Symmetry-adapted representation learning

Fabio Anselmi [a,b,*], Georgios Evangelopoulos [a,1,*], Lorenzo Rosasco [a,b], Tomaso Poggio [a]

[a] The Center for Brains, Minds, and Machines, MIT and McGovern Institute for Brain Research at MIT, Cambridge, MA, USA
[b] Laboratory for Computational and Statistical Learning (LCSL), Istituto Italiano di Tecnologia, Genova, Italy

Keywords:
Representation learning
Equivariant representations
Invariant representations
Dictionary learning
Convolutional neural networks
Regularization
Data transformations

A B S T R A C T

In this paper, we propose the use of data symmetries, in the sense of equivalences under signal transformations, as priors for learning symmetry-adapted data representations, i.e., representations that are equivariant to these transformations. We rely on a group-theoretic definition of equivariance and provide conditions for enforcing a learned representation, for example the weights in a neural network layer or the atoms in a dictionary, to have the structure of a group and specifically the group structure in the distribution of the input. By reducing the analysis of generic group symmetries to permutation symmetries, we devise a regularization scheme for representation learning algorithm, using an unlabeled training set. The proposed regularization is aimed to be a conceptual, theoretical and computational proof of concept for symmetry-adapted representation learning, where the learned data representations are equivariant or invariant to transformations, without explicit knowledge of the underlying symmetries in the data.

1. Introduction

Symmetry is ubiquitous from subatomic particles to natural patterns, man-made design, art and mathematics. Invariance to symmetries in pattern recognition and computational neuroscience is an old challenging problem [1–12]. More recently, in the context of machine learning, data symmetries have been used to derive data representations with the properties of equivariance and invariance [13–17] to unknown, symmetry-generating transformations, for example geometric transformations. These properties are reflected as structure in the representation atoms and can be explicitly used for reducing the complexity of downstream supervised learning. This is achieved, for example, by constructing representations that are invariant to transformations irrelevant for the learning task, that preserve the data distribution and prediction function [18–20]. For example, for image classification, object position or scale are data symmetries that are irrelevant.

Representations that reflect symmetries inherent in the data distribution define a quotient representation space where points are equivalent up to transformations [21]. In this space, the sample complexity of learning (the size of the labeled training set) [16,22,23] can be reduced by pooling on the representation coefficients. Indeed, the pooling operation has a crucial role in Convolutional Neural Networks (CNNs) for enforcing stability to small, local perturbations [24,25]. On the other hand, learning symmetry-adapted representations is in the direction of (a) generalizing CNNs to arbitrary weight sharing schemes and invariances by learning the symmetry group from the data and (b) learning, as opposed to informed designing, network architectures and feature map properties, such as locality, connectivity patterns and weight-sharing topologies.

CNNs and Convolutional Sparse Coding schemes [26] have an explicit parameterization for equivariance and robustness to shifts in the input (translations) through convolutions and pooling respectively (see also [27]). However, data symmetries extend to more general transformations depending on the data domain, for example geometric changes such as scaling, rotation or affine maps for the case of images, which, in general are unknown or complex to model. A symmetry-blind data representation will have to compensate for transformation variability using more parameters, and as a result an increased demand for labeled examples or data augmentation and adaptation assuming simple and known transformation models [28,29]. Extensions of CNNs to known transformations, beyond translations, were explored with scale-space pooling [30], convolutional maxout networks [31], pooling over neighboring values and similar filters [32], tiled CNNs [33], cyclic weight sharing and pooling [34], and wavelet *scattering networks* [35,36], (but see also [37]). In particular *symmetry networks* [23] and *group-equivariant networks* [17,38] highlighted the complexity gains of incorporating other symmetries in the representations at each layer

* Corresponding author at: Department of Brain and Cognitive Science, 43 Vassar Street, Cambridge, MA 02139, United States.
E-mail addresses: anselmi@mit.edu (F. Anselmi), gevang@mit.edu (G. Evangelopoulos), lrosasco@mit.edu (L. Rosasco), tp@mit.edu (T. Poggio).
[1] Current affiliation: X (Alphabet Inc.).

of CNNs. In [39] the parameters of a neural network are tied to achieve equivariance with respect to a known group. Weight sharing over transformations capturing perceptual changes, such as speaker characteristics, were used for speech representations through *group-CNNs* [40].

The more general problem of *learning symmetries* has been previously approached as estimating the infinitesimal generators of Lie groups generating data transformations [41–44]. *Symmetries in learning* have been used in the context of categorizing symmetry groups (mirror, roto-translation) in random patterns [45]. The relations between typical, e.g., l_1, l_2, l_∞, and group-based regularization schemes, for known groups, have been explored in [46].

The contributions of this work are: **(1)** Outlining principles for learning symmetries in data, and learning equivariant representations, without explicit knowledge of the symmetry group. As opposed to *learning with known symmetries*, like many existing methods, we propose *learning the symmetries*; **(2)** Reducing the analysis of generic group symmetries to permutation symmetries; **(3)** Deriving an analytic expression for a regularization term acting on the representation matrix that promotes a group structure and specifically the group structure in an unlabeled observation set.

The rest of the paper is organized as follows: In Section 2, we briefly recall the setting of representations with equivariance or invariance to transformations captured by group symmetries [14,16,17,21,47]. In Section 3 we formulate the problem of symmetry-adapted representation learning and provide a general principle (Section 4) for designing the regularization term. The main theoretical contributions are stated in Section 5 along with a computable, analytic form for a regularization term. Section 6 provides proof of concept results on learning exact, analytic, group-transformations.

2. Equivariant and invariant representations

We briefly recall the setting of constructing data representations with equivariance or invariance to transformations using group symmetries [14,16,17,21,47]. Intuitively, a representation is equivariant with respect to a transformation on the input space, if it can be equivalently expressed as some transformation on the representation space. If this is the identity, the representation is invariant to the transformation.

Let the input space be a vector space endowed with dot product $\langle \cdot, \cdot \rangle$, $\mathcal{X} = \mathbb{R}^d$. We denote the transformations of a point $x \in \mathcal{X}$ through a representation of a group $\mathcal{G}$, of order $|\mathcal{G}| < \infty$ in d dimensions. In the following, we identify $\mathcal{G}$ with the set of $d \times d$ matrices $\{g\}$, (e.g. rotation or reflection matrices), and each transformation by gx. The *group orbit* of $x \in \mathcal{X}$ is the set of transformed signals

$$O_x = \{gx \in \mathcal{X} \,|\, g \in \mathcal{G}\},\ x \in \mathcal{X}. \tag{1}$$

Thus the action of the group, and the associated orbit definition, induces a partition of $\mathcal{X}$ into orbit sets. More precisely, it defines an equivalence relation

$$x \sim x' \Leftrightarrow \exists g \in \mathcal{G}\colon x' = gx,\ \forall x, x' \in \mathcal{X}, \tag{2}$$

that separates points in $\mathcal{X}$ on the basis of x being a transformation of x'.

Orbits can be used to derive a representation $\Phi : \mathcal{X} \to \mathcal{F}$ (e.g. $\mathbb{R}^k$) selected from some hypothesis space of maps with a specific parametrization. A parametrization for an equivariant, or invariant as a special case, representation is a nonlinear measurements of the projections on the orbit elements [16,21]. More specifically, given an orbit of group $\mathcal{G}$ of some vector $t_j \in \mathcal{X}$ written in a matrix form as

$$W_j = \left[g_1 t_j, \ldots, g_{|\mathcal{G}|} t_j\right],\ t_j \in \mathcal{X}, \tag{3}$$

the nonlinear projection of x on W_j is

$$\Phi_{j,\alpha}(x) = \sigma_\alpha(W_j^T x) = \left(\sigma_\alpha(\langle g_1 t_j, x\rangle), \ldots, \sigma_\alpha(\langle g_{|\mathcal{G}|} t_j, x\rangle)\right)^T, \tag{4}$$

where $\{\sigma_\alpha : \mathbb{R} \to \mathbb{R} \ |\alpha \in \mathbb{R}\}$ is a set of nonlinear functions with a scalar parameter α, e.g. sigmoids $\sigma_\alpha(\cdot) = (1 + e^{-\alpha})^{-1}$ or rectifier functions $\sigma_\alpha(\cdot) = \max(0, \cdot - \alpha)$.

It is easy to see, using the closure property of the group composition, that $\Phi_{j,\alpha} : \mathcal{X} \to \mathbb{R}^{|\mathcal{G}|}$ is a *permutation-equivariant* representation w.r.t. the transformations in $\mathcal{G}$ i.e.:

$$\Phi_{j,\alpha}(gx) = P_g \Phi_{j,\alpha}(x),\ g \in \mathcal{G}, \forall j, \alpha \tag{5}$$

where $P_g \in \mathbb{R}^{|\mathcal{G}| \times |\mathcal{G}|}$ is a permutation matrix that depends on the transformation g. Indeed, transforming x amounts to simply reordering the entries in vector (4). This can be easily seen noting that each entry i in the vector in (4) is

$$(\Phi_{j,\alpha}(gx))_i = \sigma_\alpha(\langle g_i t, gx\rangle) = \sigma_\alpha\left(\langle g^{-1} g_i t, x\rangle\right),$$

and due to the closure property of the group composition, i.e. $g_i g_k = g_l,\ \forall\, i, k\, \exists\, l$, we have that setting $x \to gx$ in $\Phi_{j,\alpha}(x)$ amounts to a permutation of the vector entries.

Further, by summing the components of $\Phi_{j,\alpha}(x)$ (or any pointwise function of them), we obtain a representation $\bar{\Phi}_{j,\alpha} : \mathcal{X} \to \mathbb{R}$ *invariant* to transformations in $\mathcal{G}$:

$$\bar{\Phi}_{j,\alpha}(x) = \sum_{i=1}^{|\mathcal{G}|} \sigma_\alpha\left(\langle g_i t_j, x\rangle\right) = \mathbf{1}_{|\mathcal{G}|}^T \Phi_{j,\alpha}(x), \tag{6}$$

where $\alpha \in \mathbb{R}$ and $\mathbf{1}_{|\mathcal{G}|}$ is the all-ones vector of dimension $|\mathcal{G}|$. Indeed, using the permutation equivariance in (5) we have for all $g \in \mathcal{G}$:

$$\begin{aligned}\bar{\Phi}_{j,\alpha}(gx) &= \mathbf{1}_{|\mathcal{G}|}^T \Phi_{j,\alpha}(gx) = \mathbf{1}_{|\mathcal{G}|}^T P_g \Phi_{j,\alpha}(x) \\ &= \mathbf{1}_{|\mathcal{G}|}^T \Phi_{j,\alpha}(x) = \bar{\Phi}_{j,\alpha}(x)\end{aligned} \tag{7}$$

since $\mathbf{1}_{|\mathcal{G}|}^T P_g = \mathbf{1}_{|\mathcal{G}|}^T$. Another example of invariant quantity, being P_g unitary, is the ℓ^2 norm of the representation vector:

$$\left\| \Phi_{j,\alpha}(gx) \right\|_2^2 = \left\| P_g \Phi_{j,\alpha}(x) \right\|_2^2 = \left\| \Phi_{j,\alpha}(x) \right\|_2^2 \tag{8}$$

Furthermore, given a set of Q such orbits $\left\{W_j\right\}_{j=1}^Q$ of the same group $\mathcal{G}$, the map $\Phi : \mathcal{X} \to \mathbb{R}^{Q \times |\{\sigma_\alpha\}|}$ obtained by concatenating $\{\bar{\Phi}_{j,\alpha}\}$ can be shown to be *selective* to the partition of $\mathcal{X}$ by $\mathcal{G}$, or equivalently, sufficient for separating the equivalence classes induced by the group action [16].

3. Problem formulation

In the above sense, deriving equivariant and invariant representations is conditioned upon having access to an orbit set W, or learning a set W such that it reflects the same generating symmetries of the target group. In this paper, we focus on learning such a symmetry set, without supervision, from an observation set for which we make the following simplifying assumption:

Assumption 1. The observation set $S_N = \{x_i\}_{i=1}^N \subset \mathcal{X}$ is a finite collection of Q orbits in $\mathcal{X} = \mathbb{R}^d$ w.r.t. a finite group $\mathcal{G}$

$$S_N = \{x_i\}_{i=1}^N = \{gx_j,\ \forall g \in \mathcal{G}\}_{j=1}^Q,\ x_i, x_j \in \mathcal{X} \tag{9}$$

where $N = |\mathcal{G}|Q$ and $|\mathcal{G}|$ is the group cardinality.

A symmetry-adapted representation learning problem will be formulated as learning W from data S_N such that

$$\underset{W \in \mathbb{R}^{d \times |\mathcal{G}|}}{\arg\min} \left(\mathcal{L}(W, S_N) + \gamma \mathcal{L}'(W, S_N) + \beta \mathcal{R}(W)\right) \tag{10}$$

with $\beta, \gamma \in \mathbb{R}_+$ where $\mathcal{L}(W, S_N)$ is the representation loss selected to satisfy some objective criterion, e.g. reconstruction, similarity or

empirical error (supervised) [24]; $\mathcal{R} : \mathbb{R}^{d\times|\mathcal{G}|} \to \mathbb{R}_+$ is a regularization term, controlled by β, that enforces the columns of W to be an orbit of a finite group; $\mathcal{L}'(W, S_N)$ is a data-dependent regularization term to further restrict the group in W to be the same as the latent group in the observation set.

The contribution of this paper is to develop analytic formulations for $\mathcal{L}'$ and $\mathcal{R}$ assuming S_N as above, given $|\mathcal{G}|$ and no form of supervision on the partition of S_N or the group identity.

4. The Gram matrix of orbits

The simple observation for designing $\mathcal{R}(W)$ comes from the inspection of the matrix of all inner products of the vectors of an orbit (the so called Gram matrix). If the columns of W correspond to an orbit of a vector $t \in \mathbb{R}^d$:

$$W = \left[g_1 t, \dots, g_{|\mathcal{G}|} t\right], \tag{11}$$

then the associated Gram matrix $G = W^T W \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{G}|}$ has entries of the form:

$$(G)_{ij} = \left\langle g_i t, g_j t\right\rangle = \left\langle t, g_i^* g_j t\right\rangle = v_t(g_i^{-1} g_j) \tag{12}$$

where g_i^* the conjugate transpose of g_i and $v_t : \mathcal{G}\times\mathcal{G} \to \mathbb{R}$ is an injective function that depends on the vector t. Assuming a unitary group, i.e. $g^* = g^{-1}$, a hypothesis that can be relaxed, and using the closure property of group composition, we have the following proposition (see also [48], Ch5, Th.2 and [49]).

Proposition 1. *If a set of vectors forms an orbit w.r.t. a finite, unitary group, then their Gramian matrix G is a permuted matrix i.e. a matrix whose columns are permutations of a single vector.*

Proof. Each entry of G is a pointwise function, v_t, of the multiplication table of $\mathcal{G}$. By the Cayley theorem [50], the columns of the multiplication table of a group are permuted versions one of the other. In other words the statement that $G = W^T W$ is a permuted matrix follows from the property that the group is closed under composition. □

As an intuition for the proposition, consider the simple example of the cyclic group of order 3. Let $\mathbb{Z}_3 = \{0, 1, 2\}$ with law $i \circ j = (i + j \pmod 3)$. Each column in the multiplication table associated to the elements of $\mathbb{Z}_3$ is a permutation of the others.

$\circ$	**0**	**1**	**2**
0	0	1	2
1	1	2	0
2	2	0	1

Therefore a function of the multiplication table entries will give a permuted matrix.

To further give an intuition of the idea behind the construction of the regularization term we consider the group of rotations of angles $\theta = \{0, 120, 240\}$. In dimension two the elements of the group are 2×2 matrices that we indicate with R_θ. The orbit matrix of a vector $t \in \mathbb{R}^2$ is then

$$W = (et, R_{120}t, R_{240}t)$$

with $e = R_0$ the matrix implementing a rotation of a angle zero i.e. the identity matrix. The associated gramian elements are given by

$$(G)_{ij} = \left\langle t, R_{\theta_i}^* R_{\theta_j} t\right\rangle = \left\langle t, R_{-\theta_i} R_{\theta_j} t\right\rangle = \left\langle t, R_{\theta_j - \theta_i} t\right\rangle.$$

The explicit computation of the gramian shows:

$$G = \begin{bmatrix} \langle t,t\rangle & \langle t, R_{120}t\rangle & \langle t, R_{240}t\rangle \\ \langle t, R_{-120}t\rangle & \langle t,t\rangle & \langle t, R_{120}t\rangle \\ \langle t, R_{-240}t\rangle & \langle t, R_{-120}t\rangle & \langle t,t\rangle \end{bmatrix} = \begin{bmatrix} \langle t,t\rangle & \langle t, R_{120}t\rangle & \langle t, R_{240}t\rangle \\ \langle t, R_{240}t\rangle & \langle t,t\rangle & \langle t, R_{120}t\rangle \\ \langle t, R_{120}t\rangle & \langle t, R_{240}t\rangle & \langle t,t\rangle \end{bmatrix}$$

i.e. the columns of the gramian are permuted versions of each other.

5. Analytic expressions for regularization

In this section we provide closed-form expressions for two conditions: a) the *permuted matrix condition*, from Proposition 1, enforcing a group orbit structure on the representation matrix W and b) the *same-symmetry condition* penalizing groups different than the generating group $\mathcal{G}$ of an observation set S_N.

5.1. Permuted matrix condition

Following Proposition 1, we provide an analytical condition for the Gram matrix G to be a permuted matrix, by imposing the columns (or rows) of G to be permuted versions of the same vector. A straightforward way of doing this is to impose the Euclidean distance between the distribution of column values, for all column pairs, to be null. Mathematically the condition can be written as:

$$\sum_{i,j=1}^{|\mathcal{G}|} \int d\lambda\ (h_i(\lambda) - h_j(\lambda))^2 = 0, \tag{13}$$

$$h_i(\lambda) = \sum_{k=1}^{|\mathcal{G}|} \delta(G_{ki} - \lambda)$$

with $G_{ki} = (G)_{ki}$ the element in row k and column i of G and $\delta(a) = 1,\ if\ a = 0$ and 0 otherwise. This is a necessary and sufficient condition for the columns of G to be permutations of a single vector, as the distribution of values is a *maximal invariant* [19,51] with respect to the permutation group. The following theorem provides an explicit expression for computing (13) given a matrix G.

Theorem 1 (Symmetry regularization)**.** *Let matrix $W \in \mathbb{R}^{d\times|\mathcal{G}|}$, $G = W^T W$ and $r : \mathbb{R}^{|\mathcal{G}|\times|\mathcal{G}|} \to \mathbb{R}_+$ such that*

$$r(G) = \tau^T \delta(C\mathrm{vec}(G)) \tag{14}$$

where $\mathrm{vec}(G) \in \mathbb{R}^{|\mathcal{G}|^2}$ is the vectorization of G, δ is the elementwise function $\delta(a) = 1,\ if\ a = 0$ (otherwise 0), and τ, C are respectively a constant weight vector and matrix that depend only on $|\mathcal{G}|$. If $r(G) = 0$, then G is a permuted matrix and viceversa.

Proof. A direct calculation of (13) for all pairs of columns in G gives:

$$\sum_{i,j,k,l=1}^{|\mathcal{G}|} (|\mathcal{G}|\delta(i-j) - 1)\delta(G_{ki} - G_{lj}) = 0, \tag{15}$$

which can be rewritten in a compact way by vectorizing G:

$$\tau^T \delta(C\mathrm{vec}(G)) = 0, \tag{16}$$

where τ is the $|\mathcal{G}|^2(|\mathcal{G}|^2 - 1)/2 \times 1$ vector of weights $(|\mathcal{G}|\delta(i-j) - 1), i, j = 1, \dots, |\mathcal{G}|$ and C is a $(|\mathcal{G}|^2 - 1)|\mathcal{G}|^2/2 \times |\mathcal{G}|^2$ sparse, con-

stant matrix that encodes all pairwise differences in a right multiplied vector of size $|\mathcal{G}|^2 \times 1$, namely:

$$C = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & -1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix}$$

Since the condition is derived by a chain of equalities, the converse is also true. □

The regularizer $\mathcal{R}(W)$ in (10) can then be chosen to be the function of G that minimizes (14), namely $\mathcal{R}(W) = r(W^TW)$. The size of W, or equivalently the orbit size $|\mathcal{G}|$, which would correspond to the size of the representation matrix is left as a hyperparameter or can be chosen so that the size of the associated Gram matrix allows for a fast and memory-efficient computation of $\mathcal{R}(W)$. Further, using G instead of the representation matrix W, allows us to work with the permutation group only, and thus for a uniform treatment of arbitrary, finite groups. In representations with additional structure, e.g. multilayer as in deep neural networks, the regularizer can be the sum over subsets of weights in the same layer.

We make a few additional remarks on Theorem 1: (a) Making the regularizer (14) null is a necessary and sufficient condition to have an orbit with respect to an Abelian group; see Proposition 2 in [52] and [49] where these orbits are also called *geometrically uniform frames*. (b) In addition, it can be used, as a pseudo-metric, to test the equivariance and invariance properties of a representation (Eq. (30) and Appendix B). (c) Choosing different nonlinear functions than δ [51] or distances in (13) can give alternative expressions for regularization. (d) For the case of permutation groups, the result can be generalized to a class of regularization terms. In this case, any elementwise function of the orbit matrix, i.e. $f((W)_{ij}), f : \mathbb{R} \to \mathbb{R}$, will induce a Gram matrix with entries $(G)_{ij} = \langle f(g_i t), f(g_j t) \rangle$ which is also a permuted matrix and the following holds

Corollary 1. *For any permutation group $\mathcal{G}$ and any $f : \mathbb{R} \to \mathbb{R}$ acting pointwise on the matrix W, if W^TW is a permuted matrix then $r(f(W^T)f(W)) = 0$ and viceversa.*

In fact, the action of the function on the orbit W is equivalent to a change of the vector t since for a permutation group $(f(g_i t))_q = (g_i f(t))_q$, $\forall\, i, q$. This observation will be useful for having additional constraints for representation learning (see (29)).

5.2. Same-symmetry condition

The permuted-matrix condition in Proposition 1 is not conditioned on an observation set, i.e. the group generating the orbit for W and the data could be different. One way to constrain the solutions to those with same symmetries as the observation set S_N comes from looking at the data covariance matrix. In fact if $X = [x_1, \cdots, x_N] \in \mathbb{R}^{d \times N}$ is the nonzero matrix storing S_N (as in (9)), a simple calculation shows:

$$XX^T = \sum_{i=1}^{|\mathcal{G}|} g_i T T^T g_i^T \tag{17}$$

where $T = [t_1, \dots, t_Q]$ is the $d \times Q$ matrix of representatives for each orbit in S_N (e.g. an arbitrary element of each). The matrix in (17) has a high degree of symmetry that can be used to prove the following Lemma. Let $[A, B] = AB - BA$ denote the commutator of matrices A, B.

Lemma 1. *Let $W \in \mathbb{R}^{d \times |\tilde{\mathcal{G}}|}$ a nonzero orbit matrix w.r.t. a finite group $\tilde{\mathcal{G}}$ and $X = [x_1, \cdots, x_N] \in \mathbb{R}^{d \times N}$, $x_i \in S_N$ (as in 9) with $|\tilde{\mathcal{G}}| = |\mathcal{G}|$. If $[XX^T, WW^T] = \mathbf{0}$ then $\tilde{\mathcal{G}}$ and $\mathcal{G}$ are the same up to a fixed unitary conjugation.*

Towards proving this Lemma, we start with two preparatory Lemmas: the first states that the elements of a $\mathcal{G}-$orbit of an eigenvector are also eigenvectors with the same eigenvalue; the second states that the linear span of the $\mathcal{G}-$orbit of eigenvector with eigenvalue λ coincides with the eigenspace associated to λ.

Lemma 2. *If $v \in \mathbb{R}^d$ is an eigenvector of XX^T with eigenvalue λ, the set $\{gv | g \in \mathcal{G}\}$ is a set of eigenvectors of XX^T with the same eigenvalue.*

Proof. Multiplying the right hand side of Eq. (17) by some $g_j \in \mathcal{G}$ we have

$$XX^T g_j = \sum_{i=1}^{|\mathcal{G}|} g_i T T^T g_i^T g_j = g_j \sum_{k=1}^{|\mathcal{G}|} g_k T T^T g_k^T = g_j XX^T, \tag{18}$$

where we used the change of variables $g_k = g_j^T g_i$, the unitary group assumption and the closure property of the group. Thus

$$[XX^T, g] = \mathbf{0}, \quad \forall g \in \mathcal{G}. \tag{19}$$

Let $v \in \mathbb{R}^d$ be an eigenvector of XX^T with eigenvalue λ, i.e. $XX^T v = \lambda v$. Then, using (19), for any element of the $\mathcal{G}$-orbit of v, $O_v = \{gv | g \in \mathcal{G}\}$:

$$XX^T gv = gXX^T v = \lambda gv, \quad \forall g \in \mathcal{G}, \tag{20}$$

i.e. any element of the orbit O_v is an eigenvector of XX^T with eigenvalue λ. □

Lemma 3. *Let E_λ the eigenspace of XX^T associated to eigenvalue λ and v an arbitrary vector in E_λ. Then the eigenspace coincides with the linear span of the $\mathcal{G}$-orbit of v, i.e. $\mathrm{span}(O_v) = E_\lambda$. Further the representation of $\mathcal{G}$ on E_λ is irreducible, i.e. E_λ is the smallest subspace containing v that is left invariant by the action of the group $\mathcal{G}$.*

Proof. Let $B = \{b_i\}_{i=1}^{|B|}$ an orthogonal basis in E_λ with $|B| = \dim(E_\lambda)$. Any $v \in E_\lambda$ can be expressed as a linear combination of the basis elements

$$v = \sum_{i=1}^{|B|} \alpha_i b_i, \tag{21}$$

and an element of its $\mathcal{G}$-orbit O_v as:

$$gv = \sum_{i=1}^{|B|} \alpha_i g b_i, \quad \forall g \in \mathcal{G}. \tag{22}$$

Since each gb_i is an eigenvector (by Lemma 2), gv is a linear combination of eigenvectors. Clearly this holds also for any linear combination of the orbit elements, i.e., we have $\mathrm{span}(O_v) \subseteq E_\lambda$. Note now that for any two u, $v \in E_\lambda$:

$$\mathrm{span}(O_u) = \mathrm{span}(O_v), \tag{23}$$

i.e., $\mathrm{span}(O_u)$ and $\mathrm{span}(O_v)$ coincide since they both consist of all linear combinations of the set $\{gb_i\}$. This implies:

$$\bigcup_{u \in E_\lambda} \mathrm{span}(O_u) = \mathrm{span}(O_v). \tag{24}$$

However:

$$E_\lambda \subseteq \bigcup_{u \in E_\lambda} \mathrm{span}(O_u) = \mathrm{span}(O_v) \subseteq E_\lambda \tag{25}$$

which implies that $\mathrm{span}(O_v) = E_\lambda$ for any $v \in E_\lambda$ and clearly is the smallest invariant space w.r.t. $\mathcal{G}$ (since any such space is a span of

an orbit of a vector in E_λ). In other words $\mathcal{G}$ acts irreducibly on E_λ. □

We can now prove Lemma 1:

Proof. Let E_λ and $\tilde{E}_{\tilde{\lambda}}$ the eigenspaces and eigenvalues respectively of the matrices XX^T and WW^T. From Lemma 2 we have

$$gE_\lambda = E_\lambda, \;\; \tilde{g}\tilde{E}_{\tilde{\lambda}} = \tilde{E}_{\tilde{\lambda}}, \;\forall\, g \in \mathcal{G}, \;\tilde{g} \in \tilde{\mathcal{G}}. \tag{26}$$

Note that XX^T, WW^T are Hermitian matrices. Now, if two Hermitian matrices commute they have the same eigenspaces and eigenvalues [53], so $E_\lambda = \tilde{E}_{\tilde{\lambda}}$. Thus,

$$\tilde{g}E_\lambda = g\tilde{E}_{\tilde{\lambda}} = gE_\lambda = E_\lambda, \;\forall\, \lambda, \tilde{\lambda}, \;\forall\, g \in \mathcal{G}, \;\forall\, \tilde{g} \in \tilde{\mathcal{G}} \tag{27}$$

or in other words $\tilde{\mathcal{G}}$ and $\mathcal{G}$ preserve, respectively, the eigenspaces of XX^T, WW^T i.e. $[g, WW^T] = [\tilde{g}, XX^T] = 0$. Further, by Lemma 3, the action of the representation of $\mathcal{G}$ and $\tilde{\mathcal{G}}$ restricted to E_λ is irreducible. Since $\mathcal{G}$ and $\tilde{\mathcal{G}}$ act irreducibly on the same subspaces E_λ they are equivalent up to a unitary conjugation, i.e. $\tilde{g} = U^T g U, \;\exists\, U$. □

5.3. Representation learning with symmetry conditions

The main result of this work is summarized in the following Theorem 2, which puts together Proposition 1, Theorem 1 and Lemma 1.

Theorem 2 (Symmetry-adapted regularization). *Let $W \in \mathbb{R}^{d\times|\tilde{\mathcal{G}}|}$ a nonzero matrix whose columns are the elements of an orbit of a finite group $\tilde{\mathcal{G}}$ with cardinality $|\tilde{\mathcal{G}}|$. Then W^TW is permuted and $r(W^TW) = 0$. Further, let $X \in \mathbb{R}^{d\times Q|\mathcal{G}|}$ a matrix whose columns are all elements from the union of Q orbits of a finite group $\mathcal{G}$, with $|\mathcal{G}| = |\tilde{\mathcal{G}}|$. If $[XX^T, WW^T] = \mathbf{0}$ then $\tilde{\mathcal{G}}$ and $\mathcal{G}$ are the same up to a fixed unitary conjugation.*

Proof. Follows by putting together Theorem 1 and Lemma 1. □

Theorem 2 provides the two regularization conditions for the symmetry-adapted representation learning problem (10):

$$\arg\min_{W\in\mathbb{R}^{d\times|\mathcal{G}|}} \mathcal{L}(W, S_N) + \gamma \left\| \left[XX^T, WW^T\right] \right\|_F^2 + \beta r(W^TW) \tag{28}$$

where $\beta, \gamma \in \mathbb{R}_+$ are regularization parameters, $\mathcal{L}'(W, S_N) = \left\| \left[XX^T, WW^T\right] \right\|_F^2$, $\mathcal{R}(W) = r(W^TW)$ and the loss $\mathcal{L}(W, S_N)$ is task-dependent. We use a smooth version of (14), substituting the δ function with a Gaussian function $g_\sigma : \mathbb{R} \to \mathbb{R}$ of width $\sigma \ll 1$; the regularizer is then $r(W^TW) = \tau^T g_\sigma(\mathrm{Cvec}(W^TW))$. The analytic form of the gradients, for use in gradient-based methods for minimizing (28), is provided in Appendix A.

6. Results on unsupervised orbit learning

As a proof of concept, we pose the following unsupervised learning problem: *Given an unlabeled observation set as in (9), namely a union of orbits of the same, finite, unknown group $\mathcal{G}$, learn a single orbit W, of an arbitrary vector $t \in \mathbb{R}^d$, with respect to the (latent) group that generated the data.*

6.1. Synthetic group data

We use synthetic data generated by known permutation groups, of different orders $|\mathcal{G}|$, acting on $\mathbb{R}^6$:

- Cyclic group (C_6), Abelian group, $|\mathcal{G}| = 6$
- Dihedral group (D_6), non-Abelian group, $|\mathcal{G}| = 12$
- Pyritohedral group (T_h), non-Abelian group, $|\mathcal{G}| = 24$

The data X were generated using Q vectors uniformly sampled at random from the unit ball in $\mathbb{R}^6$, for train and validation sets of sizes $Q = 1000$ and $Q = 200$ orbits (or $Q|\mathcal{G}|$ observations).

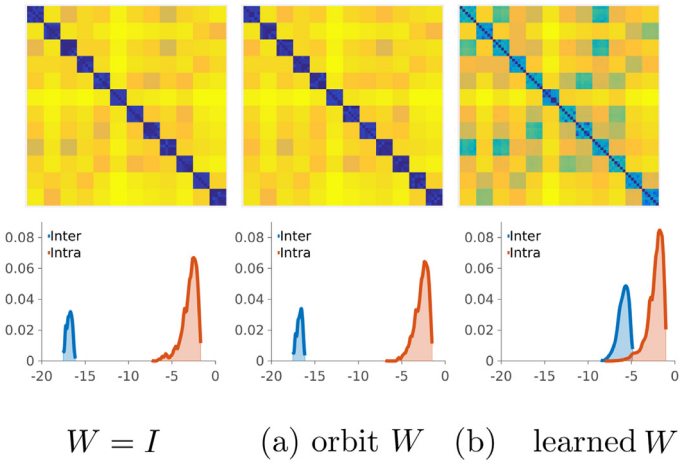

Fig. 1. Permutation sensitivity of the pseudo-metric for a permutation group (Abelian group, Cyclic C_6) and (a) representation through a random orbit of C_6, (b) representation through (unsupervised) learned orbit. *Details on the distance matrices and plots are as in Fig. 2.*

6.2. Optimization for learning

We use (28), without a loss $\mathcal{L}(W, S_N)$, i.e. the commutator norm drives the representation, and $\gamma = 1$. Instead of $r(W^TW)$, since we are testing permutation group, following Corollary 1, we use a sum of regularizers $r(\sigma_j(W)^T\sigma_j(W))$, $j = 1 \ldots J$, where $\sigma_j(x) = \max(0, x - \alpha_j)$ the rectifier function applied pointwise. Then, the minimization problem takes the form

$$\arg\min_{W\in\mathbb{R}^{d\times|\mathcal{G}|}} \left(\left\| [XX^T, WW^T] \right\|_F^2 + \frac{\beta}{J} \sum_{j=1}^{J} r(\sigma(W^T - \alpha_j)\sigma(W - \alpha_j)), \quad \beta \in \mathbb{R}_+ \right. \tag{29}$$

where X is the $d \times N$ matrix storing the observation set, $\sigma(x - \alpha_j) = \max(x - \alpha_j, 0)$ is the rectifier nonlinearity with parameter α_j, J is the number of nonlinearities and β is the regularization constant. Note that the scheme with a single $r(W^TW)$ term can be derived as a special case by setting $J = 1$ and $\alpha_1 < < 0$.

Minimization was performed using quasi-Newton iterative optimization with a cubic line search, using Broyden-Fletcher-Goldfarb-Shanno (BFGS) [54] for the Hessian matrix approximation. The matrix W is set to size $d \times |\mathcal{G}|$, assuming the target orbit size $|\mathcal{G}|$ is given, and initialized at random. For each β, we initialize and run the minimization process m times to obtain m orbit solutions, corresponding to different local minima of the loss. For the results we used schemes with J set to 100 and 300, selecting α_j values uniformly spaced in $[-1, 1]$. For β, we used a range of orders of magnitude, namely $\log_{10}\beta = \{0, \ldots, -7\}$. For the learned weights (Figs. 1(b) and 2(b)) correspond to the minimum loss (commutator norm) solution for $m = 50$ random initializations of W. In Fig. 1 (C_6), we used $\beta = 10^{-4}, J = 100, |\mathcal{G}| = 6$. In Fig. 2 we used $J = 300$, $\beta = 10^{-5}, |\mathcal{G}| = 12$ (for D_6) and $\beta = 10^{-6}, |\mathcal{G}| = 24$ (for T_h).

6.3. Equivariance of learned data representations

To test symmetry structure in a representation W for the solutions given by the algorithm we used the fact that (Section 2) a representation Φ of the form of (4) is permutation-equivariant and thus if two signals $x, x' \in \mathbb{R}^d$ are part of an orbit O_x then their representations $\Phi(x)$, $\Phi(x')$, with respect to an orbit dictionary W, are permuted vectors if W is an orbit of a vector generated by the same group. We can then define the following pseudometric

$$D(x, x') = r([\Phi(x), \Phi(x')]) = r(W^T[x, x']), \tag{30}$$

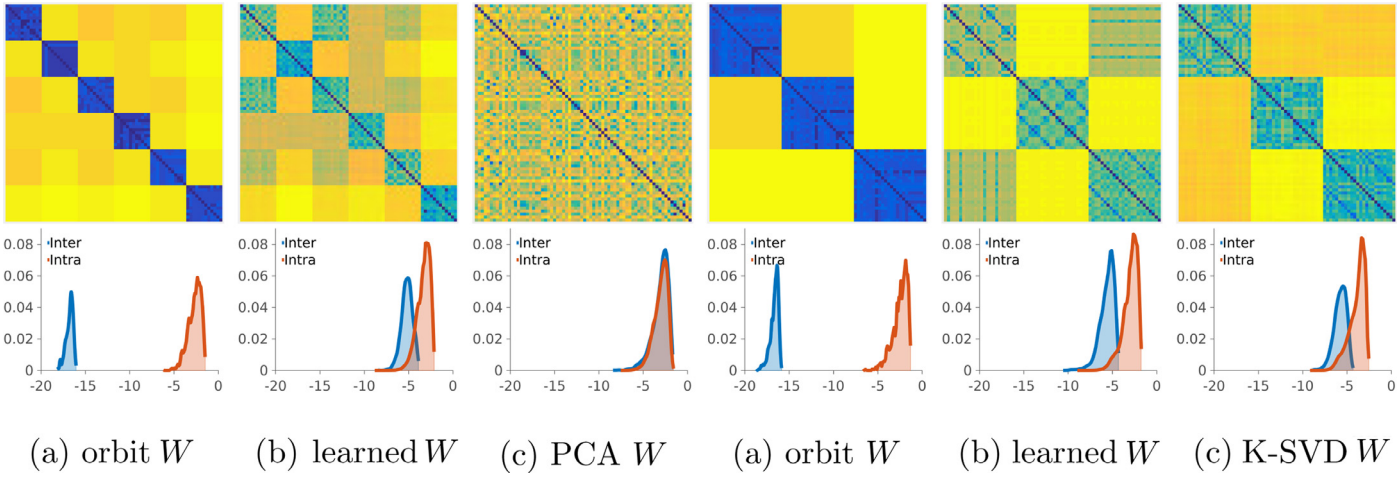

Fig. 2. Dihedral group D_6 of order 12 (a-c, left) and Pyritohedral group T_h of order 24 (a-c, right): pseudo-distance $r([\Phi(x_i), \Phi(x_j)])$ (30), where $\Phi(x_i) = W^T x_i$ and W a 6×12 and 6×24 dictionary, for different representations. Learning performed from a training set of 1000 orbits. Shown are the distance matrices (top) for 6 orbits (arranged in 6 blocks of 12 elements) and 3 orbits (arranged in 3 blocks of 24 elements) and inter/intra-orbit distributions on log x-axis (bottom) for 50 orbits from a validation set.

where the input to $r : \mathbb{R}^{m\times 2} \to \mathbb{R}_+$ is now the $m \times 2$ matrix $(\Phi(x), \Phi(x'))$. Details are provided in Appendix B. Using D we can check if W induces an equivariant representation and, for the case of matrices learned with the proposed method, if the equivariance is induced by a matrix with the symmetries in the data. For evaluation purposes only, we consider a labeled validation set, i.e. the orbit assignment for each point is known.

In Figs. 1 and 2 we evaluate the following cases: a) W is an orbit selected from the same distribution as the training set, i.e. a ground truth W; this serves as a sanity check that the same symmetry will indeed induce equivariance in $\Phi(x)$, b) W is a learned representation via minimization of (29), and c) W are the eigenvectors of the data covariance (PCA), to check if symmetries can be discovered through high-variance directions or the result of a sparse dictionary learning algorithm (K-SVD [55]). The last two are included as baseline comparisons to check if symmetries can be discovered through projections in max-variance directions or sparsity/reconstruction constraints. In Fig. 1 we also depict the case $W = I_{d\times|\mathcal{G}|}$, corresponding to $\Phi(x) = x$. This serves as a sanity check for the validity of the pseudodistance as, for the case of permutation groups, each orbit is composed from permuted vectors.

The figures depict the distance matrices (of 12, 6 and 3 random orbits resp.) for C_6, D_6 and T_h and the probability distributions of pairwise distances for the validation set (50 orbits and 300 points in total), estimated for inter- and intra-orbit pairs separately. The block diagonal structure of the distance matrices support the main claims of this work (Theorem 2), and the separability of the inter- and intra- distributions suggest that the learned W implies a quotient representation with respect to the unknown group of transformations.

7. Conclusion

We studied the problem of learning data symmetries, in particular group symmetries, as a prior on structure due to transformations in the data. Our motivation was to derive representations that are adapted to symmetries and reduce the sample complexity of downstream supervised learning. In particular we explored mathematical conditions that can drive, in an unsupervised way, a learned representation to reflect the symmetries in an unlabeled training set.The approach is particularly relevant for data that have a low-dimensional intrinsic structure (e.g. transforming images or sounds) and can be applied to any finite group since it reduces their analysis to permutation groups whose size |\mathcall(G)|can be chosen to be efficiently computable. This is in contrast with other proposed methods, [41,42] where transformations matrices of size $d^2 \gg$ |\mathcall(G)^2| are learned from data.

In this work, we focused on global group transformations as a proof of concept. However, the same theoretical framework can be applied to real data, for dealing with non-group, arbitrary transformations. The key notion for relaxing the assumptions on the observation set is locality. Arbitrary transformations can be approximated by smooth transformations of a number of instances, on local neighborhoods (of each instance). The transformation can then be described in terms of a group, i.e. the Lie group associated to the tangent space of the manifold of the signal transformations. The proposed regularization scheme can then be applied for learning localized filters, each representing a small, local signal neighborhood. This would be applicable to a general family of CNNs, where local filters are employed at each layer for local, position-robust feature detection.

Additional directions for future work include addressing the dependency on noise (both in the unsupervised setting and supervised/semisupervised setting, [56–58]), size of the dictionary and partial orbits. For the latter, Lemma 1 can be restated in terms of empirical covariance matrices and the validity of the approximation can be measured using concentration inequalities. Interestingly, group codes (first studied by Slepian in his seminal paper [59]) or dictionaries ([60]) tend to have low self-coherence which is related to the exact recovery condition of a signal x of given sparsity [61].

The long-term goal of this work is to learn data-symmetries driven group convolutions and network topologies in CNNs. Imposing structure in subsets of weights, corresponding to multiple filters transforming under one or more groups, will necessitate an extension to learning multiple orbits or multiple symmetries. Moreover, extensions to multilayer representations, e.g. deep networks, can be formulated by noting that the representation will be permutation-equivariant after the first layer, given a map as in (5). Subsequent layers will then process permutation-transformed signals.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and the Italian Institute of Technology. The authors gratefully acknowledge Alan Yuille, Maximilian Nickel, Silvia Villa and Carlo Cilliberto, for the insightful discussions and feedback on

early versions of this work, and Saverio Salzo for pointing to the proof of Lemma 3.

Appendix

In the following we let $\mathcal{G}, \tilde{\mathcal{G}}$ two finite groups with cardinality $|\mathcal{G}|$ and $X \in \mathbb{R}^{d\times Q|\mathcal{G}|}$ a matrix whose columns are all elements from the union of orbits of Q vectors in $\mathbb{R}^d$ w.r.t. $\mathcal{G}$.

Appendix A. Analytic gradients

Regularizer gradient:
We calculate the derivative of $r(W^TW) = \tau^T g_\sigma(C\text{vec}(W^TW))$ with respect to the vectorization of W; using the chain rule we have:

$$\frac{\partial r(W^TW)}{\partial \text{vec}(W)} = \left(\frac{\partial\langle\tau, p\rangle}{\partial p}\frac{\partial p}{\partial q}\frac{\partial q}{\partial s}\frac{\partial s}{\partial \text{vec}(W)}\right)^T \tag{A.1}$$

with $p = g_\sigma(q)$, $q = Cs$ and $s = \text{vec}(W^TW)$. The first three factors are easy to calculate and the right side of Eq. (A.1) is:

$$\frac{\partial r(W^TW)}{\partial \text{vec}(W)} = -\frac{2}{\sigma^2}\Big(\tau^T \text{diag}\Big((C\text{vec}(G)) \odot g_\sigma(C\text{vec}(W^TW))\Big) \times C\frac{\partial\ \text{vec}(G)}{\partial \text{vec}(W)}\Big)^T \tag{A.2}$$

where $\odot$ denotes the Hadamard product and $\text{diag}(\cdot)$ a $(|\mathcal{G}|^2(|\mathcal{G}|^2-1)/2) \times (|\mathcal{G}|^2(|\mathcal{G}|^2-1)/2)$ diagonal matrix. For the last partial derivative we have:

$$\frac{\partial \text{vec}(W^TW)}{\partial \text{vec}(W)} = (\mathbb{I}_{|\mathcal{G}|^2} + T)(\mathbb{I}_{|\mathcal{G}|} \otimes W^T) \tag{A.3}$$

where the matrix T is defined as $T\text{vec}(v) = \text{vec}(v^T)$ and $\mathbb{I}_{|\mathcal{G}|^2}$, $\mathbb{I}_{|\mathcal{G}|}$ are the identity matrices of dimensions $|\mathcal{G}|^2 \times |\mathcal{G}|^2$, $|\mathcal{G}| \times |\mathcal{G}|$. It can be shown that T can be written explicitly as:

$$\begin{cases} T_{ij} = 1, \ if\ j = 1 + |\mathcal{G}|(i-1) - (|\mathcal{G}|^2-1)\text{floor}\left(\frac{i-1}{|\mathcal{G}|}\right) \\ T_{ij} = 0, \ \text{otherwise.} \end{cases} \tag{A.4}$$

Putting everything together, right hand side of (A.1) becomes:

$$\frac{\partial r(W^TW)}{\partial \text{vec}(W)} = -\frac{2}{\sigma^2}(\mathbb{I}_{|\mathcal{G}|} \otimes W)P^T(C\text{vec}(W^TW)) \odot g_\sigma \times (C\text{vec}(W^TW)) \odot \tau \tag{A.5}$$

with $P = C(\mathbb{I}_{|\mathcal{G}|^2} + T)$ a constant matrix, which can be pre-calculated given the group cardinality. Correctness of the gradient has been tested with numerical simulations.

Commutator gradient: It is easy to derive the analytic gradient of the norm $\|M(X,W)\|_F^2 = \text{trace}(M(X,W)^TM(X,W))$, where $M(X,W) = [XX^T, WW^T]$ is the commutator of the data and representation covariance matrices. A simple calculation shows that the gradient has the analytic form:

$$\frac{\partial \|M(X,W)\|_F^2}{\partial W} = -4[M(X,W), XX^T]W, \tag{A.6}$$

by expanding the gradient as follows:

$$\begin{aligned} &\tfrac{\partial}{\partial W}\left(-2Tr(XX^TWW^TXX^TWW^T) + 2Tr(XX^TWW^TWW^TXX^T)\right) \\ &= -8XX^TWW^TXX^T + 4XX^TXX^TWW^TW + 4WW^TXX^TXX^TW \\ &= -4[XX^T, WW^T]XX^TW + 4XX^T[XX^T, WW^T]W \\ &= -4[[XX^T, WW^T], XX^T]W \\ &= -4[M(X,W), XX^T]W. \end{aligned}$$

Appendix B. Permutation invariant pseudometric

Quantifying equivalence under permutations is a way to test for equivalence under general symmetries, provided that W has the same symmetries, and thus a way to quantify same-symmetry, orbit structure in a given W. To measure equivalence under permutations, we can use a modified version of the regularizer function (14), where the input is a 2-column matrix instead of the Gram matrix. Indeed the representations $\Phi(x)$, $\Phi(x')$ of two elements $x, x' \in \mathbb{R}^d$ of the same orbit O_x, w.r.t. a same-group orbit dictionary W are permuted vectors.

Given matrix $W \in \mathbb{R}^{d\times m}$, and the corresponding representation $\Phi(x) = W^Tx = (\langle w_1, x\rangle, \dots \langle w_m, x\rangle)$, $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, let

$$D(x,x') = r([\Phi(x), \Phi(x')]) = \tau^T\delta(C[\Phi(x), \Phi(x')])$$

where the input to r is now the $m \times 2$ matrix $[\Phi(x), \Phi(x')]$. The quantity $r : \mathbb{R}^{m\times 2} \to \mathbb{R}_+$ is a pseudometric and

$$\Phi(x) = P_g\Phi(x') \iff r([\Phi(x), \Phi(x')]) = 0. \tag{B.1}$$

In addition, if the conditions $\left\|[XX^T, WW^T]\right\|_F^2 = 0$ and $r(W^TW) = 0$ are also satisfied, as is the case for W learned by minimizing (29), we have the following.

Lemma 4. *If W satisfies the conditions of Theorem 2 and is invertible, then:*

$$r(W^T[x,x']) = 0 \iff x = gx',\ \forall x, x' \in \mathcal{X},\ \ \forall g \in \mathcal{G}. \tag{B.2}$$

Proof. First note that the Gram matrix W^TW identifies W up to an arbitrary (but fixed) unitary transformation U; together with the condition $r(W^TW) = 0$, this implies that W can be written as $W = UW_t$, where W_t is a group orbit matrix of some vector $t \in \mathbb{R}^d$ and $W_t^TW_t = W^TW$.

For the direct implication, if $r(W^T[x,x']) = 0$ then there exists a permutation matrix P_g such that

$$W^Tx = P_gW^Tx' \Rightarrow W_t^TU^Tx = P_gW_t^TU^Tx' = W_t^TgU^Tx' \tag{B.3}$$

since the action of g on W_t is a permutation of its columns i.e. $gW_t = W_tP_g$. If W is invertible from the previous equation we have

$$U^Tx = gU^Tx', \tag{B.4}$$

which suffices since the rotation U^T is irrelevant from the learning point of view since $U^T\mathcal{X} = \mathcal{X}$. For the inverse implication we proceed by contradiction. Suppose $x = gx'$ but $r([\Phi(x), \Phi(x')]) \neq 0$. Then:

$$W^Tx \neq PW^Tx', \quad x = gx' \tag{B.5}$$

for all permutations P. By the assumptions

$$(W)_{:,i} = (UW_t)_{:,i} = Ug_it = Ug_iU^TUt = \tilde{g}_i\tilde{t} = (\widetilde{W}_{\tilde{t}})_{:,i}, \tag{B.6}$$

i.e. the columns of matrix W are an orbit of vector $\tilde{t}$ w.r.t. the conjugate representation $\tilde{\mathcal{G}}$ of $\mathcal{G}$ induced by U, namely $W_t = \widetilde{W}_{\tilde{t}}$. We can then write (B.5) as

$$\widetilde{W}_{\tilde{t}}^Tx \neq P\widetilde{W}_{\tilde{t}}^Tx' = \widetilde{W}_{\tilde{t}}^T\tilde{g}x',\ \ \forall \tilde{g} \in \tilde{\mathcal{G}}. \tag{B.7}$$

As W_t is invertible, so is $\widetilde{W}_{\tilde{t}}$. Thus, from (B.7) we have that $x \neq \tilde{g}x'$ and we can conclude as in the direct implication. □

Lemma 4 states that r is zero for elements of the same orbit (inter-orbit distance) and large for elements of different orbits (intra-orbit distance). Examples of the distributions of these distances are shown in Figs. 1 and 2. Concretely, the value of r for pairs of vectors can be used to test the following: (a) are $\Phi(x)$ and $\Phi(x')$ permuted vectors (from (B.1))? and (b) for $\Phi(x) = W^Tx$ and W satisfying the symmetry conditions, are x, x' on the same orbit ? And, as an extension, does the representation matrix W provide an equivariant map, i.e. are the column vectors an orbit *of the same group*?

References

[1] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133.
[2] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 193–202.
[3] T.M. Caelli, Z.-Q. Liu, On the minimum number of templates required for shift, rotation and size invariant pattern recognition, Pattern Recognit. 21 (3) (1988) 205–216.
[4] R. Lenz, Group invariant pattern recognition, Pattern Recognit. 23 (1) (1990) 199–217.
[5] P. Foldiak, Learning invariance from transformation sequences, Neural Comput. 3 (2) (1991) 194–200.
[6] A. Grace, M. Spann, A comparison between Fourier-Mellin descriptors and moment based features for invariant object recognition using neural networks, Pattern Recognit. Lett. 12 (10) (1991) 635–643.
[7] J. Flusser, T. Suk, Pattern recognition by affine moment invariants, Pattern Recognit. 26 (1) (1993) 167–174.
[8] B.A. Olshausen, C.H. Anderson, D.C. Van Essen, A multiscale dynamic routing circuit for forming size- and position-invariant object representations, J. Comput. Neurosci. 2 (1) (1995) 45–62.
[9] L. Van Goola, T. Moons, E. Pauwels, A. Oosterlinck, Vision and Lie's approach to invariance, Image Vis. Comput. 13 (4) (1995) 259–277.
[10] M. Michaelis, G. Sommer, A Lie group approach to steerable filters, Pattern Recognit. Lett. 16 (11) (1995) 1165–1174.
[11] J. Wood, Invariant pattern recognition: a review, Pattern Recognit. 29 (1) (1996) 1–17.
[12] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nat. Neurosci. 2 (11) (1999) 1019–1025.
[13] M. Lessmann, R.P. Wrtz, Learning invariant object recognition from temporal correlation in a hierarchical network, Neural Netw. 54 (2014) 70–84.
[14] K. Lenc, A. Vedaldi, Understanding image representations by measuring their equivariance and equivalence, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.
[15] Z. Shao, Y. Li, Integral invariants for space motion trajectory matching and recognition, Pattern Recognit. 48 (8) (2015) 2418–2432.
[16] F. Anselmi, L. Rosasco, T. Poggio, On invariance and selectivity in representation learning, Inf. Inference 5 (2) (2015) 134–158.
[17] T.S. Cohen, M. Welling, Group equivariant convolutional networks, in: International Conference on Machine Learning (ICML), 2016.
[18] A. Achille, S. Soatto, Emergence of invariance and disentangling in deep representations, in: Proceedings of the ICML Workshop on Principled Approaches to Deep Learning, 2017.
[19] S. Soatto, Steps towards a theory of visual information: active perception, signal-to-symbol conversion and the interplay between sensing and control, arXiv:1110.2053, 2011.
[20] B. Haasdonk, H. Burkhardt, Invariant kernel functions for pattern analysis and machine learning, Mach. Learn. 68 (1) (2007) 35–61.
[21] F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio, Unsupervised learning of invariant representations, Theor. Comput. Sci. 633 (2016) 112–121.
[22] Y.S. Abu-Mostafa, Learning from hints in neural networks, J. Complex. 6 (2) (1990) 192–198.
[23] R. Gens, P.M. Domingos, Deep symmetry networks, in: Advances in Neural Information Processing System (NIPS), 2014, pp. 2537–2545.
[24] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[25] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[26] V. Papyan, Y. Romano, M. Elad, Convolutional neural networks analyzed via convolutional sparse coding, J. Mach. Learn. Res. 18 (1) (2017) 2887–2938.
[27] S. Zhang, J. Wang, X. Tao, Y. Gong, N. Zheng, Constructing deep sparse coding network for image classification, Pattern Recognit. 64 (2017) 130–140.
[28] M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, Advances in Neural Information Processing Systems (NIPS), 2015.
[29] J.J. Kivinen, C.K.I. Williams, Transformation equivariant Boltzmann machines, in: International Conference on Artificial Neural Networks (ICANN), 2011.
[30] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.
[31] L. Toth, Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 190–194.
[32] K. Kavukcuoglu, M. Ranzato, R. Fergus, Y. LeCun, Learning invariant features through topographic filter maps, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, 2009, pp. 1605–1612.
[33] Q.V. Le, J. Ngiam, Z. Chen, D. Chia, P.W. Koh, A.Y. Ng, Tiled convolutional neural networks, in: Advances in Neural Information Processing Systems 23, 2010, pp. 1279–1287.
[34] S. Dieleman, J. De Fauw, K. Kavukcuoglu, Exploiting cyclic symmetry in convolutional neural networks, in: International Conference on Machine Learning (ICML), 2016, pp. 1889–1898.
[35] L. Sifre, S. Mallat, Rotation, scaling and deformation invariant scattering for texture discrimination, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 1233–1240.
[36] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886.
[37] M.I. Khalil, M.M. Bayoumi, Invariant 2d object recognition using the wavelet modulus maxima, Pattern Recognit. Lett. 21 (9) (2000) 863–872.
[38] T.S. Cohen, M. Welling, Steerable CNNs, in: International Conference on Learning Representations (ICLR), 2017.
[39] S. Ravanbakhsh, J. Schneider, B. Poczos, Equivariance through parameter-sharing, Proceedings of the 34th International Conference on Machine Learning 70 (2017) 2892–2901.
[40] C. Zhang, S. Voinea, G. Evangelopoulos, L. Rosasco, T. Poggio, Discriminative template learning in group-convolutional networks for invariant speech representations, in: INTERSPEECH 2015, Annual Conf. of the International Speech Communication Association, 2015, pp. 3229–3233.
[41] X. Miao, R.P.N. Rao, Learning the Lie groups of visual invariance, Neural Comput. 19 (10) (2007) 2665–2693.
[42] R.P.N. Rao, D.L. Ruderman, Learning Lie groups for invariant visual perception, Adv. Neural Inf. Process. Syst. 11 (1999) 810–816.
[43] J. Sohl-Dickstein, C.M. Wang, B.A. Olshausen, An unsupervised algorithm for learning lie group transformations, 2010, arXiv:1001.1027.
[44] C.M. Wang, J. Shol-Dickstein, I. Tosic, B.A. Olshausen, Lie group transformation models for predictive video coding, in: Data Compression Conference Proceedings, IEEE, 2011, pp. 83–92.
[45] T.J. Sejnowski, P.K. Kienker, G.E. Hinton, Learning symmetry groups with hidden units: beyond the perceptron, Physica D 22 (1–3) (1986) 260–275.
[46] R. Negrinho, A. Martins, Orbit regularization, in: Advances in Neural Information Processing Systems (NIPS), 2014, pp. 3221–3229.
[47] S. Mallat, Group invariant scattering, Commun. Pure Appl. Math. 65 (10) (2012) 1331–1398.
[48] P.G. Casazza, G. Kutyniok, Finite Frames: Theory and Applications, Birkhauser Basel, 2012.
[49] Y. Eldar, Least-squares inner product shaping, Linear Algebra Appl. 348 (1–3) (2002) 153–174.
[50] A. Cayley, On the theory of groups, as depending on the symbolic equation $\theta^n = 1$, Philos. Mag. Series 4 VII (42) (1854) 40–47,408–409.
[51] J.J. Rodrigues, P.M.Q. Aguiar, J.M.F. Xavier, ANSIG - an analytic signature for permutation-invariant two-dimensional shape representation, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[52] Y. Eldar, H. Bolcskei, Geometrically uniform frames, IEEE Trans. Inf. Theory 49 (4) (2003) 993–1006.
[53] A. Laub, Matrix Analysis, Cambridge University Press, 2005.
[54] J. Nocedal, S.J. Wright, Numerical Optimization, second ed., Springer, 2006.
[55] M. Aharon, M. Elad, A.M. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[56] B. Bo Han, I. Tsang, L. Chen, C. P., C. Yu, S. Fung, Progressive stochastic learning for noisy labels, IEEE Trans. Neural Netw. Learn. Syst. (2018).
[57] B. Han, Y. Pan, I. Tsang, Robust Plackettluce model for k-ary crowdsourced preferences, Machine Learning 107 (2018) 675–702.
[58] C. Gong, T. Liu, D. Tao, K. Fu, E. Tu, J. Yang, Deformed graph Laplacian for semisupervised learning, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015) 2261–2274.
[59] D. Slepian, Group codes for the Gaussian channel, Bell Syst. Tech. J. 47 (4) (1968) 575–602.
[60] M. Thill, B. Hassibi, Group frames with few distinct inner products and low coherence, IEEE Trans. Signal Process. 63 (19) (2015) 5222–5237.
[61] D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. Inf. Theory 47 (7) (2001) 2845–2862.

Anselmi Fabio is a postdoctoral fellow in the Istituto Italiano di Tecnologia and the Laboratory for Computational and Statistical Learning at MIT and part of the Center for Brains, Minds, and Machines.