17}$$

for any i. First, observe that $\delta_i \to 0$. If not, there would exist a subsequence ξ_{i_k}, $k \in \mathbb{N}$, such that $\delta_{i_k} \geq \delta > 0$. However, since $\xi_{i_k} \in \mathcal{B}(x;\epsilon)$ for some sufficiently large ϵ, and $\mathcal{B}(x;\epsilon)$ is a compact subset of $\mathcal{X}$, there must be a further subsequence $\xi_{i_{k_\ell}}$, $\ell \in \mathbb{N}$, that converges. By the same argument as before, $\xi_{i_{k_\ell}} \to \hat{x}$ as $\ell \to \infty$ for some $\hat{x} \in \Pi(x)$, which contradicts the assumption that δ is positive.

Thus, as $\delta_i \to 0$, we need to show that $f(\pi(\xi_i)) \to f(x)$. Note that $\Pi(x)$ is a closed and bounded (i.e., compact) set, and by assumption, f is continuous at every $\hat{x} \in \Pi(x)$. Therefore, for any $\hat{x} \in \Pi(x)$ and any $\gamma > 0$, there exists $\rho(\hat{x}) > 0$ such that if $\operatorname{dist}(\pi(\xi_i), \hat{x}) \leq \rho(\hat{x})$, then $|f(\pi(\xi_i)) - f(\hat{x})| \leq \gamma$. Since $f(\hat{x}) = f(x)$ and $\Pi(x)$ is compact, we can conclude that for any $\gamma > 0$, there exists $\rho > 0$ such that if $\delta_i = \min_{\hat{x} \in \Pi(x)} \operatorname{dist}(\pi(\xi_i), \hat{x}) \leq \rho$, then $|f(\pi(\xi_i)) - f(x)| \leq \gamma$. Hence, as $\delta_i \to 0$, there exists a sufficiently large N such that $|f(\pi(\xi_i)) - f(x)| \leq \gamma$ for all $i \geq N$. This demonstrates that $f(\pi(\xi_i)) \to f(x)$ and completes the proof.

Part 2 ($\Longleftarrow$). Assume now that $f(\pi(x))$ is a continuous function on $\mathcal{X}$. Similar to the previous argument, we can assume without loss of generality that $\pi(x) = x$. For any $\hat{x} \in \Pi(x)$, by definition, there exists a sequence ξ_i, $i \in \mathbb{N}$, such that $\xi_i \to x$ and $\pi(\xi_i) \to \hat{x}$. Given the continuity of the canonicalized function $f(\pi(x))$, we have $f(\pi(\xi_i)) \to f(\pi(x)) = f(x)$. Since $\pi(\xi_i) \to \hat{x}$ and f is continuous, it follows that $f(\pi(\xi_i)) \to f(\hat{x})$. Therefore, $f(\hat{x}) = f(x)$, which completes the proof.

□

C Proof of Theorem 3

Proof. We divide the proof into two parts.

Part 1 ($\Longrightarrow$). Fix an arbitrary $x \in \mathcal{X}$ and a function $f \in \mathcal{F}$. We need to show that $\nabla^k f_{\text{can}}(x)$ exists continuously on $\mathcal{X}$. Similar to the proof of Theorem 1, this trivially holds if $\pi(x) \in (\mathcal{X}/G)^\circ$. So without loss of generality, we assume that $\pi(x) = x$ and $x \in \partial(\mathcal{X}/G)$.

We prove the theorem by induction on $k \in \mathbb{N} \cup \{0\}$ and claim that $\nabla^\ell f_{\text{can}}(x) = \nabla^\ell f(x)$ for any $\ell \in [k]$ and any $x \in \mathcal{X}/G$. The case $k = 0$ of the theorem corresponds to the continuity of the canonicalized model, and it is addressed in Theorem 1. Thus, assume that $k \in \mathbb{N}$ and the function $\nabla^{k-1} f_{\text{can}}(x)$ is defined continuously on $\mathcal{X}$. We claim that $\nabla^k f_{\text{can}}(x)$ exists for any $x \in \partial(\mathcal{X}/G)$, and also $\nabla^k f_{\text{can}}(x) = \nabla^k f(x)$. To prove this, according to the definition, we need to show that

$$\lim_{\xi \to x} \frac{\left\| \nabla^{k-1} f_{\text{can}}(\xi) - \nabla^{k-1} f_{\text{can}}(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' \right\|_2}{\left\| \xi - x \right\|_2} = 0, \tag{18}$$

where

$$\left\langle \nabla^k f(x), \xi - x \right\rangle' := \sum_{i=1}^{d} \nabla^k f(x)_{i,:} (\xi - x)_i. \tag{19}$$

In particular, note that $\nabla^k f(x) \in \mathbb{R}^{d^k}$, $(\xi - x) \in \mathbb{R}^d$, and $\left\langle \nabla^k f(x), \xi - x \right\rangle' \in \mathbb{R}^{d^{k-1}}$. We need to consider all the scenarios that can happen for the above limit and show that it will always converge to zero. As the first possible case, assume that as $\xi \to x$, we observe that $\xi \in \partial(\mathcal{X}/G)$ infinitely often. Passing to this specific subsequence, we have that

$$\left\| \nabla^{k-1} f_{\text{can}}(\xi) - \nabla^{k-1} f_{\text{can}}(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' \right\|_2 \tag{20}$$

$$= \left\| \nabla^{k-1} f(\xi) - \nabla^{k-1} f(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' \right\|_2 \ll \left\| \xi - x \right\|_2, \tag{21}$$

as $\xi \to x$, where the latter holds from the differentiability of $\nabla^k f$ at x.

Thus, we can assume that as $\xi \to x$, we never have $\xi \in \partial(\mathcal{X}/G)$. Consider $\pi(\xi) = D(g_\xi)\xi$. According to the assumption, the group G is finite, so $D(g_\xi)$ can take only finitely many values. By passing to a subsequence, we can assume that $D(g_\xi) = D(g)$ for some fixed $g \in G$, along the way that $\xi \to x$. Note that since always $\xi \notin \partial(\mathcal{X}/G)$, we have $\pi(\xi) = D(g)\xi$ locally around each ξ. This is because the mapping $\xi \to D(g)\xi$ is continuous. Also, note that according to the definition, we have $D(g)\xi \to D(g)x = \hat{x}$ for some $\hat{x} \in \Pi(x)$.

Therefore, we conclude that $f_{\text{can}} = f \circ D(g)$ locally around each ξ. Note that we can compute derivatives of $\tilde{f}(x) := f(D(g)x)$ using the chain rule:

Lemma 2. *If $\tilde{f}(x) := f(D(g)x)$ for some fixed $g \in G$, then for any $x \in \mathcal{X}$ and any $\ell \in [k]$,*

$$\nabla^\ell \tilde{f}(x) = D(g)^{\otimes \ell} \nabla^\ell f(D(g)x). \tag{22}$$

The proof of Lemma 2 is presented in Appendix D. Now using the above lemma and the fact that $f_{\text{can}} = \tilde{f}$ locally around each ξ, we have

$$\begin{aligned}
\nabla^{k-1} f_{\text{can}}(\xi) - \nabla^{k-1} f_{\text{can}}(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' & \tag{23}\\
\overset{(a)}{=} D(g)^{\otimes(k-1)} \nabla^{k-1} f(D(g)\xi) - \nabla^{k-1} f(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' & \tag{24}\\
\overset{(b)}{=} D(g)^{\otimes(k-1)} \nabla^{k-1} f(D(g)\xi) - D(g)^{\otimes(k-1)} \nabla^{k-1} f(\hat{x}) - \left\langle \nabla^k f(x), \xi - x \right\rangle', & \tag{25}
\end{aligned}$$

where in above, (a) follows from the assumption that $\nabla^{k-1} f_{\text{can}}(x) = \nabla^{k-1} f(x)$ (which holds from the induction hypothesis). Moreover, (b) holds from the assumption in the theorem.

Now we use the following claim:

Claim 1. $\left\langle \nabla^k f(x), \xi - x \right\rangle' = D(g)^{\otimes(k-1)} \left\langle \nabla^k f(\hat{x}), D(g)(\xi - x) \right\rangle'$.

The proof of Claim 1 is presented in Appendix E.

Thus, using the above claim and noting that $D(g)x = \hat{x}$ and $D(g)\xi \to \hat{x}$, we conclude

$$\begin{aligned}
& \left\| \nabla^{k-1} f_{\text{can}}(\xi) - \nabla^{k-1} f_{\text{can}}(x) - \left\langle \nabla^k f(x), \xi - x \right\rangle' \right\|_2 & \tag{26}\\
& = \left\| D(g)^{\otimes(k-1)} \left(\nabla^{k-1} f(D(g)\xi) - \nabla^{k-1} f(\hat{x}) - \left\langle \nabla^k f(\hat{x}), D(g)\xi - \hat{x} \right\rangle' \right) \right\|_2 & \tag{27}\\
& \ll \left\| D(g)^{\otimes(k-1)} \right\|_{\text{op}} \left\| D(g)\xi - \hat{x} \right\|_2 \leq \left\| D(g)^{\otimes(k-1)} \right\|_{\text{op}} \left\| D(g) \right\|_{\text{op}} \left\| \xi - x \right\|_2 & \tag{28}\\
& \leq \left\| D(g) \right\|_{\text{op}}^k \left\| \xi - x \right\|_2, & \tag{29}
\end{aligned}$$

which completes the proof of the claim that $\nabla^k f_{\text{can}}(x)$ exists for any $x \in \mathcal{X}/G$, and also that $\nabla^k f_{\text{can}}(x) = \nabla^k f(x)$ for all $x \in \mathcal{X}/G$. Note that the last inequality holds from the sub-multiplicity of the operator norm under tensor products.

Now that we have proved that f_{can} is differentiable up to order k, we need to show that its k-th order derivative is also continuous, which completes the proof.

Again, we only need to consider the case that $x \in \partial(\mathcal{X}/G)$. Let $\xi \to x$ be an arbitrary sequence. Our goal is to prove that

$$\lim_{\xi \to x} \nabla^k f_{\text{can}}(\xi) = \nabla^k f_{\text{can}}(x) = \nabla^k f(x). \tag{30}$$

If we observe that $\xi \in \mathcal{X}/G$ infinitely often, then along that specific subsequence, the above condition holds. This is since $\nabla^k f$ exists and is continuous on $\mathcal{X}$ according to the assumption.

Therefore, assume that $\xi \notin \mathcal{X}/G$. Note that $\pi(\xi) = D(g_\xi)\xi$ and since the group G is finite, by passing to a subsequence, we can assume that $D(g_\xi) = D(g)$ for some fixed $g \in G$. Similar to the proof of the existence of derivatives, we can assume that $\pi(\xi) = D(g)\xi$ locally around each ξ, and also that $D(g)\xi \to D(g)x = \hat{x}$ for some $\hat{x} \in \Pi(x)$.

Thus, $f_{\text{can}}(\xi) = f(D(g)\xi)$ locally around each ξ. By Lemma 2, we conclude

$$\lim_{\xi \to x} \nabla^k f_{\text{can}}(\xi) = \lim_{\xi \to x} D(g)^{\otimes k} \nabla^k f(D(g)\xi) = \lim_{\xi \to x} D(g)^{\otimes k} \nabla^k f(\hat{x}) = \nabla^k f(x), \tag{31}$$

where in the last step, we used the assumption in the theorem. The proof is thus complete.

Part 2 ($\Longleftarrow$). Assume that $f_{\text{can}} = f \circ \pi$ is continuously differentiable up to order k at any point $x \in \mathcal{X}$. Note that we have $f_{\text{can}}(x) = f_{\text{can}}(gx) = f_{\text{can}}(D(g)x)$ for any $g \in G$.

Now, let $x \in \partial(\mathcal{X}/G)$ and $(g, \hat{x}) \in \Xi(x)$ be arbitrary. According to the definition, and similar to the proof of Part 3 in there exists a sequence $\xi_i \in \mathcal{X}$, $i \in \mathbb{N}$, such that $\xi_i \to x$ and $\pi(\xi_i) \to \hat{x}$. Note that $D(g)\xi_i \to D(g)x = \hat{x}$ as well. Without loss of generality, we can assume that $\pi(\xi) = D(g)\xi$ for any $\xi \in \mathcal{X}$ being sufficiently close to ξ_i, for any $i \in \mathbb{N}$. This means that we have $f_{\text{can}}(\xi) = f(D(g)\xi)$ around ξ_i, for any $i \in \mathbb{N}$. Now, let $x \in \partial(\mathcal{X}/G)$ and $(g, \hat{x}) \in \Xi(x)$ be arbitrary. According to the definition, and similar to the proof of Part 3 in Proposition 1, there exists a sequence $\xi_i \in \mathcal{X}$, $i \in \mathbb{N}$, such that $\xi_i \to x$ and $\pi(\xi_i) \to \hat{x}$. Note that $D(g)\xi_i \to D(g)x = \hat{x}$ as well. Without loss of generality, we can assume that $\pi(\xi) = D(g)\xi$ for any $\xi \in \mathcal{X}$ being sufficiently close to ξ_i, for any $i \in \mathbb{N}$. This means that we have $f_{\text{can}}(\xi) = f(D(g)\xi)$ around ξ_i, for any $i \in \mathbb{N}$.

Therefore, according to Lemma 2, we have

$$\nabla^\ell f_{\text{can}}(\xi) = D(g)^{\otimes \ell} \nabla^\ell f(D(g)\xi), \tag{32}$$

for any ξ sufficiently close to ξ_i, and any $i \in \mathbb{N}$. Specifically, if we choose $\xi = \xi_i$, and take the limit as $\xi_i \to x$, we conclude that

$$\nabla^\ell f_{\text{can}}(x) = \lim_{\xi_i \to x} \nabla^\ell f_{\text{can}}(\xi_i) = D(g)^{\otimes \ell} \lim_{\xi_i \to x} \nabla^\ell f(D(g)\xi_i) = D(g)^{\otimes \ell} \nabla^\ell f(D(g)x) \tag{33}$$

$$= D(g)^{\otimes \ell} \nabla^\ell f(\hat{x}), \tag{34}$$

and this completes the proof.

□

D Proof of Lemma 2

Proof. Fix $g \in G$ and define $y := D(g)x \in \mathbb{R}^d$. Let us evaluate derivatives of $f(D(g)x)$. Note that

$$y_i = \sum_{j=1}^{d} D(g)_{ij} x_j \implies \frac{\partial y_i}{\partial x_j} = D(g)_{ij}, \tag{35}$$

for any $i, j \in [d]$. This means that

$$\frac{\partial}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial y_i}{\partial x_j} \frac{\partial}{\partial y_i} = \sum_{i=1}^{d} D(g)_{ij} \frac{\partial}{\partial y_i}, \tag{36}$$

for any $j \in [d]$. Similarly, we can extend this to arbitrary higher-order partial derivatives.

Lemma 3. *For any $\ell \in [k]$, and any $j_1, j_2, \ldots, j_\ell \in [d]$, one has*

$$\frac{\partial^\ell}{\partial x_{j_1} \partial x_{j_2} \ldots \partial x_{j_\ell}} = \sum_{i_1, i_2, \ldots, i_\ell = 1}^{d} \prod_{t=1}^{\ell} D(g)_{i_t j_t} \frac{\partial^\ell}{\partial y_{i_1} \partial y_{i_2} \ldots \partial y_{i_\ell}}. \tag{37}$$

The proof of Lemma 3 is presented in Appendix F. Now consider an arbitrary $\ell \in [k]$, and apply the above differentiation formula to $\tilde{f}(x) = f(D(g)x)$ to get

$$\frac{\partial^\ell \tilde{f}}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_\ell}} = \sum_{i_1, i_2, \ldots, i_\ell = 1}^{d} \prod_{t=1}^{\ell} D(g)_{i_t j_t} \frac{\partial^\ell \tilde{f}}{\partial y_{i_1} \partial y_{i_2} \ldots \partial y_{i_\ell}}, \tag{38}$$

for any $j_1, j_2, \ldots, j_\ell \in [d]$. Now note that $D(g)^{\otimes \ell} \in \mathbb{R}^{d^\ell \times d^\ell}$ is a matrix with entries:

$$D(g)^{\otimes \ell}_{i,j} = \prod_{t=1}^{\ell} D(g)_{i_t j_t}, \tag{39}$$

where $i = (i_1, i_2, \ldots, i_\ell)$ and $j = (j_1, j_2, \ldots, j_\ell)$. This means that, for any $x \in \mathcal{X}$ and any $g \in G$, if we use the vector representation of the ℓ-th order derivatives, we get

$$\nabla^\ell \tilde{f}(x) = D(g)^{\otimes \ell} \nabla^\ell f(y) = D(g)^{\otimes \ell} \nabla^\ell f(D(g)x). \tag{40}$$

The proof is thus complete. □

E Proof of Claim 1

Note that both the left-hand side and the right-hand side of the desired identity are vectors in $\mathbb{R}^{d^{k-1}}$. We start from the right-hand side of the claim. For any arbitrary $i_2, i_3, \dots, i_k \in [d]$,

$$\Big(D(g)^{\otimes(k-1)}\big\langle\nabla^k f(\hat{x}),D(g)(\xi-x)\big\rangle'\Big)_{i_2,i_3,\dots,i_k} \tag{41}$$

$$= \sum_{j_2,j_3,\dots,j_k=1}^{d} \prod_{t=2}^{k} D(g)_{i_t j_t} \sum_{i=1}^{d} \nabla^k f(\hat{x})_{i,j_2,j_3,\dots,j_k} \sum_{j=1}^{d} D(g)_{ij}(\xi-x)_j \tag{42}$$

$$= \sum_{j=1}^{d}(\xi-x)_j \sum_{i=1}^{d} D(g)_{ij} \sum_{j_2,j_3,\dots,j_k=1}^{d} \prod_{t=2}^{k} D(g)_{i_t j_t} \nabla^k f(\hat{x})_{i,j_2,j_3,\dots,j_k}. \tag{43}$$

Note that according to the assumption, we have

$$\nabla^k f(x)_{j,i_2,i_3,\dots,i_k} = \sum_{i=1}^{d} D(g)_{ji} \sum_{j_2,j_3,\dots,j_k=1}^{d} \prod_{t=2}^{k} D(g)_{i_t j_t} \nabla^k f(\hat{x})_{i,j_2,j_3,\dots,j_k} \tag{44}$$

Thus, since $D(g)^{\mathsf{T}} = D(g)$, we conclude that

$$\Big(D(g)^{\otimes(k-1)}\big\langle\nabla^k f(\hat{x}),D(g)(\xi-x)\big\rangle'\Big)_{i_2,i_3,\dots,i_k} \tag{45}$$

$$= \sum_{j=1}^{d} \nabla^k f(x)_{j,i_2,i_3,\dots,i_k}(\xi-x)_j = \Big(\big\langle\nabla^k f(x),\xi-x\big\rangle'\Big)_{i_2,i_3,\dots,i_k}, \tag{46}$$

and this completes the proof.

F Proof of Lemma 3

Proof. The case $\ell = 1$ is already proved in Equation (36). We use induction on ℓ to prove the general case. Suppose $\ell \geq 2$ and the identity holds up to $\ell - 1$. We have

$$\frac{\partial^{\ell-1}}{\partial x_{j_1}\partial x_{j_2}\dots\partial x_{j_{\ell-1}}} = \sum_{i_1,i_2,\dots,i_{\ell-1}=1}^{d} \prod_{t=1}^{\ell-1} D(g)_{i_t j_t} \frac{\partial^{\ell-1}}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_{\ell-1}}}, \tag{47}$$

for any $j_1, j_2, \dots, j_{\ell-1} \in [d]$. Now for any $j_\ell \in [d]$, we take partial derivatives from the above identity, with respect to x_{j_ℓ}, to get

$$\frac{\partial^{\ell}}{\partial x_{j_1}\partial x_{j_2}\dots\partial x_{j_\ell}} = \frac{\partial}{\partial x_{j_\ell}}\Big(\frac{\partial^{\ell-1}}{\partial x_{j_1}\partial x_{j_2}\dots\partial x_{j_{\ell-1}}}\Big) \tag{48}$$

$$= \frac{\partial}{\partial x_{j_\ell}}\Big(\sum_{i_1,i_2,\dots,i_{\ell-1}=1}^{d} \prod_{t=1}^{\ell-1} D(g)_{i_t j_t} \frac{\partial^{\ell-1}}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_{\ell-1}}}\Big) \tag{49}$$

$$= \sum_{i_1,i_2,\dots,i_{\ell-1}=1}^{d} \prod_{t=1}^{\ell-1} D(g)_{i_t j_t} \frac{\partial}{\partial x_{j_\ell}}\Big(\frac{\partial^{\ell-1}}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_{\ell-1}}}\Big). \tag{50}$$

Note that from the identity for $\ell = 1$, we have

$$\frac{\partial}{\partial x_{j_\ell}}\Big(\frac{\partial^{\ell-1}}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_{\ell-1}}}\Big) = \sum_{i_\ell=1}^{d} D(g)_{i_\ell j_\ell} \frac{\partial^{\ell}}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_\ell}}. \tag{51}$$

Therefore,

$$\frac{\partial^\ell}{\partial x_{j_1}\partial x_{j_2}\dots\partial x_{j_\ell}} = \sum_{i_1,i_2,\dots,i_{\ell-1}=1}^{d}\prod_{t=1}^{\ell-1} D(g)_{i_t j_t}\sum_{i_\ell=1}^{d} D(g)_{i_\ell j_\ell}\frac{\partial^\ell}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_\ell}} \tag{52}$$

$$= \sum_{i_1,i_2,\dots,i_{\ell}=1}^{d}\prod_{t=1}^{\ell} D(g)_{i_t j_t}\frac{\partial^\ell}{\partial y_{i_1}\partial y_{i_2}\dots\partial y_{i_\ell}}, \tag{53}$$

and this completes the proof. □