# A COLLECTED PROOFS OF THEORETICAL RESULTS

In this appendix, we collect proofs of the theoretical results that were not given in Section 3. Most of the claims are well-known, and we include them mainly to keep the article self-contained.

We begin by proving a small claim we will use implicitly throughout.

**Lemma 7.** *The lifted representations $\overline{\rho}_i$ on $\mathrm{Hom}(X_i, X_{i+1})$, $\overline{\rho}$ on $\mathcal{L}$ and $\overline{\rho}^{\otimes 2}$ on $\mathcal{L}^{\otimes 2}$ are unitary with respect to the canonical inner products.*

*Proof.* Let us begin by quickly establishing that the $\overline{\rho}_i$ are representations:

$$\overline{\rho}_i(g)\overline{\rho}_i(h)A_i = \rho_{i+1}(g)\rho_{i+1}(h)A_i\rho_i(h)^{-1}\rho_i(g)^{-1} = \rho_{i+1}(gh)A_i\rho_i(gh)^{-1} = \overline{\rho}(gh)A_i \tag{28}$$

Now let us move on to the unitarity. We have

$$\begin{aligned}\langle \overline{\rho}_i(g)A_i, \overline{\rho}_i(g)B_i\rangle &= \mathrm{tr}((\rho_{i+1}(g)A_i\rho_i(g)^{-1})^*\rho_{i+1}(g)B_i\rho_i(g)^{-1}) \\ &= \mathrm{tr}(\rho_i(g)A_i^*\rho_{i+1}(g)^{-1}\rho_{i+1}(g)B_i\rho_i(g)^{-1}) \\ &= \mathrm{tr}(A_i^*B_i\rho_i(g)^{-1}\rho_i(g)) = \mathrm{tr}(A_i^*B_i) = \langle A_i, B_i\rangle\end{aligned} \tag{29}$$

which immediately implies

$$\langle \overline{\rho}(g)A, \overline{\rho}(g)B\rangle = \sum_{i\in[L]} \langle \overline{\rho}_i(g)A_i, \overline{\rho}_i(g)B_i\rangle = \sum_{i\in[L]} \langle A_i, B_i\rangle = \langle A, B\rangle \tag{30}$$

proving the unitarity of $\overline{\rho}_i$ and then $\overline{\rho}$. Unitary of $\overline{\rho}^{\otimes 2}$ then follows from the fundamental properties of the tensor product:

$$\begin{aligned}\langle \overline{\rho}^{\otimes 2}(g)(A\otimes B), \overline{\rho}^{\otimes 2}(g)(C\otimes D)\rangle &= \langle (\overline{\rho}(g)A\otimes\overline{\rho}(g)B), (\overline{\rho}(g)C\otimes\overline{\rho}(g)D)\rangle \\ &= \langle \overline{\rho}(g)A, \overline{\rho}(g)C\rangle\langle \overline{\rho}(g)B, \overline{\rho}(g)D\rangle = \langle A, C\rangle\langle B, D\rangle \\ &= \langle (A\otimes B), (C\otimes D)\rangle\,.\end{aligned} \tag{31}$$

□

Next, we prove the lemmas concerning the lifted representation $\overline{\rho}$.

*Proof of Lemma 1.* The statement follows immediately from the fact the a tuple of layers $A \in \mathcal{L}$ is in $\mathcal{E}$ if and only if $\rho_{i+1}(g)A_i = A_i\rho_i(g)$, which is equivalent to $A_i = \rho_{i+1}(g)A_i\rho_i(g)^{-1} = (\overline{\rho}(g)A)_i$ for all $i \in [L]$ and $g \in G$. □

*Proof of Lemma 2.* Let $P$ be the operator on $\mathcal{L}$ defined by

$$PA = \int_G \overline{\rho}(g)A\,\mathrm{d}\mu(g). \tag{32}$$

To show that $P = \Pi_{\mathcal{E}}$, we first need to show that if $PA \in \mathcal{E}$ for any $A \in \mathcal{L}$. To do this, it suffices by Lemma 1 to prove that $\overline{\rho}(g)PA = PA$ for any $g \in G$. Using the fact that $\overline{\rho}$ is an representation of $G$, and the invariance of the Haar measure, we obtain

$$\overline{\rho}(g)PA = \int_G \overline{\rho}(g)\overline{\rho}(h)A\,\mathrm{d}\mu(h) = \int_G \overline{\rho}(gh)A\,\mathrm{d}\mu(h) = \int_G \overline{\rho}(h')A\,\mathrm{d}\mu(h') = PA. \tag{33}$$

Next, we need to show that $PA = A$ for any $A \in \mathcal{E}$. But since $\overline{\rho}(g)A = A$ for such $A$, we immediately obtain

$$PA = \int_G \overline{\rho}(g)A\,\mathrm{d}\mu(g) = \int_G A\,\mathrm{d}\mu(g) = A. \tag{34}$$

Consequently, $P : \mathcal{L} \to \mathcal{E}$ is a projection. Finally, to establish that $P$ is also orthogonal, we need to show that $\langle A - PA, B\rangle = 0$ for all $A \in \mathcal{L}$, $B \in \mathcal{E}$. This is a simple consequence of the unitarity of $\overline{\rho}$ and Lemma 1,

$$\langle PA, B\rangle = \int_G \langle \overline{\rho}(g)A, B\rangle\,\mathrm{d}\mu(g) = \int_G \langle \overline{\rho}(g)A, \overline{\rho}(g)B\rangle\,\mathrm{d}\mu(g) = \int_G \langle A, B\rangle\,\mathrm{d}\mu(g) = \langle A, B\rangle\,, \tag{35}$$

which completes the proof that $P = \Pi_{\mathcal{E}}$. □

*Proof of Lemma 5.* Replacing $\mathcal{L}$ with $\mathcal{L} \otimes \mathcal{L}$, $\mathcal{E}$ with $\mathcal{E}^{\otimes 2}$ and $\overline{\rho}$ with $\overline{\rho}^{\otimes 2}$, the proof is identical to that of Lemma 2. □

*Proof of Lemma 6.* For $A \in \mathcal{E}$ we have $\overline{\rho}(g)A = A$ for any $g \in G$ according to Lemma 1. Consequently,

$$\left(\Pi_{\mathcal{E}^{\otimes 2}} M\right)[A, A] = \int_G \overline{\rho}^{\otimes 2}(g) M[A, A] d\mu(g) = \int_G M[\overline{\rho}(g)A, \overline{\rho}(g)A] d\mu(g) \tag{36}$$

$$= \int_G M[A, A] d\mu(g) = M[A, A]\,. \tag{37}$$

which proves the first statement. To prove the second, we use that for $B \in \mathcal{E}^{\perp}$, $\Pi_{\mathcal{E}} B = 0$ by definition. A similar calculation as above now yields

$$\left(\Pi_{\mathcal{E}^{\otimes 2}} M\right)[A, B] = \int_G \overline{\rho}^{\otimes 2}(g) M[A, B] d\mu(g) = \int_G M[\overline{\rho}(g)A, \overline{\rho}(g)B] d\mu(g) \tag{38}$$

$$= \int_G M[A, \overline{\rho}(g)B] d\mu(g) = M[A, \Pi_{\mathcal{E}} B] = 0\,, \tag{39}$$

where we have used the bilinearity of $M$ repeatedly to complete the proof. □

## B Examples of representations

In addition to the examples provided in the main part of the paper, the following representations will feature in the experimental design described below.

**Example 3.** *The canonical action of the permutation group $S_N$ on $\mathbb{R}^N$ is defined through $(\rho^{\mathrm{perm}}(\pi)v)_i = v_{\pi^{-1}(i)}, i \in [n]$, i.e., an element acts via permuting the elements of a vector. This action induces an action on the tensor space $(\mathbb{R}^N)^{\otimes k} = \mathbb{R}^N \otimes \mathbb{R}^N \otimes \cdots \otimes \mathbb{R}^N$: $(\rho^{\mathrm{perm}}(\pi)T)_{i_0,\ldots,i_{k-1}} = T_{\pi^{-1}(i_0),\ldots,\pi^{-1}(i_{k-1})}$ which is important for graphs. For instance, when applied to $\mathbb{R}^N \otimes \mathbb{R}^N$, it encodes the effect a re-ordering of the nodes of a graph has on the adjacency matrix on the graph.*

**Example 4.** *$C_4 \cong \mathbb{Z}_4$ acts on images $x \in \mathbb{R}^{N,N}$ through rotations by multiples of $90°$. That is, $\rho^{\mathrm{rot}}(k) = \rho^{\mathrm{rot}}(1)^k$, $k = 0, 1, 2, 3$ where $(\rho^{\mathrm{rot}}(1)x)_{i,j} = x_{-j,i}$.*

## C The converse of Proposition 3 does not hold

To construct a counterexample, it is clearly enough to construct $U \in \mathcal{L} \otimes \mathcal{L}$ such that $U$ is not positive definite but $\Pi_{\mathcal{E}^{\otimes 2}} U$, and $V \in \mathcal{E}^{\otimes 2}$ such that $V$ is not positive definite but $\Pi_{\mathcal{E}}^{\otimes 2} V$ is.

We choose $\mathcal{L} = \mathbb{R}^N$ and $G = S_N$ with the canonical representation $\overline{\rho} = \rho^{\mathrm{perm}}$ as in Example 3. With $e_i$, $i = 1, \ldots, N$, the standard ONB of $\mathcal{L}$ we can construct the equivariant subspaces of $\mathcal{L}$ and $\mathcal{L} \otimes \mathcal{L}$ using

$$\mathbb{1} = \sum_{i=1}^{N} e_i \in \mathcal{L}\,, \quad \mathrm{id} = \sum_{i=1}^{N} e_i \otimes e_i \in \mathcal{L}^{\otimes 2} \tag{40}$$

as $\mathcal{E} = \mathrm{Span}\{\mathbb{1}\}$ and $\mathcal{E}^{\otimes 2} = \mathrm{Span}\{\mathrm{id}, \mathbb{1} \otimes \mathbb{1}\}$ Maron et al. (2019a). Furthermore, ONBs of the two spaces are given by

$$B_{\mathcal{E}}^1 = \tfrac{1}{\sqrt{N}}\mathbb{1} \qquad \text{and } B_{\mathcal{E}^{\otimes 2}}^1 = \tfrac{1}{\sqrt{N}}\mathrm{id}, B_{\mathcal{E}^{\otimes 2}}^2 = \tfrac{1}{\sqrt{N(N-1)}}(\mathbb{1} \otimes \mathbb{1} - \mathrm{id})\,,$$

respectively.

Now consider the matrix $U = (N+1)e_1 \otimes e_1 - e_2 \otimes e_2 \in \mathcal{L}^{\otimes 2}$. It is not positive definite, since $U[e_2, e_2] = -1$. However, as a direct calculation reveals, its projection to $\mathcal{E}^{\otimes 2}$, $\Pi_{\mathcal{E}^{\otimes 2}} U = \mathrm{id}$, clearly is.

Second, we can construct a matrix $V = -\mathrm{id} + \frac{2}{N}(\mathbb{1} \otimes \mathbb{1}) \in \mathcal{E}^{\otimes 2}$ which is not positive definite, since its restriction to $\mathcal{E}^{\perp}$ is $-\mathrm{id}$. However, the projection $\Pi_{\mathcal{E}}^{\otimes 2} V = \frac{1}{N}\mathbb{1} \otimes \mathbb{1}$ is positive definite when restricted to $\mathcal{E}$.

# D EXPERIMENTAL DETAILS

Here, we provide a more detailed description of the experiments in Section 5 in the main paper. Each experiment used a single NVIDIA Tesla T4 GPU with 16GB RAM, but many of them were performed in parallell on a cluster. The experiments presented here took in total about 75 GPU hours.

The code is provided in the supplementary material.

## D.1 THE PERM EXPERIMENT

In the PERM experiment, we train a model to detect whether a graph is connected or not. The model takes adjacency matrices $M \in \mathbb{R}^N \otimes \mathbb{R}^N$ of the graph as input. The training task is obviously invariant to permutations of the adjacency matrix (where the action of $S_N$ on the matrix space was defined in Example 3 in the main paper), since such a permutation does not change the underlying graph. The invariance of the network is conveniently expressed by letting $S_N$ acting trivially on the output space $Y = \mathbb{R}$.

**Data** We consider graphs of size $N = 10$ drawn according to a stochastic block model: We divide the nodes into two randomly drawn clusters $I, J$, $I \cap J = \emptyset$, $I \cup J = \{0, \dots, 10\}$. Within each cluster, a bidirected edge is added between each pair of nodes with a probability of $0.5$, and in between the clusters with a probability of $0.05$. The graphs are subsequently checked for connectivity by checking that all entries of $\sum_{i=0}^{N} A^i$ are strictly positive, which clearly is sufficient. In this manner, we generate a $1000$ graphs and labels.

**Model** The model set up is as follows: Before the first layer, the inputs are 'normalized' by subtracting .5 from each entry. Then, the first layer is chosen to map from the input space $\mathbb{R}^N \otimes \mathbb{R}^N$ into another space of adjacency matrices matrices $(\mathbb{R}^N \otimes \mathbb{R}^N)^{32}$, (on which $S_N$ is acting according to $\rho^{\text{perm}}$, defined in Example 3, on each component). Then, a group pooling layer is used, that is, a map into $\mathbb{R}^{64}$ on which $S_N$ is acting trivially. This is followed by a layer normalization layer, and then two layers $\mathbb{R}^{64} \to \mathbb{R}^{32}$ and $\mathbb{R}^{32} \to \mathbb{R}$. The group acts trivially on each of the final spaces. In other words, all models are equipped a fully connected head. All but the last non-linearities are chosen as leaky ReLUs, whereas the last one is a sigmoid function. Note that the non-linearities are applied pointwise in the early layers, so that they are trivially equivariant to the group action. We use a binary loss. A graphical depiction of the architecture is given in Figure 2.

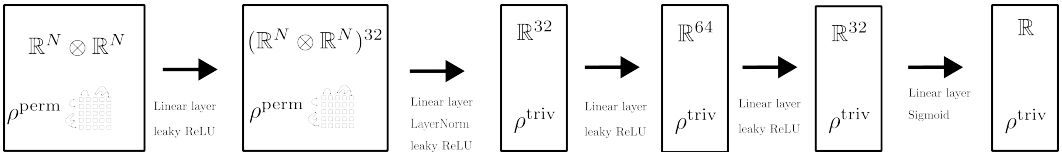


Figure 2: The setup for the models of PERM.

**Projection operators** To project the layers onto $\mathcal{E}$, we use the results of Maron et al. (2019a). In said paper, bases for the space invariant under the (lifted) representation $\rho^{\text{perm}}$ on $(\mathbb{R}^N)^{\otimes k}$ are given. In this context, the first layer consists of a $(32 \times 1)$ array of elements in $(\mathbb{R}^N \otimes \mathbb{R}^N) \otimes (\mathbb{R}^N \otimes \mathbb{R}^N) = (\mathbb{R}^N)^{\otimes 4}$, the second is an $(32 \times 32)$ array of elements in $\mathbb{R} \otimes (\mathbb{R}^N \otimes \mathbb{R}^N) = (\mathbb{R}^N)^{\otimes 2}$, and the final ones are simply arrays of real numbers (and in particular, $\mathcal{E}$ is the entire space).

## D.2 THE TRANS EXPERIMENT

In the TRANS experiment, we train a model for classification on MNIST. The input space here is $X_0 = \mathbb{R}^{N,N}$, where $N$ is the width/height of the image. On $X_0$, we let the translation group $\mathbb{Z}_N^2$ act as in Example 2. The classification task is invariant to this action, whence we again let $\mathbb{Z}_N^2$ act trivially on the output space $\mathbb{R}^{10}$ of probability distribution on the ten classes.

**Data** We use the public dataset MNIST Lecun et al. (1998) in our experiments, but modify it in two ways to keep the experiments light. First, we train our models only on the 10000 test examples

(instead of the 60000 training samples). Secondly, we subsample the $28 \times 28$ images to images of size $14 \times 14$ (and hence set $N = 14$) using `opencv`'s Bradski (2000) built-in RESIZE function. This simply to reduce the size of the networks. Note that the size of the early (non-equivariant) layers of the model are proportional to the $(\text{image width})^4$.

**Model** We again begin by 'normalizing' the output by subtracting $.5$ from each pixels. The actual architecture then begins with two initial layers mapping between spaces on which $\mathbb{Z}_N^2$ acts according to $\rho^{\text{tr}}$ (in each component) : Layer 1 maps from $\mathbb{R}^{N,N} \to (\mathbb{R}^{N,N})^{32}$ and layer 2 from $(\mathbb{R}^{N,N})^{32}$ to $(\mathbb{R}^{N,N})^{32}$. We then again apply a group pooling layer followed by a 'fully connected head': The third layer maps into $\mathbb{R}^{32}$, on which $\mathbb{Z}_N^2$ is acting trivially, and the final one between $\mathbb{R}^{32}$ and $\mathbb{R}^{10}$. After the pooling layer, we add a layer normalization layer. The non-linearities are again chosen as leaky ReLU:s, except for the last one, which is a SoftMax. The equivariance of the non-linearities are again lifted from them acting elementwise on the first three spaces. We use a cross-entropy loss. A graphical depiction of the setup is given in 3.

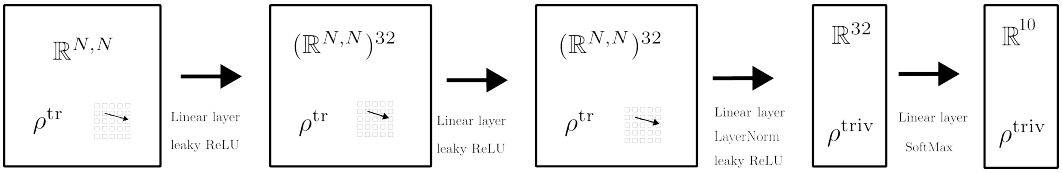


Figure 3: The setup for the models of TRANS.

**Projection operators** It is well-known that a linear operator on $\mathbb{R}^{N,N}$ is equivariant if and only if it is a convolutional operator. Hence, for the first two layers, the projection is carried out via projecting each component onto the space spanned by

$$C^{k\ell} = \sum_{i,j \in [N]} (e_i \otimes e_j) \otimes (e_{i+k} \otimes e_{i+\ell}), \quad k, \ell \in [N]^2. \tag{41}$$

The third layers consist of arrays of functionals on $\mathbb{R}^{N,N}$. The only linear invariant such is (up to scale and interpreted as a member of $\mathbb{R}^{N,N}$) the constant matrix. Hence, the projection in this space simply consists of averaging the layer. After that, just as above, $\mathcal{E}$ is the entire space, and the projection is trivial.

The convolution operators are furthermore well-known to be characterized as diagonal operators in the Fourier domain. We may therefore implement the projection onto $\mathcal{E}$ efficiently by first transforming into Fourier domain, extract the diagonal, and then transforming back. For completeness, let us quickly prove that this really yields the orthogonal projection.

**Proposition 4.** *For $N \in \mathbb{N}$, let the translation group $G = \mathbb{Z}_N^2$ act on the space $\mathbb{C}^{N,N} \sim \mathbb{C}^{[N]^2}$. If we let $\mathcal{F} : \mathbb{C}^{N,N} \to \mathbb{C}^{N,N}$ denote the orthogonal two-dimensional Fourier transform, and $\text{diag} : \mathbb{C}^{N,N} \to \mathbb{C}^{N,N}$ the operator that kills all off-diagonal values, the orthogonal projection onto $\text{Hom}_G(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$ is given by*

$$\Pi_{\mathcal{E}}(A) = \mathcal{F}^{-1}\text{diag}(\mathcal{F}A\mathcal{F}^{-1})\mathcal{F}. \tag{42}$$

*Proof.* Let us begin by showing that $P = \mathcal{F}^{-1}\text{diag}(\mathcal{F}A\mathcal{F}^{-1})\mathcal{F}$ defines an orthogonal projection. As for the 'projection' part, note that for any $A$, the matrix $D = \text{diag}(\mathcal{F}A\mathcal{F}^{-1})$ will of course be diagonal, so that $\text{diag}(D) = D$. Consequently,

$$P^2(A) = P(\mathcal{F}^{-1}D\mathcal{F}) = \mathcal{F}^{-1}\text{diag}(D)\mathcal{F} = \mathcal{F}^{-1}D\mathcal{F} = P(A). \tag{43}$$

The orthogonality can now be shown via arguing that $P$ is self-adjoint. This is furthermore not hard: Due to the orthogonality of $\mathcal{F}$, $A \mapsto \mathcal{F}A\mathcal{F}^{-1}$ and $A \mapsto \mathcal{F}^{-1}A\mathcal{F}$ are self-adjoint, as is obviously $\text{diag}$ (with respect to the standard scalar product on $\mathbb{C}^{N,N}$).

It is now only left to show that the range of $P$ is equal to the space $\text{Hom}_G(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$, which is the space of translations equivariant operators $\mathbb{C}^{N,N} \to \mathbb{C}^{N,N}$. It is however well-known that those exactly correspond to convolutional operators $C_\phi(v) = \phi * v$. The convolution theorem now states

$$\mathcal{F}(C_\phi v) = \mathcal{F}(\phi * v) = \mathcal{F}(\phi) \cdot \mathcal{F}(v) = \mathcal{D}(\mathcal{F}(\phi))\mathcal{F}v, \tag{44}$$

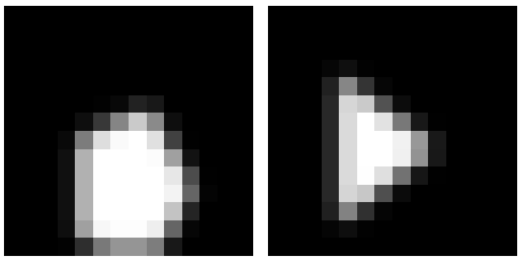

Figure 4: Two examples of generated images for ROT.

where $\mathcal{D} : \mathbb{C}^N \mapsto \mathbb{C}^{N,N}$ is the operator that inserts a vector into the diagonal of a matrix. In other words, $C_\phi = \mathcal{F}^{-1}\mathcal{D}(\mathcal{F}(\phi))\mathcal{F}$, which makes it obvious that $\mathrm{ran}P \subseteq \mathcal{E}$. Furthermore, by setting $A = \mathcal{F}^{-1}\mathcal{D}(\mathcal{F}(\phi))\mathcal{F}$, we see that we can write every convolution operator as an image under $P$. In conclusion, $P = \Pi_\mathcal{E}$.

□

### D.3 THE ROT EXPERIMENT

In the final experiment, we consider a simple, rotation invariant, segmentation task for synthetically generated data. The input space consist of images in $\mathbb{R}^{N,N}$, and the output space $(\mathbb{R}^{N,N})^2$ of pairs of segmentation masks. This task is hence 'truly' equivariant.

**Data** We generate $3000$ synthetic images as follows: With equal probability, we generate a regular pentagon or an equilateral triangle (with vertices on the unit circle), scale it using a uniformly randomly chosen factor in $[0.7, 0.8]$, and place it randomly on a canvas. The resulting image, along with two segmentation masks, each indicating where each shape is present in the image (and in particular zero everywhere if the particular object is not present at all) are converted to $14 \times 14$ pixels images. Importantly, all triangles, and pentagons, have the same orientation, so that the image distribution is not invariant under discrete rotations of $90°$. Examples of generated images are showcased in Figure 4.

**Model** In contrast to the previous examples, the action on the output space is not trivial. Therefore, the model set up is less convoluted : We use hidden spaces $(\mathbb{R}^{N,N})^{32}$,$(\mathbb{R}^{N,N})^{32}$ and $(\mathbb{R}^{N,N})^{16}$ on which the rotation group acts according to $\rho^{\mathrm{rot}}$ (on each component). Before the final non-linearity, we use a batch norm layer. All non-linearities are leaky ReLUs, except for the last one, which is a sigmoid non-linearity. A visualization is given in Figure 5. The loss function is

$$\begin{aligned}\ell((y_{\mathrm{pent}}, y_{\mathrm{tri}}), y'_{\mathrm{pent}}, y'_{\mathrm{tri}}) = \frac{1}{N^2}\sum_{k\in[N]^2} \mathrm{BCL}((y_{\mathrm{pent}}(k), y'_{\mathrm{pent}}(k))) \\ + \frac{1}{N^2}\sum_{k\in[N]^2} \mathrm{BCL}(y_{\mathrm{tri}}(k), y'_{\mathrm{tri}}(k)))\end{aligned} \tag{45}$$

where $\mathrm{BCL}$ is the binary cross-entropy loss, and $y_{\mathrm{pent}}, y_{\mathrm{tri}}$ are the pentagon and triangle segmentation masks, respectively. Note that the pixel-wise nature of the loss implies that it is invariant under rotations of the masks.

**Projection operators** Since the group in this case only consists of four elements, and the Haar measure is the uniform one, we can calculate the projection of the layers by explicitly carrying out the integration

$$\Pi_\mathcal{E} A = \tfrac{1}{4}\sum_{k\in\mathbb{Z}_4} \overline{\rho}(k)A. \tag{46}$$

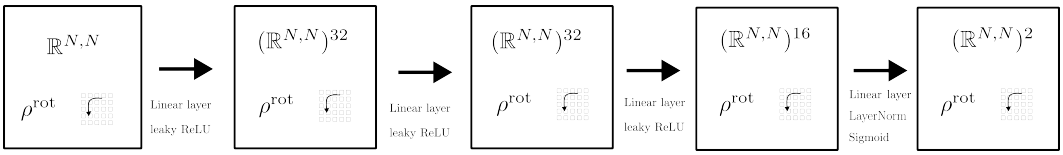


Figure 5: The setup for the models of Experiment 3

# E THE RELATION BETWEEN $\dim(\mathcal{E})$ AND THE 'EMPIRICAL STABILITY' OF $\mathcal{E}$.

In the main text, we sketched a heuristic argument relating the relative dimension of the space $\mathcal{E}$ in the space $\mathcal{L}$ with how close the AUG model stayed to $\mathcal{E}$. In this section, we first want to carry out these computations. To further investigate the link, we also provide an additional experiment. A quick glance of the calculations are given in Tables 1 and 2

## E.1 THE EXPERIMENTS IN SECTION 5

**PERM** As mentioned before, the spaces of equivariant linear maps was described in Maron et al. (2019a). Therein, it was shown that the dimension of (under the permutation group) equivariant linear maps from $\mathbb{R}^N \otimes \mathbb{R}^N$ to itself is equal to 15, and from $\mathbb{R}^N \otimes \mathbb{R}^N$ to itself is 2 (as long as $N \geq 2$, as was pointed out in Finzi et al. (2021)). The dimensions of the corresponding whole spaces are $N^4$ and $N^2$, respectively. In the architecture we are proposing, we use 32 maps $\mathbb{R}^N \otimes \mathbb{R}^N$ in the first layer, and $32 \cdot 64$ maps $\mathbb{R}^N \otimes \mathbb{R}^N \to \mathbb{R}$ in the second layer. In the layers thereafter, all $64 \cdot 32 + 32 \cdot 1$ maps are equivariant. Hence

$$\frac{\dim \mathcal{E}}{\dim \mathcal{L}} = \frac{15 \cdot 32 + 5 \cdot 32 \cdot 64 + 64 \cdot 32 + 32 \cdot 1}{N^4 \cdot 32 + N^2 \cdot 32 \cdot 64 + 64 \cdot 32 + 32 \cdot 1} \approx 2.43 \cdot 10^{-2}. \tag{47}$$

where we plugged in $N = 10$ in the final step.

**TRANS** Here, the space of equivariant linear maps $\mathbb{R}^{N,N} \to \mathbb{R}^{N,N}$ is the space of convolution operators. On $\mathbb{R}^{N,N}$, there are $N^2$ of those. There is up to scale only one equivariant map $\mathbb{R}^{N,N} \to \mathbb{R}$, given by taking the mean of the matrix entries. The space of all linear maps in the two cases of course have dimension $N^4$ and $N^2$, respectively. Since we use two initial layers of 32 and $32 \cdot 32$ maps $\mathbb{R}^{N,N} \to \mathbb{R}^{N,N}$, respectively, one layer of $32 \times 32$ maps $\mathbb{R}^{N,N} \to \mathbb{R}$ and $32 \cdot 10$ maps from $\mathbb{R} \to \mathbb{R}$ (where again all maps are equivariant), we obtain

$$\frac{\dim \mathcal{E}}{\dim \mathcal{L}} = \frac{N^2 \cdot 32 + N^2 \cdot 32 \cdot 32 + 1 \cdot 32 \cdot 32 + 32 \cdot 10}{N^4 \cdot 32 + N^4 \cdot 32 \cdot 32 + N^2 \cdot 32 \cdot 32 + 32 \cdot 10} \approx 5.1 \cdot 10^{-3}. \tag{48}$$

where we plugged in $N = 14$ in the final step.

**ROT** Here, a map from $\mathbb{R}^{N,N}$ to $\mathbb{R}^{N,N}$ is clearly equivariant if and only if its matrix representation, as a function from $([N]^2)^2$ to $R$, is constant along all orbits

$$\{(\iota_0, \iota_1), (r^{-1}(\iota_0), r(\iota_1)), (r^{-2}(\iota_0), r^2(\iota_1)), (r^{-3}(\iota_0), r^3(\iota_1))\}, \tag{49}$$

| Experiment | $U$ | $V$ | $\dim \mathrm{Hom}(U,V)$ | $\dim \mathrm{Hom}_G(U,V)$ | # in model |
|---|---|---|---|---|---|
| PERM | $\mathbb{R}^N \otimes \mathbb{R}^N$ | $\mathbb{R}^N \otimes \mathbb{R}^N$ | $N^4$ | 15 | 32 |
| | $\mathbb{R}^N \otimes \mathbb{R}^N$ | $\mathbb{R}$ | $N^2$ | 2 | $32 \cdot 64$ |
| | $\mathbb{R}$ | $\mathbb{R}$ | 1 | 1 | $64 \cdot 32 + 32 \cdot 1$ |
| TRANS | $\mathbb{R}^{N,N}$ | $\mathbb{R}^{N,N}$ | $N^4$ | $N^2$ | $32 + 32 \cdot 32$ |
| | $\mathbb{R}^{N,N}$ | $\mathbb{R}$ | $N^2$ | 1 | $32 \cdot 32$ |
| | $\mathbb{R}$ | $\mathbb{R}$ | 1 | 1 | $32 \cdot 10$ |
| ROT | $\mathbb{R}^{N,N}$ | $\mathbb{R}^{N,N}$ | $N^4$ | $\frac{1}{4} \cdot N^4$ | $32 + 32 \cdot 32 +$ $32 \cdot 16 + 16 \cdot 2$ |

Table 1: Dimension calculations for the experiments PERM , TRANS and ROT from the main paper.

where $\iota_i = [\iota_i(0), \iota_i(1)] \in [N]^2$ and

$$r(\iota) = (-\iota(1), \iota(0)) \tag{50}$$

Since $N = 14$, i.e. even, each such orbit has $4$ elements, and they are of course disjoint. Therefore, the relative dimension between the space of equivariant maps $\mathbb{R}^{N,N} \to \mathbb{R}^{N,N}$ and the space of all such maps is equal to $1/4$. Since it is only such maps that appear in the model used in the experiment,

$$\frac{\dim \mathcal{E}}{\dim \mathcal{L}} = \frac{1}{4} = 0.25 \tag{51}$$

in this case.

### E.2 AN ADDITIONAL EXPERIMENT

In the main paper, not only $G$ (and accordingly $\mathcal{E}$) was varying in between the experiment, but also the nominal model architectures and datasets, which makes it unclear if it is only $G$ that plays a role in the different speeds of drift of the augmented model from (or say 'empirical stability' of) $\mathcal{E}$ . We therefore perform an additional experiment in which all models use the same underlying architecture and dataset, namely the one of the TRANS experiment, and only vary the underlying symmetry group acting on $\mathbb{R}^{N,N}$. We use four groups and actions.

- TRANS $\mathbb{Z}_N^2$ acting through translations, as in the TRANS experiment.
- ROT $\mathbb{Z}_4$ acting through rotations, as in the ROT experiment.
- ONEDTRANS $\mathbb{Z}_N$ acting through translations in the $x$-direction, i.e.

$$(\rho^{\mathrm{tr}_0}(k)x)_{i,j} = x_{i-k,j} \tag{52}$$

- TRANSROT The semi-direct product $\mathbb{Z}_N^2 \rtimes \mathbb{Z}^4$ acting through

$$\rho^{\mathrm{trot}}(\iota, k)x = \rho^{\mathrm{rot}}(k)\rho^{\mathrm{tr}}(\iota)x, \quad \iota \in \mathbb{Z}_N^2, k \in \mathbb{Z}_4. \tag{53}$$

  This can, in the same way $\mathbb{Z}_4$ is a discretization of the full group $\mathrm{SO}(2)$ of rotations in the plane, be thought of as a discretization of the group $\mathrm{SE}(2)$ of isometries in the plane.

The relative dimensions $\dim \mathcal{E} / \dim \mathcal{L}$ are in these cases (approximately) given by

$$\begin{array}{llll} \text{TRANS}: & 5.1 \cdot 10^{-3} & \text{ROT}: & 0.25 \\ \text{ONEDTRANS}: & 7.1 \cdot 10^{-2} & \text{TRANSROT}: & 1.3 \cdot 10^{-3} \end{array} . \tag{54}$$

These numbers are the results of similar calculations as above, with the data from Table (2) – the (conceptually simple but technical) arguments for the validity of non-trivial entries are given in Section E.2.1. According to it, we expect the relative drift of the augmented model compared to the nominal one from $\mathcal{E}$ should be the largest for ROT, followed by ONEDTRANS, TRANS and TRANSROT, in decreasing order.

| Experiment | $U$ | $V$ | $\dim \mathrm{Hom}(U,V)$ | $\dim \mathrm{Hom}_G(U,V)$ | # in model |
|---|---|---|---|---|---|
| ROT | $\mathbb{R}^{N,N}$ | $\mathbb{R}^{N,N}$ | $N^4$ | $\frac{1}{4}N^4$ | $32 + 32 \cdot 32$ |
| | $\mathbb{R}^{N,N}$ | $\mathbb{R}$ | $N^2$ | $\frac{1}{4}N^2$ | $32 \cdot 32$ |
| | $\mathbb{R}$ | $\mathbb{R}$ | 1 | 1 | $32 \cdot 10$ |
| ONEDTRANS | $\mathbb{R}^{N,N}$ | $\mathbb{R}^{N,N}$ | $N^4$ | $N^3$ | $32 + 32 \cdot 32$ |
| | $\mathbb{R}^{N,N}$ | $\mathbb{R}$ | $N^2$ | $N$ | $32 \cdot 32$ |
| | $\mathbb{R}$ | $\mathbb{R}$ | 1 | 1 | $32 \cdot 10$ |
| TRANSROT | $\mathbb{R}^{N,N}$ | $\mathbb{R}^{N,N}$ | $N^4$ | $\frac{1}{4}N^2$ | $32 + 32 \cdot 32$ |
| | $\mathbb{R}^{N,N}$ | $\mathbb{R}$ | $N^2$ | 1 | $32 + 32 \cdot 32$ |
| | $\mathbb{R}$ | $\mathbb{R}$ | 1 | 1 | $32 \cdot 10$ |

Table 2: Dimension calculations for the experiments ROT , ONEDTRANS and TRANSROT in the appendix.

As mentioned, we repeat the exact same experiment as TRANS for the new groups[2]. We then plot the results similarly as in the experiments in the main paper, in Figure 6. Note that the scales in the figures are different, to assess the amount the augmented model drifts *relative* to the other ones. We have chosen the limits for the axes as follows:

- The $x_{\text{left}}$-limit in both subplots is chosen as $1.5$ times the maximal (with respect to the $50$ training epochs) median (with respect to the $30$ runs) value of $\|A - A_0\|$ for the EQUI model.
- The $y_{\text{left}}$-limit in the left subplot is chosen as $1.5$ times the maximal (with respect to the $50$ training epochs) median (with respect to the $30$ runs) value of $\|\Pi_{\mathcal{E}^\perp} A\|$ for the NOM model.
- The $y_{\text{right}}$-limit in the left subplot is given by $\lambda \cdot y_{\text{left}}$, where $\lambda > 0$ is a factor common for all four groups. $\lambda$ is chosen so that $y_{\text{right}}$ is equal to $1.5$ times the maximal (with respect to the $50$ training epochs) median (with respect to the $30$ runs) value of $\|\Pi_{\mathcal{E}^\perp} A\|$ for the AUG model *for the* TRANS *experiment*.

In this way, we ensure that the coordinate systems are on the same scale relative to the NOM and EQUI models. It is hence not surprising that the NOM curves look the same in all the plots – it is very much that way by design. The same is however not true for the AUG curve, and that is telling us something – in fact, this behaviour is in accordance with our hypothesis. This further strengthens the argument that the dimension of the subspace $\mathcal{E}$ plays a crucial role in the amount of regularizing effect augmentation has. Judging by the quite small difference between TRANS and TRANSROT , although the relative dimensions differ by a factor $4$, further suggests that hypothesizing a simple linear relationship is to naive – more work needs to be done here.

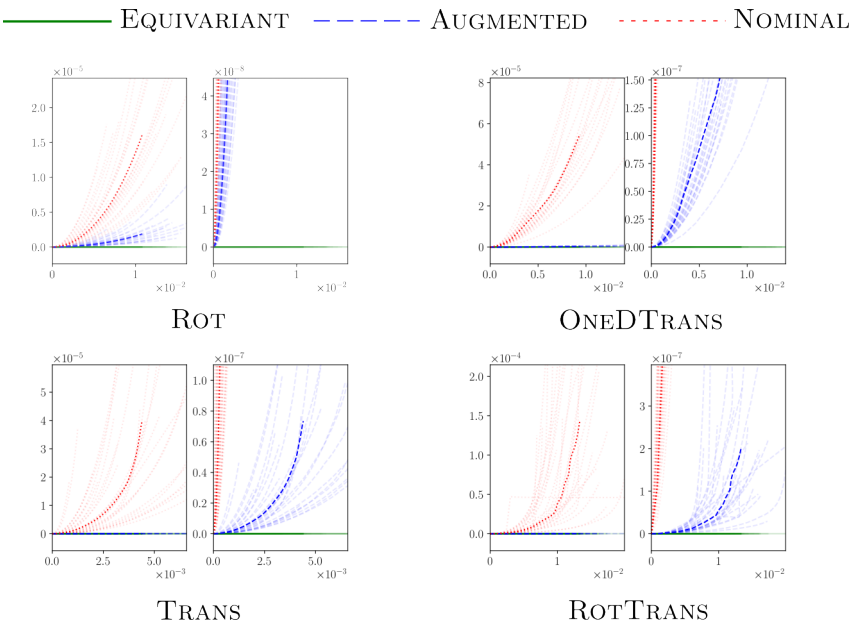


Figure 6: Results from the experiments of the appendix.

### E.2.1 THE SPACES $\text{Hom}_G(U, V)$ AND PROJECTION OPERATORS

Here, we derive what $\text{Hom}_G(U, V)$ look like for the groups $\mathbb{Z}_N$ and $\mathbb{Z}^4 \rtimes \mathbb{Z}_N$, and the corresponding projection operators. These derivations are all conceptually easy, but technical, and are only included for completeness.

**$\mathbb{Z}_N$ and ONEDTRANS** To describe the space $\text{Hom}_{\mathbb{Z}_N}(\mathbb{R}^{N,N}, \mathbb{R}^{N,N})$, let us begin by introducing some notation. First, every element in $\mathbb{R}^{N,N}$ can equivalently be described as a collection of rows.

[2]Resulting in about 60 more hours of GPU time, on Tesla A40 GPUs situated on a cluster

This can be written compactly as

$$X = \sum_{k\in[N]} e_k x_k^* \tag{55}$$

where $x_k \in \mathbb{R}^N$ are the rows, and $e_k$ is the $k$:th canonical unit vector. Correspondingly, each operator $A : \mathbb{R}^{N,N} \to \mathbb{R}^{N,N}$ decomposes into an array of $N^2$ operators $A_{\ell,k} : \mathbb{R}^N \to \mathbb{R}^N$:

$$A(x) = \sum_{k,\ell\in[N]} e_\ell (A_{\ell,k} x_k)^*. \tag{56}$$

With this notation introduced, we can conveniently describe the space $\mathrm{Hom}_{\mathbb{Z}_N}(\mathbb{R}^{N,N}, \mathbb{R}^{N,N})$

**Proposition 5.** *Using the notation* (56)*,* $\mathrm{Hom}_{\mathbb{Z}_N}(\mathbb{R}^{N,N}, \mathbb{R}^{N,N})$ *is characterized as the set of operators for which each* $A_{\ell,k}$ *is a convolutional operator. In particular, the dimension of the space is* $N^2 \cdot N = N^3$.

*Proof.* Somewhat abusing notation, let us denote the action of $\mathbb{Z}_N$ on $\mathbb{R}^N$ also as $\rho^{\mathrm{tr}_0}$. We then have

$$\rho^{\mathrm{tr}_0}(n)(e_k x^*) = e_k \rho^{\mathrm{tr}_0}(n)(x)^* \tag{57}$$

Hence,

$$A(\rho^{tr_0}(n)x) = \sum_{k,\ell\in[N]} e_\ell (A_{\ell,k}\rho^{\mathrm{tr}_0}(n)x_k)^* \tag{58}$$

$$\rho^{\mathrm{tr}_0(n)}(Ax) = \sum_{k,\ell\in[N]} e_\ell (\rho^{\mathrm{tr}_0}(n)A_{\ell,k}x_k)^* \tag{59}$$

These two expression are equal for any $x$ if and only if $A_{\ell,k}\rho^{\mathrm{tr}_0}(n)x_k = \rho^{\mathrm{tr}_0}(n)A_{\ell,k}x_k$ for any $x_k$ and $\ell, k$, i.e., when every $A_{\ell,k}$ is translation equivariant, which is equivalent to each $A_{\ell,k}$ being a convolution operator. □

Not surprisingly, we can again calculate the projection onto the space of equivariant operators with the help of the Fourier transform. However, the Fourier transform should here only be applied partially. Let us make this explicit. Let $\mathcal{F} : \mathbb{C}^N \to \mathbb{C}^N$ be the one-dimensional orthogonal Fourier transform. We then define the row-wise partial Fourier transformation $\mathbb{F} : \mathbb{C}^{N,N} \to \mathbb{C}^{N,N}$ through

$$\mathbb{F}\left(\sum_{k\in[N]} e_k x_k^*\right) = \sum_{k\in\mathbb{N}} e_k (\mathcal{F}x_k)^*. \tag{60}$$

We can now show that $A \in \mathrm{Hom}_{\mathbb{Z}^N}(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$ is related to $\mathbb{F}A\mathbb{F}^{-1}$ being 'partially diagonal', in the following sense.

**Lemma 8.** *Let* $\mathcal{F} : \mathbb{C}^N \to \mathbb{C}^N$ *be the one-dimensional Fourier transform. Then, an operator* $A$ *is in* $\mathrm{Hom}_{\mathbb{Z}^N}(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$ *if and only if, in the notation 56, each* $\mathcal{F}A_{\ell,k}\mathcal{F}^{-1}$ *is diagonal.*

*Proof.* It is not hard to show that

$$(\mathbb{F}A\mathbb{F}^{-1})_{\ell,k} = \mathcal{F}A_{\ell,k}\mathcal{F}^{-1}. \tag{61}$$

Now, it is well known that $A_{\ell,k}$ is convolutional if and only if $\mathcal{F}A_{\ell,k}\mathcal{F}^{-1}$ is diagonal. The claim follows. □

By applying exactly the same argument as in Proposition 4, we may now derive

**Proposition 6.** *The orthogonal projection from* $\mathrm{Hom}(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$ *onto* $\mathrm{Hom}_{\mathbb{Z}^N}(\mathbb{C}^{N,N}, \mathbb{C}^{N,N})$ *is given by*

$$\mathbb{F}^{-1}\mathrm{diag}_0(\mathbb{F}A\mathbb{F}^{-1})\mathbb{F}, \tag{62}$$

*where* $\mathrm{diag}_0$ *is the operator that kills all off-diagonal elements in all operators* $A_{\ell,k}$ *in the decomposition 56.*

As before, it is not hard to realize what the space $\mathrm{Hom}_{\mathbb{Z}^N}(\mathbb{R}^{N,N}, 1)$ must be: Interpreted as a matrix $X = \sum_{k\in[N]} e_k x_k^*$, $X$ is in $\mathrm{Hom}_{\mathbb{Z}^N}(\mathbb{R}^{N,N}, 1)$ if and only if each $x_k$ is constant. Correspondingly, the projection is given by taking means along the $x$-direction, and the dimension of the space in particular is $N$.

**$\mathbb{Z}_4$ and ROT** Here, we have in fact already described the space $\mathrm{Hom}_{\mathbb{Z}_4}(\mathbb{R}^{N,N}, \mathbb{R}^{N,N})$ and the projection operator in Appendix D. Let us here just record that the exact same idea goes through for the space $\mathrm{Hom}_{\mathbb{Z}_4}(\mathbb{R}^{N,N}, 1)$ – the projection can again be calculated via explicit integration, and its elements are again characterized by having constant values on orbits of length $4$. Correspondingly, $\dim \mathrm{Hom}_{\mathbb{Z}_4}(\mathbb{R}^{N,N}, 1) = \frac{1}{4}N^2$

**TRANSROT and $\mathbb{Z}_N^2 \rtimes \mathbb{Z}_4$** Here, the semi-direct product structure immediately implies a relationship between the projection operators. First, let us notice that the lift $\overline{\rho}^{\mathrm{trot}}$ is given by $\overline{\rho}^{\mathrm{trot}}(\iota, k) = \overline{\rho}^{\mathrm{rot}}(k)\overline{\rho}^{\mathrm{tr}}(\iota)$. This fact together with the explicit integration formula

$$\Pi_{\mathcal{E}} = \int_{\mathbb{Z}_N^2 \rtimes \mathbb{Z}_4} \overline{\rho}^{\mathrm{trot}}(g)\mathrm{d}\mu(g) = \frac{1}{4N^2} \sum_{k \in Z_4, \iota \in \mathbb{Z}_N^2} \overline{\rho}^{\mathrm{rot}}(k)\overline{\rho}^{\mathrm{tr}}(\iota), \tag{63}$$

where we explicitly used that the Haar measure is the uniform one, implies the following

**Proposition 7.** *For $G$, let $P_G$ denote the projection operators onto $\mathrm{Hom}_G(\mathbb{R}^{N,N}, V)$ where $V$ is either $\mathbb{R}$ or $\mathbb{R}^{N,N}$. Then,*

$$P_{\mathbb{Z}_N^2 \rtimes \mathbb{Z}^4} = P_{\mathbb{Z}_4} P_{\mathbb{Z}_N^2}. \tag{64}$$

The above in particular means that the space $\mathcal{E}$ is the intersection between the space $\mathcal{E}$ for the respective groups. Specifically, this means that the space $\mathrm{Hom}_{\mathbb{Z}_N \rtimes \mathbb{Z}_4}(\mathbb{R}^{N,N}, \mathbb{R}^{N,N})$ is the space of convolution operators whose convolutional filter is constant on orbits under $\mathbb{Z}_4$, and therefore (in the even case) has dimension $\frac{N^2}{4}$. For $\mathrm{Hom}_{\mathbb{Z}_N \rtimes \mathbb{Z}_4}(\mathbb{R}^{N,N}, \mathbb{R})$, the space is still one-dimensional – it is given by the convolution with the constant filter (which in particular is constant on orbits)..¨

## F INCLUDING BIAS TERMS IN THE FRAMEWORK

Let us here show how bias terms can be incorporated into our framework by applying the standard trick of writing an affine map as a linear one on an extended space. Indeed, given an affine map $D : U \to V, x \mapsto Ax + b$ for some $A \in \mathrm{Hom}(U, V)$ and $b \in V$, we may define a linear map $D^\circ : U^\circ \to V^\circ$, where for a vector space $X$, we wrote $X^\circ = X \oplus \mathbb{R}$, through

$$D^\circ(x, \lambda) = (Ax + \lambda \cdot b, \lambda). \tag{65}$$

Note that $D^\circ(x, 1) = (D(x), 1)$. Hence, $D^\circ$ restricted to $U \times \{1\}$ acts exactly as $D$. These maps form a linear space that we call $\mathrm{Hom}^\circ(U^\circ, V^\circ)$. We can further extend a representation $\rho$ of $G$ on $U$ to one on $U^\circ$ via $\rho^\circ(g)(u, \lambda) = (\rho(g)u, \lambda)$. If $\rho$ is unitary, $\rho^\circ$ also is.

It is not hard to show that the affine map $D : X_i \to X_{i+1}$ is equivariant with respect to $\rho_i, \rho_{i+1}$ if and only if the linear map $D^\circ$ is equivariant with respect to $\rho_i^\circ, \rho_{i+1}^\circ$. In particular, the lifted representation $\overline{\rho}_i^\circ$ maps $\mathrm{Hom}^\circ(X^i, X^{i+1})$ onto itself, and $D^\circ$ is equivariant if and only if $\overline{\rho}(g)D^\circ = D^\circ$ for all $g$. Finally, the non-linearities $\sigma_i$ and loss $\ell$ can be extended to $\sigma_i^\circ : X_i^\circ \to X_{i+1}^\circ, (x, \lambda) \to (\sigma(x), \lambda)$, and $\ell : Y^\circ \times Y^\circ \to R$, $\ell^\circ((x, \lambda), (y, \kappa)) = \ell(x, y)$. Equivariance of $\sigma_i$ and $\ell$ is equivalent to equivariance of $\sigma^\circ$ and $\ell^\circ$, respectively.

Hence, in short, by extending all hidden spaces $X_i$ to $X_i^\circ$, the non-linearities, losses and representations accordingly, and restricting the $A_i$ to live in $\mathrm{Hom}^\circ(X_i^\circ, X_{i+1}^\circ)$, we obtain a framework that for all intents and purposes works exactly like the one we study in this paper, but also allows for the inclusion of bias terms. All arguments in the paper carry over verbatim in this extended framework.