A Proof of Proposition 4.1

A.1 A Primer on Representation Theory

Let $\mathrm{GL}(V)$ denote the general linear group of invertible linear transformations on the vector space V. Contrarily to the notations defined in Section 2, the vector of the vector of N ones and the all-ones $N \times M$ matrix are denoted by $\mathbf{1}_N$ and $\mathbf{1}_{N,M}$, respectively.

A group homomorphism. Let G and H be groups. A *group homomorphism* is a function $\phi : G \to H$ that preserves the group operation, meaning that for all $g_1, g_2 \in G$, $\phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2)$. Additionally, a homomorphism satisfies: For all $g \in G$, $\phi(g^{-1}) = (\phi(g))^{-1}$ and $\phi(e_G) = e_H$, where e_G and e_H are the identity elements of G and H, respectively.

A representation. Let G be a group and let V be a vector space over a field $\mathbb{F}$. Let $\rho : G \to \mathrm{GL}(V)$ be a group homomorphism. Then a *representation* of G is the pair (V, ρ).

An invariant subspace. Let G be a group. Let (V, ρ) be a representation. A subspace $W \subseteq V$ is called *an invariant subspace for* ρ if for every $g \in G$ and $w \in W$, $\rho(g)w \in W$ holds.

A subrepresentation. Let G be a group. Let (V, ρ) be a representation. Let $W \subseteq V$ be an invariant subspace for ρ. Then, (W, ρ_W) is a subrepresentation of (V, ρ), where the restriction of ρ to W defines the representation $\rho_W : G \to \mathrm{GL}(W)$. **Irreducible and completely reducible representations.** If V has no nontrivial subrepresentations (i.e., the only invariant subspaces are $\{0\}$ and V), then (V, ρ) is called **irreducible**. The representation (V, ρ) is said to be **completely reducible** if $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$, where each (V_i, ρ_{V_i}) is an irreducible representations. For brevity, we refer to a representation (V_i, ρ_{V_i}) by (V_i, ρ).

G-equivariance over representations. Let G be a group. Let (V, ρ_V) and (W, ρ_W) be representations of G. We say that a function $f : V \to W$ is *G-equivariant with respect to ρ_V and ρ_W* if

$$\forall g \in G, v \in V \quad f(\rho_V(g)v) = \rho_W(g)f(v).$$

We can now detail a classical result in representation theory that is later used to derive the irreducible representations that are $(S_F \times S_C)$-equivariant.

Lemma A.1. *The vector space $\mathbb{R}^N$ can be decomposed the S_N-invariant subspaces $V_1^N \oplus V_0^N$ such that*

$$\mathbb{R}^N \cong V_1^N \oplus V_0^N,$$

where the one-dimensional $V_1^N = \{c \cdot \mathbf{1}_N \mid c \in \mathbb{R}\}$ is denoted as the trivial irreducible representation *and the $(N-1)$-dimensional $V_0^N = \{\boldsymbol{x} \in \mathbb{R}^N \mid \mathbf{1}_N^\top \boldsymbol{x} = 0\}$ is denoted as* the standard irreducible representation*, which is the orthogonal complement to V_1^N.*

After we obtain all the irreducible representations we employ Schur's lemma to derive all linear layers that are $(S_F \times S_C)$-equivariant.

Lemma A.2 (Schur's lemma)**.** *Let G be a group. Let (V, ρ_V) and (W, ρ_W) be finite-dimensional irreducible representations of G. Suppose a linear map $L : V \to W$ is G-equivariant with respect to ρ_V and ρ_W. Then,*

1. *If V and W are non-isomorphic, then $L = 0$.*
2. *If $V = W$, then T is a scalar multiple of the identity, i.e., there exists $\lambda \in \mathbb{N}$ such that $L = \lambda I_V$.*

A.2 Proof of Proposition 4.1

In this section we prove an extension of Proposition 4.1.

Proposition 4.1. *A linear function of the form*

$$T = (T_1, T_2)\colon \mathbb{R}^{K_1\times N\times F} \times \mathbb{R}^{K_1\times N\times C} \to \mathbb{R}^{K_2\times N\times F} \times \mathbb{R}^{K_2\times N\times C}$$

is $(S_N \times S_F \times S_C)$-equivariant if and only if there exist $\mathbf{\Lambda}^{(1)}, \dots, \mathbf{\Lambda}^{(12)} \in \mathbb{R}^{K_2\times K_1}$ such that for every $\boldsymbol{X} \in \mathbb{R}^{K_1\times N\times F}$, $\boldsymbol{Y} \in \mathbb{R}^{K_1\times N\times C}$ and $k_2 \in [K_2]$, we have

$$\forall i \in [6], \quad \boldsymbol{X}^{(i)}_{k_2} = \sum_{k_1=0}^{K_1} \Lambda^{(i)}_{k_2,k_1}\boldsymbol{X}_{k_1}, \quad \boldsymbol{Y}^{(i)}_{k_2} = \sum_{k_1=0}^{K_1} \Lambda^{(i+6)}_{k_2,k_1}\boldsymbol{Y}_{k_1}$$

$$T_1(\boldsymbol{X},\boldsymbol{Y})_{k_2} = \left(\mathbf{1}_{N,N}\boldsymbol{X}^{(1)}_{k_2} + \boldsymbol{X}^{(2)}_{k_2}\right)\mathbf{1}_{F,F} + \mathbf{1}_{N,N}\boldsymbol{X}^{(3)}_{k_2} + \boldsymbol{X}^{(4)}_{k_2} + \left(\mathbf{1}_{N,N}\boldsymbol{Y}^{(5)}_{k_2} + \boldsymbol{X}^{(6)}_{k_2}\right)\mathbf{1}_{C,F},$$

$$T_2(\boldsymbol{X},\boldsymbol{Y})_{k_2} = \left(\mathbf{1}_{N,N}\boldsymbol{Y}^{(1)}_{k_2} + \boldsymbol{Y}^{(2)}_{k_2}\right)\mathbf{1}_{C,C} + \mathbf{1}_{N,N}\boldsymbol{Y}^{(3)}_{k_2} + \boldsymbol{Y}^{(4)}_{k_2} + \left(\mathbf{1}_{N,N}\boldsymbol{X}^{(5)}_{k_2} + \boldsymbol{X}^{(6)}_{k_2}\right)\mathbf{1}_{F,C},$$

where $T_1(\boldsymbol{X},\boldsymbol{Y}) \in \mathbb{R}^{K_2\times N\times F}$ and $T_2(\boldsymbol{X},\boldsymbol{Y}) \in \mathbb{R}^{K_2\times N\times C}$.

Proof. **The irreducibles of $\mathbb{R}^{K\times N\times (F+C)}$.** We view $\mathbb{R}^{F+C}$ as the direct product of three coordinate subspaces

$$\mathbb{R}^{K\times N\times (F+C)} = \mathbb{R}^K \otimes \mathbb{R}^N \otimes \mathbb{R}^{F+C} \tag{4}$$

The group $S_N \times (S_F \times S_C)$ acts naturally on this space, where S_N acts by permuting the coordinates of $\mathbb{R}^N$, and $S_F \times S_C$ acts on $\mathbb{R}^{F+C}$.

By Lemma A.1, the vector space $\mathbb{R}^N$ is composed of the irreducibles V_1^N and V_0^N, i.e.

$$\mathbb{R}^N = V_1^N \oplus V_0^N \tag{5}$$

Since no symmetry group acts on the channel axis $\mathbb{R}^K$, the corresponding representation is trivial. Consequently, $\mathbb{R}^K$ decomposes as the direct sum of K one-dimensional trivial representations

$$\mathbb{R}^K = \bigoplus_{k=1}^{K} E^k, \tag{6}$$

where each $E^k = \{c \cdot e_k \mid c \in \mathbb{R}\} \subset \mathbb{R}^K$ is a 1-dimensional subspace invariant under the identity action, and $e_k{}_{k=1}^{K}$ denotes the standard basis of $\mathbb{R}^K$.

The irreducibles of $\mathbb{R}^{F+C}$. We view $\mathbb{R}^{F+C}$ as the direct sum of two coordinate subspaces: $\mathbb{R}^F \oplus \mathbb{R}^C$. The group $S_F \times S_C$ acts naturally on this space, where S_F acts by permuting the coordinates of $\mathbb{R}^F$, and S_C acts similarly on $\mathbb{R}^C$.

By Lemma A.1, the vector space $\mathbb{R}^F$ is composed of the irreducibles V_1^F and V_0^F. Similarly, the vector space $\mathbb{R}^C$ is composed of the irreducibles V_1^C and V_0^C. Thus,

$$\mathbb{R}^{F+C} = V_1^F \oplus V_0^F \oplus V_1^C \oplus V_0^C. \tag{7}$$

Denote, for all $k \in [K]$ the following irreducible representations

$$\begin{aligned}
V_{1,1}^{N,F,k} &= V_1^N \oplus V_1^F \oplus E^k, & V_{1,0}^{N,F,k} &= V_1^N \oplus V_0^F \oplus E^k,\\
V_{1,1}^{N,C,k} &= V_1^N \oplus V_1^C \oplus E^k, & V_{1,0}^{N,C,k} &= V_1^N \oplus V_0^C \oplus E^k,\\
V_{0,1}^{N,F,k} &= V_0^N \oplus V_1^F \oplus E^k, & V_{0,0}^{N,F,k} &= V_0^N \oplus V_0^F \oplus E^k,\\
V_{0,1}^{N,C,k} &= V_0^N \oplus V_1^C \oplus E^k, & V_{0,0}^{N,C,k} &= V_0^N \oplus V_0^C \oplus E^k,
\end{aligned}$$

with dimensions

$$\begin{aligned}
\dim\left(V_{1,1}^{(N,F,k)}\right) &= 1, & \dim\left(V_{1,0}^{(N,F,k)}\right) &= F-1,\\
\dim\left(V_{1,1}^{(N,C,k)}\right) &= 1, & \dim\left(V_{1,0}^{(N,C,k)}\right) &= C-1,\\
\dim\left(V_{0,1}^{(N,F,k)}\right) &= N-1, & \dim\left(V_{0,0}^{(N,F,k)}\right) &= (N-1)(F-1),\\
\dim\left(V_{0,1}^{(N,C,k)}\right) &= N-1, & \dim\left(V_{0,0}^{(N,C,k)}\right) &= (N-1)(C-1).
\end{aligned}$$

By substituting Equations (5) to (7) into Equation (4), we get

$$\mathbb{R}^{K\times N\times(F+C)} = \left(V_1^N \oplus V_0^N\right) \otimes \left(V_1^F \oplus V_0^F \oplus V_1^C \oplus V_0^C\right) \bigoplus_{k=1}^{K} E^k$$

$$= \bigoplus_{k=1}^{K} \left(V_{1,1}^{N,F,k} \oplus V_{1,0}^{N,F,k} \oplus V_{1,1}^{N,F,k} \oplus V_{1,0}^{N,F,k} \oplus V_{1,1}^{N,C,k} \oplus V_{1,0}^{N,C,k} \oplus V_{1,1}^{N,C,k} \oplus V_{1,0}^{N,C,k}\right).$$

The equivariant linear maps from $\mathbb{R}^{K_1\times N\times(F+C)}$ to $\mathbb{R}^{K_2\times N\times(F+C)}$. We denote the irreducible components of $\mathbb{R}^{K_1\times N\times(F+C)}$ by

$$V_{1,1}^{N,F,k_1}, V_{1,0}^{N,F,k_1}, V_{1,1}^{N,F,k_1}, V_{1,0}^{N,F,k_1}, V_{1,1}^{N,C,k_1}, V_{1,0}^{N,C,k_1}, V_{1,1}^{N,C,k_1}, V_{1,0}^{N,C,k_1}$$

for each $k_1 \in [K_1]$ and similarly for $\mathbb{R}^{K_2\times N\times(F+C)}$.

By Schur's lemma Lemma A.2, any equivariant linear map between non-isomorphic irreducible representations must be the zero map. Therefore, we restrict our attention to the linear maps between isomorphic irreducibles for each $k_1 \in [K_1], k_2 \in [K_2]$

$$\begin{aligned}
&L_1^{k_1,k_2} : V_{1,1}^{N,F,k_1} \to V_{1,1}^{N,F,k_2}, && L_2^{k_1,k_2} : V_{0,1}^{N,F,k_1} \to V_{0,1}^{N,F,k_2},\\
&L_3^{k_1,k_2} : V_{1,0}^{N,F,k_1} \to V_{0,0}^{N,F,k_2}, && L_4^{k_1,k_2} : V_{0,0}^{N,F,k_1} \to V_{0,0}^{N,F,k_2},\\
&L_5^{k_1,k_2} : V_{1,1}^{N,F,k_1} \to V_{1,1}^{N,C,k_2}, && L_6^{k_1,k_2} : V_{0,1}^{N,F,k_1} \to V_{0,1}^{N,C,k_2},\\
&L_7^{k_1,k_2} : V_{1,1}^{N,C,k_1} \to V_{1,1}^{N,C,k_2}, && L_8^{k_1,k_2} : V_{0,1}^{N,C,k_1} \to V_{0,1}^{N,C,k_2},\\
&L_9^{k_1,k_2} : V_{1,0}^{N,C,k_1} \to V_{0,0}^{N,C,k_2}, && L_{10}^{k_1,k_2} : V_{0,0}^{N,C,k_1} \to V_{0,0}^{N,C,k_2},\\
&L_{11}^{k_1,k_2} : V_{1,1}^{N,C,k_1} \to V_{1,1}^{N,F,k_2}, && L_{12}^{k_1,k_2} : V_{0,1}^{N,C,k_1} \to V_{0,1}^{N,F,k_2},
\end{aligned}$$

Following the respective definitions of the irreducibles these linear layers take the forms for every $\boldsymbol{X} \in \mathbb{R}^{K_1\times N\times F}, \boldsymbol{Y} \in \mathbb{R}^{K_1\times N\times C}$:

$$\begin{aligned}
&L_1^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_1^{k_1,k_2}\mathbf{1}_{N,N}\boldsymbol{X}\mathbf{1}_{F,F}, && L_2^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_2^{k_1,k_2}\left(\boldsymbol{X} - \mathbf{1}_{N,N}\boldsymbol{X}\right)\mathbf{1}_{F,F},\\
&L_3^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_3^{k_1,k_2}\mathbf{1}_{N,N}\left(\boldsymbol{X} - \boldsymbol{X}\mathbf{1}_{F,F}\right), && L_4^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_4^{k_1,k_2}\left(\boldsymbol{X} - \mathbf{1}_{N,N}\boldsymbol{X}\mathbf{1}_{F,F}\right),\\
&L_5^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_5^{k_1,k_2}\mathbf{1}_{N,N}\boldsymbol{X}\mathbf{1}_{F,C}, && L_6^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_6^{k_1,k_2}\left(\boldsymbol{Y} - \mathbf{1}_{N,N}\boldsymbol{X}\right)\mathbf{1}_{F,C},\\
&L_7^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_7^{k_1,k_2}\mathbf{1}_{N,N}\boldsymbol{Y}\mathbf{1}_{C,C}, && L_8^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_8^{k_1,k_2}\left(\boldsymbol{Y} - \mathbf{1}_{N,N}\boldsymbol{Y}\right)\mathbf{1}_{C,C},\\
&L_9^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_9^{k_1,k_2}\mathbf{1}_{N,N}\left(\boldsymbol{Y} - \boldsymbol{Y}\mathbf{1}_{C,C}\right), && L_{10}^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_{10}^{k_1,k_2}\left(\boldsymbol{Y} - \mathbf{1}_{N,N}\boldsymbol{Y}\mathbf{1}_{C,C}\right),\\
&L_{11}^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_{11}^{k_1,k_2}\mathbf{1}_{N,N}\boldsymbol{Y}\mathbf{1}_{C,F}, && L_{12}^{k_1,k_2}(\boldsymbol{X},\boldsymbol{Y}) = \lambda_{12}^{k_1,k_2}\left(\boldsymbol{Y} - \mathbf{1}_{N,N}\boldsymbol{Y}\right)\mathbf{1}_{C,F},
\end{aligned}$$

Thus, the general form can be written as:

$$\begin{aligned}
&\forall i \in [6], \quad \boldsymbol{X}_{k_2}^{(i)} = \sum_{k_1=0}^{K_1} \lambda_{k_2,k_1}^{(i)}\boldsymbol{X}_{k_1}, \quad \boldsymbol{Y}_{k_2}^{(i)} = \sum_{k_1=0}^{K_1} \lambda_{k_2,k_1}^{(i+6)}\boldsymbol{Y}_{k_1}\\
&T_1(\boldsymbol{X},\boldsymbol{Y})_{k_2} = \left(\mathbf{1}_{N,N}\boldsymbol{X}_{k_2}^{(1)} + \boldsymbol{X}_{k_2}^{(2)}\right)\mathbf{1}_{F,F} + \mathbf{1}_{N,N}\boldsymbol{X}_{k_2}^{(3)} + \boldsymbol{X}_{k_2}^{(4)} + \left(\mathbf{1}_{N,N}\boldsymbol{Y}_{k_2}^{(5)} + \boldsymbol{X}_{k_2}^{(6)}\right)\mathbf{1}_{C,F},\\
&T_2(\boldsymbol{X},\boldsymbol{Y})_{k_2} = \left(\mathbf{1}_{N,N}\boldsymbol{Y}_{k_2}^{(1)} + \boldsymbol{Y}_{k_2}^{(2)}\right)\mathbf{1}_{C,C} + \mathbf{1}_{N,N}\boldsymbol{Y}_{k_2}^{(3)} + \boldsymbol{Y}_{k_2}^{(4)} + \left(\mathbf{1}_{N,N}\boldsymbol{X}_{k_2}^{(5)} + \boldsymbol{X}_{k_2}^{(6)}\right)\mathbf{1}_{F,C},
\end{aligned}$$

where $T = (T_1, T_2)$ are $(S_N \times S_F \times S_C)$-equivariant linear function of the form

$$T = (T_1, T_2)\colon \mathbb{R}^{K_1\times N\times F} \times \mathbb{R}^{K_1\times N\times C} \to \mathbb{R}^{K_2\times N\times F} \times \mathbb{R}^{K_2\times N\times C}.$$

□

B Proof of Theorem 4.2

B.1 Background and Definitions

Contrarily to the notations defined in Section 2, we denote $\mathbf{1}_N$ as the vector of N ones.

Polynomials. Let $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$. We denote the first F columns by $\boldsymbol{X}^{(F)} \in \mathbb{R}^{N\times F}$ and the last C columns by $\boldsymbol{X}^{(C)} \in \mathbb{R}^{N\times C}$, such that $\boldsymbol{X} = (\boldsymbol{X}^{(F)}, \boldsymbol{X}^{(C)})$. We also denote that all permutation subgroup act on vectors by permuting their indices.

Let $\boldsymbol{x} \in \mathbb{R}^D$ be a vector. For a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^D$, define the monomial $\boldsymbol{x}^{\boldsymbol{\alpha}} = \boldsymbol{x}_1^{\alpha_1} \cdots \boldsymbol{x}_D^{\alpha_D}$ and its total degree $\|\boldsymbol{\alpha}\|_1$. Let $\{\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(T)}\}$ enumerate all such multi-indices $\boldsymbol{\alpha} \in \mathbb{N}_0^D$ with $\|\boldsymbol{\alpha}\|_1 \leq N$, where $T = \binom{N+D}{D}$.

Let $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ and let $p : \mathbb{R}^{N\times D} \to \mathbb{R}^{L\times K}$. For any $l \in [L]$, we define $p_l : \mathbb{R}^{N\times D} \to \mathbb{R}^K$ to be the mapping that extracts the l-th row of $p(\boldsymbol{X})$. Explicitly,

$$p_l(\boldsymbol{X}) = (p(\boldsymbol{X})_{l,1}, p(\boldsymbol{X})_{l,2}, \ldots, p(\boldsymbol{X})_{l,K}) \in \mathbb{R}^K,$$

that is, the K-dimensional feature vector corresponding to the l-th row of $p(\boldsymbol{X})$. Similarly, for all $l \in [L], k \in [K]$ we define $p_{l,k} : \mathbb{R}^{N\times D} \to \mathbb{R}$ to be the mapping that extracts the (l,k)-th element of $p(\boldsymbol{X})$.

Groups. We define an action of $S_{N-1} \times S_{F-1} \times S_C$ on $\mathbb{R}^{N\times(F+C)}$ by restricting the permutation $\sigma_{F-1} \in S_{F-1}$ to act only on the second to the F-th columns, and restricting the permutation $\sigma_{N-1} \in S_{N-1}$ to act only on last $(N-1)$-th rows, keeping element $(1,1)$ fixed. Formally,

$$((\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{X})_{n,j} = \begin{cases} \boldsymbol{X}_{1,1} & \text{if } n = 1 \text{ and } j = 1, \\ \boldsymbol{X}_{1,1+\sigma_{F-1}^{-1}(j-1)} & \text{if } n = 1 \text{ and } j \in [F] \setminus [1], \\ \boldsymbol{X}_{1,F+\sigma_C^{-1}(j-F)} & \text{if } n = 1 \text{ and } j \in [F+C] \setminus [F], \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),1} & \text{if } n \in [N] \setminus [1] \text{ and } j = 1, \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),1+\sigma_{F-1}^{-1}(j-1)} & \text{if } n \in [N] \setminus [1] \text{ and } j \in [F] \setminus [1], \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),F+\sigma_C^{-1}(j-F)} & \text{otherwise}, \end{cases}$$

for all $(\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \in S_{N-1} \times S_{F-1} \times S_C$ and $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$. In words, σ_{F-1} permutes the last $F-1$ columns, σ_C acts as before on the last C, and σ_N acts on the last $N-1$ rows, leaving element $(1,1)$ unchanged.

Similarly, the action of $S_{N-1} \times S_F \times S_{C-1}$ on $\mathbb{R}^{N\times(F+C)}$ is defined analogously: the permutation $\sigma_{C-1} \in S_{C-1}$ acts only on the last $C-1$ label columns and the permutation $\sigma_{N-1} \in S_{N-1}$ acts act only on last $N-1$-th rows, keeping element $(1, F+1)$ fixed. That is,

$$((\sigma_{N-1}, \sigma_F, \sigma_{C-1}) \cdot \boldsymbol{X})_{n,j} =$$

$$= \begin{cases} \boldsymbol{X}_{1,F+1} & \text{if } n = 1 \text{ and } j = F+1, \\ \boldsymbol{X}_{1,F+1+\sigma_{C-1}^{-1}(j-F-1)} & \text{if } n = 1 \text{ and } j \in [F+C] \setminus [F+1], \\ \boldsymbol{X}_{1,\sigma_F^{-1}(j)} & \text{if } n = 1 \text{ and } j \in [F], \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),F+1} & \text{if } n \in [N] \setminus [1] \text{ and } j = F+1, \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),F+1+\sigma_{C-1}^{-1}(j-F-1)} & \text{if } n \in [N] \setminus [1] \text{ and } j \in [F+C] \setminus [F+1], \\ \boldsymbol{X}_{1+\sigma_{N-1}^{-1}(n-1),\sigma_F^{-1}(j)} & \text{otherwise}, \end{cases}$$

for all $(\sigma_{N-1}, \sigma_F, \sigma_{C-1}) \in S_{N-1} \times S_F \times S_{C-1}$ and $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$.

B.2 Symmetric Polynomials

This subsection introduces key definitions and lemmas from symmetric polynomials and invariant theory that will be used in the proof of Theorem 4.2.

Generating set of a polynomial ring. Let $\mathbb{R}[\boldsymbol{X}]$ denote the polynomial ring over $\mathbb{R}^{N\times(F+C)}$. A finite set of polynomials $\{p_1,\dots,p_L\}\subset\mathbb{R}[\boldsymbol{X}]$ is called a *generating set* for $\mathbb{R}[\boldsymbol{X}]$ if every polynomial in $\mathbb{R}[\boldsymbol{X}]$ can be written as a polynomial in $p_1,\dots,p_L$ (with real coefficients); that is,

$$\mathbb{R}[\boldsymbol{X}]=\mathbb{R}[p_1,\dots,p_L].$$

We define the power-sum multi-symmetric polynomials for us to use a known result from invariant theory.

Definition B.1 (Power-sum Multi-symmetric Polynomials (PMPs))**.** *Let $M\in\mathbb{N}$. The* power-sum multi-symmetric polynomials $\{s_t\}_{t=1}^{T}:\mathbb{R}^{N\times D}\to\mathbb{R}$ *are defined by*

$$s_t(\boldsymbol{X})=\sum_{n=1}^{N}\boldsymbol{X}_{n,:}^{\boldsymbol{\alpha}^{(t)}}$$

and are parameterized by $\boldsymbol{\alpha}^{(t)}\in\mathbb{N}_0^D$ such that $\|\boldsymbol{\alpha}\|_1\le M$, where $T=\binom{M+D}{D}$[2]*.*

This known result from invariant theory is utilized in the proof of Lemmas B.3 and B.6.

Lemma B.2 (Corollary 8.4 in [Rydh, 2007])**.** *Let $\boldsymbol{X}\in\mathbb{R}^{N\times D}$. Let $p:\mathbb{R}^{N\times D}\to\mathbb{R}$ be an S_N-invariant polynomial with degree M. Then, it can be written as a polynomial in the PMPs*

$$p(\boldsymbol{X})=q(s_1(\boldsymbol{X}),\dots,s_T(\boldsymbol{X})),$$

for some polynomial $q:\mathbb{R}^T\to\mathbb{R}$, where $T=\binom{M+D}{D}$.

The following lemma enables the decomposition of polynomials that are not invariant under the full symmetry group S_N into a combination of polynomials that are, thereby allowing them to be approximated using our framework. The following lemma, originally stated as Lemma 2 in Segol and Lipman [2019], contained a minor error in the summation index bounds. We present the corrected version below.

Lemma B.3 (Lemma 2 in Segol and Lipman [2019])**.** *Let $p:\mathbb{R}^{N\times D}\to\mathbb{R}$ be an S_{N-1}-invariant polynomial with degree M. Then*

$$p(\boldsymbol{X})=\sum_{\|\boldsymbol{\alpha}\|_1\le M}\boldsymbol{X}_{1,:}^{\boldsymbol{\alpha}}q_{\boldsymbol{\alpha}}(\boldsymbol{X}),$$

where $q_{\boldsymbol{\alpha}}:\mathbb{R}^{N\times D}\to\mathbb{R}$ are S_N-invariant polynomials.

B.3 Triple-symmetric Polynomials

We define the doubly power-sum multi-symmetric polynomials for us to prove a similar result to Lemma B.2.

Definition B.4 (Doubly power-sum Multi-symmetric Polynomials (DMPs))**.** *Let $M\in\mathbb{N}$. The* doubly power-sum multi-symmetric polynomials $\{s_t^{(F)}:\mathbb{R}^{N\times(F+C)}\to\mathbb{R}\}_{t=1}^{T}\cup\{s_t^{(C)}:\mathbb{R}^{N\times(F+C)}\to\mathbb{R}\}_{t=1}^{T}$ *are defined by*

$$s_t^{(F)}(\boldsymbol{X})=\sum_{f=1}^{F}\boldsymbol{X}_{:,f}^{\boldsymbol{\alpha}^{(t)}},\quad s_t^{(C)}(\boldsymbol{X})=\sum_{c=1}^{C}\boldsymbol{X}_{:,F+c}^{\boldsymbol{\alpha}^{(t)}}$$

and are parameterized by $\boldsymbol{\alpha}^{(t)}\in\mathbb{N}_0^N$ such that $\|\boldsymbol{\alpha}\|_1\le M$, where $T=\binom{M+N}{N}$.

We now show that given $\boldsymbol{X}\in\mathbb{R}^{N\times(F+C)}$, the DMPs form a generating set for the ring of $(S_F\times S_C)$-invariant polynomials. This is later employed in the construction of the approximating neural networks in Theorem 4.2.

[2] We count the number of possible $\boldsymbol{\alpha}$ by introducing a slack variable $\alpha_{D+1}:=M-\sum_{i=1}^{D}\alpha_i$, so that $\sum_{i=1}^{D+1}\alpha_i=M$. The number of such non-negative integer solutions is equivalent to distributing M indistinguishable balls into $D+1$ distinguishable bins, which yields $\binom{M+D}{D}$ possibilities.

Lemma B.5. *Let $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$. Let $p : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}$ be an $S_F \times S_C$-invariant polynomial with degree M. Then, it can be written as a polynomial in the DMPs*

$$p(\boldsymbol{X}) = q(s_1^{(F)}(\boldsymbol{X}), \dots, s_T^{(F)}(\boldsymbol{X}), s_1^{(C)}(\boldsymbol{X}), \dots, s_T^{(C)}(\boldsymbol{X}),),$$

for some polynomial $q : \mathbb{R}^{2T} \to \mathbb{R}$, where $T = \binom{M+N}{N}$.

Proof. Denote the degree of p by M. For every $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$, we fix $\boldsymbol{X}^{(C)}$ and consider p as a function of $\boldsymbol{X}^{(F)}$, denoted by $p^{\boldsymbol{X}^{(C)}} : \boldsymbol{X}^{N\times F} \to \mathbb{R}$, such that

$$p(\boldsymbol{X}) = p^{\boldsymbol{X}^{(C)}}(\boldsymbol{X}^{(F)}) \tag{8}$$

The space of S_F-invariant polynomials over $\mathbb{R}^{N\times F}$ of bounded degree is a finite dimensional linear space and therefore has a basis $\{b_l\}_{l=1}^L : \mathbb{R}^{N\times F} \to \mathbb{R}$ with degree at most M, such that

$$p^{\boldsymbol{X}^{(C)}}(\boldsymbol{X}^{(F)}) = \sum_{l=1}^{L} c_l(\boldsymbol{X}^{(C)}) b_l(\boldsymbol{X}^{(F)}), \tag{9}$$

where the coefficients $\{c_l(\boldsymbol{X}^{(C)})\}_{l=1}^L \in \mathbb{R}$ depend on $\boldsymbol{X}^{(C)}$.

By Equations (8) and (9), For every $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$, we get

$$p(\boldsymbol{X}) = \sum_{l=1}^{L} c_l(\boldsymbol{X}^{(C)}) b_l(\boldsymbol{X}^{(F)}), \tag{10}$$

where the coefficients $\{c_l\}_{l=1}^L : \mathbb{R}^{N\times C} \to \mathbb{R}$ are now polynomials with degree at most M that depend on $\boldsymbol{X}^{(C)}$. Let $\sigma_C \in S_C$. Since p is S_C-invariant, we get

$$p(\boldsymbol{X}) = p(\sigma_C \cdot \boldsymbol{X}).$$

Namely,

$$\sum_{l=1}^{L} c_l(\boldsymbol{X}^{(C)}) b_l(\boldsymbol{X}^{(F)}) = \sum_{l=1}^{L} c_l(\sigma_C \cdot \boldsymbol{X}^{(C)}) b_l(\boldsymbol{X}^{(F)}).$$

Now, due to the fact that basis elements are linearly independent, we must have

$$\forall l \in [L], \quad c_l(\boldsymbol{X}^{(C)}) = c_l(\sigma_C \cdot \boldsymbol{X}^{(C)}).$$

Hence, each $\{c_l\}_{l=1}^L$ is an S_C-invariant polynomial. As a result, by Lemma B.2, each $\{c_l\}_{l=1}^L$ can be written as a polynomial $v_l : \mathbb{R}^T \to \mathbb{R}$ in the the PMPs $\{s_i\}_{i=1}^T : \mathbb{R}^{N\times C} \to \mathbb{R}$, i.e.

$$c_l(\boldsymbol{X}^{(C)}) = v_l(s_1(\boldsymbol{X}^{(C)}), \dots, s_T(\boldsymbol{X}^{(C)})) \tag{11}$$

where $T = \binom{M+N}{N}$.

Moreover, since b_l are invariant to S_F by construction, by Lemma B.2, each $\{b_l\}_{l=1}^L$ can be written as a polynomial $u_l : \mathbb{R}^T \to \mathbb{R}$ in the the multi-symmetric power-sum polynomials (PMPs) $\{s_i\}_{i=1}^T : \mathbb{R}^{N\times F} \to \mathbb{R}$, i.e.

$$b_l(\boldsymbol{X}^{(F)}) = u_l(s_1(\boldsymbol{X}^{(F)}), \dots, s_T(\boldsymbol{X}^{(F)})). \tag{12}$$

By substituting Equations (11) and (12) into Equation (10), we deduce that p is a doubly power-sum multi-symmetric polynomial. □

Similarly to Lemma B.3, the following lemma enables the decomposition of $(S_{F-1} \times S_C)$-invariant polynomials into a combination of polynomials that are $(S_F \times S_C)$-invariant. This decomposition is later used to express general $S_N \times S_F \times S_C$-equivariant polynomials, which facilitates an approximation of $S_N \times S_F \times S_C$-equivariant functions using $S_F \times S_C$-invariant functions in Lemma B.16.

Lemma B.6. *Let $p : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}$ be an $(S_{F-1}\times S_C)$-invariant polynomial with degree M. Then, there exist $S_F \times S_C$-invariant polynomials $\{q_{\boldsymbol{\kappa}} : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}\}_{\|\boldsymbol{\kappa}\|_1\le M}$ such that*

$$p(\boldsymbol{X}) = \sum_{\|\boldsymbol{\kappa}\|_1\le M} \boldsymbol{X}_{:,1}^{\boldsymbol{\kappa}} q_{\boldsymbol{\kappa}}(\boldsymbol{X}).$$

Proof. One can expand the polynomial $p(\boldsymbol{X})$ in $\boldsymbol{X}_{:,1}$ as

$$p(\boldsymbol{X}) = \sum_{\|\boldsymbol{\gamma}\|_1\le M} \boldsymbol{X}_{:,1}^{\boldsymbol{\gamma}} p_{\boldsymbol{\gamma}}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}), \tag{13}$$

where each $p_{\boldsymbol{\gamma}} : \mathbb{R}^{N\times(F-1+C)} \to \mathbb{R}$ is a polynomial of degree at most $M - \|\boldsymbol{\gamma}\|_1$.

Since derivatives commute with the action of $S_{F-1}\times S_C$, any partial derivative of an $(S_{F-1}\times S_C)$-invariant polynomial is also an invariant polynomial. Moreover, plugging $\boldsymbol{X}_{:,1} = 0$ into an $(S_{F-1}\times S_C)$-invariant polynomial $u : \mathbb{R}^{F\times C} \to \mathbb{R}$ gives an $(S_{F-1}\times S_C)$-invariant polynomial $u|_{\boldsymbol{X}_{:,1}=0} : \mathbb{R}^{(F-1)\times C} \to \mathbb{R}$. Hence, the polynomials

$$p_{\boldsymbol{\gamma}}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = \Big(\frac{\partial^{\|\boldsymbol{\gamma}\|_1}}{\partial \boldsymbol{X}_{:,1}^{\boldsymbol{\gamma}}} p\Big)(0,\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}), \quad \forall\, \|\boldsymbol{\gamma}\|_1 \le M$$

are $(S_{F-1}\times S_C)$-invariant.

Following Definition B.4 with respect to the the domain $\mathbb{R}^{N\times((F-1)+C)}$, the doubly multi-symmetric power-sum polynomials (DMPs) $\mathbb{R}^{N\times((F-1)+C)} \to \mathbb{R}$ map the variable $\boldsymbol{Z} \in \mathbb{R}^{N\times((F-1)+C)}$ to

$$\hat{s}_t^{(F)}(\boldsymbol{Z}) = \sum_{f=1}^{F-1} \boldsymbol{Z}_{:,f}^{\boldsymbol{\beta}^{(t)}} \quad \text{and} \quad \hat{s}_t^{(C)}(\boldsymbol{Z}) = \sum_{c=1}^{C} \boldsymbol{Z}_{:,F-1+c}^{\boldsymbol{\beta}^{(t)}},$$

where these polynomials are parameterized by $t \in [T]$ for $T = \binom{M+N}{N}$, in order to go through all multi-indeces $\boldsymbol{\beta}^{(t)} \in \mathbb{N}_0^N$ such that $\left\|\boldsymbol{\beta}^{(t)}\right\|_1 \le M$.

Similarly, following Definition B.4 with respect to the the domain $\mathbb{R}^{N\times(F+C)}$, the DMPs $s_t^{(C)}, s_t^{(F)} : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}$ map the variable $\boldsymbol{Y} \in \mathbb{R}^{N\times(F+C)}$ to

$$s_t^{(F)}(\boldsymbol{Y}) = \sum_{f=1}^{F} \boldsymbol{Y}_{:,f}^{\boldsymbol{\alpha}^{(t)}} \quad \text{and} \quad s_t^{(C)}(\boldsymbol{Y}) = \sum_{c=1}^{C} \boldsymbol{Y}_{:,F+c}^{\boldsymbol{\alpha}^{(t)}},$$

where these polynomials are parameterized by $t \in [T]$, in order to go through all multi-indeces $\boldsymbol{\alpha}^{(t)} \in \mathbb{N}_0^N$ such that $\left\|\boldsymbol{\alpha}^{(t)}\right\|_1 \le M$.

Without loss of generality, assume that the ordering of $\{\boldsymbol{\alpha}^{(t)}\}_{t=1}^T$ matches that of $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ over the T multi-indices that were set by $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$, i.e.

$$\forall t \in [T], \quad \boldsymbol{\alpha}^{(t)} = \boldsymbol{\beta}^{(t)}.$$

Thus, for each $t \in [T]$, we have that

$$\hat{s}_t^{(F-1)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = s_t^{(F)}(\boldsymbol{X}) - \boldsymbol{X}_{:,1}^{\boldsymbol{\beta}^{(t)}} \text{ and } \hat{s}_t^{(C)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = s_t^{(C)}(\boldsymbol{X}). \tag{14}$$

Each $p_{\boldsymbol{\gamma}}$ of (Equation (13)) is $(S_{F-1}\times S_C)$-invariant. Hence, by Lemma B.5, it can be expressed as a polynomial $\hat{q}_{\boldsymbol{\gamma}} : \mathbb{R}^{2T} \to \mathbb{R}$ in the DMPs $\mathbb{R}^{N\times((F-1)+C)} \to \mathbb{R}$ map the variable $\boldsymbol{Z} \in \mathbb{R}^{N\times((F-1)+C)}$, i.e.,

$$\begin{aligned}p_{\boldsymbol{\gamma}}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = \hat{q}_{\boldsymbol{\gamma}}\Big(&\hat{s}_1^{(F-1)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}),\dots,\hat{s}_T^{(F-1)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}),\\ &\hat{s}_1^{(C)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}),\dots,\hat{s}_T^{(C)}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C})\Big).\end{aligned}$$

and by Equation (14) it can be written as

$$p_{\boldsymbol{\gamma}}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = \hat{q}_{\boldsymbol{\gamma}}\Big(s_1^{(F)}(\boldsymbol{X}) - \boldsymbol{X}_{:,1}^{\boldsymbol{\beta}^{(1)}},\dots,s_T^{(F)}(\boldsymbol{X}) - \boldsymbol{X}_{:,1}^{\boldsymbol{\beta}^{(T)}}, \quad s_1^{(C)}(\boldsymbol{X}^{(C)}),\dots,s_T^{(C)}(\boldsymbol{X}^{(C)})\Big). \tag{15}$$

We can now expand $\hat{q}_{\boldsymbol{\gamma}}$ in the right-hand-side of (15) in monomials of $\boldsymbol{X}_{:,1}$, and write

$$p_{\boldsymbol{\gamma}}(\boldsymbol{X}_{:,2},\dots,\boldsymbol{X}_{:,F+C}) = \sum_{\|\boldsymbol{\delta}_{\boldsymbol{\gamma}}\|_1 \le M} \boldsymbol{X}_{:,1}^{\boldsymbol{\delta}_{\boldsymbol{\gamma}}} \widetilde{q}_{\boldsymbol{\delta}_{\boldsymbol{\gamma}}}\Big(s_1^{(F)}(\boldsymbol{X}),\dots,s_T^{(F)}(\boldsymbol{X}), \quad s_1^{(C)}(\boldsymbol{X}),\dots,s_T^{(C)}(\boldsymbol{X})\Big) \tag{16}$$

where $\boldsymbol{\delta}_{\boldsymbol{\gamma}}$ depends on $\boldsymbol{\gamma}$ and $\widetilde{q}_{\boldsymbol{\delta}_{\boldsymbol{\gamma}}}$ is an $S_F \times S_C$-invariant polynomial as it is a polynomial in the $S_F \times S_C$-invariant DMPs (see Definition B.4).

When substituting Equation (16) into Equation (13), we obtain

$$p(\boldsymbol{X}) = \sum_{\|\boldsymbol{\gamma}\|_1 \le M} \sum_{\|\boldsymbol{\delta}_{\boldsymbol{\gamma}}\|_1 \le M} \boldsymbol{X}_{:,1}^{\boldsymbol{\gamma}+\boldsymbol{\delta}_{\boldsymbol{\gamma}}} \widetilde{q}_{\boldsymbol{\delta}_{\boldsymbol{\gamma}}}\Big(s_1^{(F)}(\boldsymbol{X}),\dots,s_T^{(F)}(\boldsymbol{X}), \quad s_1^{(C)}(\boldsymbol{X}),\dots,s_T^{(C)}(\boldsymbol{X}),\Big) \tag{17}$$

and recall that by Equation (13), each $p_{\boldsymbol{\gamma}}$ is of degree at most $M - \|\boldsymbol{\gamma}\|_1$. Thus, by Equation (16) $\|\boldsymbol{\gamma}\|_1 + \|\boldsymbol{\delta}_{\boldsymbol{\gamma}}\|_1 \le M$.

Define $\kappa = \boldsymbol{\gamma} + \boldsymbol{\delta}_{\boldsymbol{\gamma}}$ and $q_\kappa(\boldsymbol{X})$ as a finite linear combination of the $\widetilde{q}_{\boldsymbol{\delta}}$ from Equation (16), such that Equation (17) can be written as

$$p(\boldsymbol{X}) = \sum_{\|\kappa\|_1 \le M} \boldsymbol{X}_{:,1}^{\kappa} q_\kappa(\boldsymbol{X}).$$

Note that as $q_\kappa(\boldsymbol{X})$ is defined as a finite linear combination of the $\widetilde{q}_{\boldsymbol{\delta}}$, it is also $S_F \times S_C$-invariant, concluding the proof. ☐

By the exact same proof technique a similar lemma holds.

Lemma B.7. *Let $p : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}$ be an $(S_F \times S_{C-1})$-invariant polynomial with degree M. Then, there exist $S_F \times S_C$-invariant polynomials $\{q_{\boldsymbol{\kappa}} : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}\}_{\|\boldsymbol{\kappa}\|_1 \le \max(F,C)}$ such that*

$$p(\boldsymbol{X}) = \sum_{\|\boldsymbol{\kappa}\|_1 \le M} \boldsymbol{X}_{:,F+1}^{\boldsymbol{\kappa}} q_{\boldsymbol{\kappa}}(\boldsymbol{X}).$$

B.4 Supporting Lemmas

The following Lemma B.8 is close to Lemma 4 in Segol and Lipman [2019]. Lemma B.8 extends Lemma 4 in Segol and Lipman [2019] by showing that it holds for general invariant sets rather than the cube and any subgroup of the permutations as the symmetry group. In the proof of the lemma we use the same well-known proof technique of symmetrization to enforce equivariance. This lemma is later used in the proof of Lemma B.16 to reduce the problem to approximating polynomial functions.

Lemma B.8. *Let $G \le S_K$. Let $\mathcal{K} \subset \mathbb{R}^K$ be a compact domain such that $\mathcal{K} = \cup_{g\in G} g\mathcal{K}$. Then the space of G-equivariant polynomials $p : \mathbb{R}^K \to \mathbb{R}^L$ is dense in the space of continuous G-equivariant functions $f : \mathcal{K} \to \mathbb{R}^L$ in L_∞ over $\mathcal{K}$.*

Proof. Let $\epsilon > 0$ be arbitrary, and let $f : \mathcal{K} \to \mathbb{R}^L$ be a continuous G-equivariant function. For each output coordinate $f_l : \mathcal{K} \to \mathbb{R}$, the classical Stone–Weierstrass theorem guarantees the existence of a polynomial $p_l : \mathbb{R}^K \to \mathbb{R}$ such that

$$|f_l(\boldsymbol{X}) - p_l(\boldsymbol{X})| \le \epsilon \quad \text{for all } \boldsymbol{X} \in \mathcal{K}.$$

Define the polynomial $p : \mathbb{R}^K \to \mathbb{R}^L$ by

$$p(\boldsymbol{X}) = (p_1(\boldsymbol{X}),\dots,p_L(\boldsymbol{X})) \in \mathbb{R}^L, \quad \forall \boldsymbol{X} \in \mathbb{R}^K$$

The polynomial p is not necessarily G-equivariant. We now define the symmetrization of p, denoted by $\widetilde{p} : \mathbb{R}^K \to \mathbb{R}^L$, by

$$\widetilde{p}(\boldsymbol{X}) = \frac{1}{|G|} \sum_{g \in G} g \cdot p(g^{-1} \cdot \boldsymbol{X}).$$

Now, since f is G-equivariant, we have

$$\begin{aligned} \|f(\boldsymbol{X}) - \widetilde{p}(\boldsymbol{X})\|_\infty &= \left\| \frac{1}{|G|} \sum_{g \in G} g \cdot \left(f(g^{-1} \cdot \boldsymbol{X}) - p(g^{-1} \cdot \boldsymbol{X}) \right) \right\|_\infty \\ &\leq \frac{1}{|G|} \sum_{g \in G} \left\| f(g^{-1} \cdot \boldsymbol{X}) - p(g^{-1} \cdot \boldsymbol{X}) \right\|_\infty , \end{aligned}$$

and by the construction of p

$$\|f(\boldsymbol{X}) - \widetilde{p}(\boldsymbol{X})\|_\infty = \left\| \frac{1}{|G|} \sum_{g \in G} g \cdot \left(f(g^{-1} \cdot \boldsymbol{X}) - p(g^{-1} \cdot \boldsymbol{X}) \right) \right\|_\infty \leq \epsilon.$$

Thus, $\widetilde{p}$ is an G-equivariant polynomial map that uniformly approximates f, concluding the proof. $\square$

Given G, a subgroup of the permutation group, we define a G-descriptor as a polynomial function that maps two matrices to the same output if and only if they differ by an element of G. These G-descriptors are later used to approximate invariant functions in Lemma B.16.

Definition B.9 (G-descriptor). *Let $G \leq S_N$ and $\mathcal{K} \subseteq \mathbb{R}^{N \times D}$. We call a G-invariant polynomial function $q : \mathcal{K} \to \mathbb{R}^L$, a G-*descriptor*, if it satisfies:*

$$\forall \boldsymbol{X}, \boldsymbol{Y} \in \mathcal{K}, \quad q(\boldsymbol{X}) = q(\boldsymbol{Y}) \iff \exists g \in G : \boldsymbol{Y} = g \cdot \boldsymbol{X}.$$

Next, we cite Lemma 3 in from Maron et al. [2020], which shows the existence of a descriptors given a subgroup of the permutation group. These G-descriptors are later used to approximate invariant functions in Lemma B.16.

Lemma B.10 (Lemma 3 in Maron et al. [2020]). *Let $G \leq S_N$ be a subgroup. Then, there exists a G-descriptor $u : \mathbb{R}^{N \times D} \to \mathbb{R}^L$.*

We also mention a triviality from point set topology (for example, see Lemma 5 from Maron et al. [2020]).

Lemma B.11 (Lemma 5 in Maron et al. [2020]). *Let $\mathcal{K} \subset \mathbb{R}^L$ be a compact domain and $f : \mathcal{K} \to \mathbb{R}$ be a continuous function such that $f = h \circ g$, where g is continuous. Then h is continuous on $g(\mathcal{K})$.*

B.5 Supporting Universal Approximation Properties

First, we mention a classical Universal Approximation Theorem (UAT) as it is later employed to show universality in Lemma B.16.

Theorem B.12 (Universal approximation theorem Pinkus [1999]). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous, non-polynomial function. Then, for every $M, L \in \mathbb{N}$, compact $\mathcal{K} \subseteq \mathbb{R}^M$, continuous function $f : \mathcal{K} \to \mathbb{R}^L$, and $\varepsilon > 0$, there exist $D \in \mathbb{N}, \boldsymbol{W}_1 \in \mathbb{R}^{D \times M}, \boldsymbol{b}_1 \in \mathbb{R}^D$ and $\boldsymbol{W}_2 \in \mathbb{R}^{L \times D}$ such that*

$$\sup_{\boldsymbol{x} \in \mathcal{K}} \| f(\boldsymbol{x}) - \boldsymbol{W}_2 \sigma \left(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1 \right) \| \leq \varepsilon.$$

We next study universality with respect to composition of functions. For that, we first define universal approximation rigorously.

Definition B.13. *Let $D_1, D_2 \in \mathbb{N}$. A space of continuous functions $\mathcal{N}$ from a domain $Q_1 \subset \mathbb{R}^{D_1}$ to a domain $Q_2 \subset \mathbb{R}^{D_2}$ is said to be a* universal approximator *of another space of continuous functions $\mathcal{B}$ from Q_1 to Q_2, if $\mathcal{N} \subset \mathcal{B}$, and for every $f \in \mathcal{B}$, every compact domain $C \subset \mathbb{R}^{D_1}$ such that $C \subset Q_1$, and every $\epsilon > 0$, there is a function $q \in \mathcal{N}$ such that for every $x \in C$*

$$\|f(x) - q(x)\|_\infty < \epsilon.$$

The following theorem is a trivial approximation theoretic result.

Theorem B.14. *Let $L \in \mathbb{N}$, and $D_1, \dots D_{L+1} \in \mathbb{N}$. Let $\mathcal{N}_l$ be a* universal approximator *of $\mathcal{B}_l$ for $l \in [L]$. Let the domain of the functions $\mathcal{B}_l$ be $Q_l \subset \mathbb{R}^{D_l}$ and the range be $Q_{l+1} \subset \mathbb{R}^{D_{l+1}}$, for $l \in [L]$. Consider the function spaces*

$$\mathcal{N} := \{\theta_L \circ \dots \circ \theta_1 \mid \theta_l \in \mathcal{N}_l,\ l \in [L]\},$$

and

$$\mathcal{B} := \{f_L \circ \dots \circ f_1 \mid f_l \in \mathcal{B}_l,\ l \in [L]\}.$$

Then $\mathcal{N}$ is a universal approximator of $\mathcal{B}$.

B.6 The proof of Theorem 4.2

We define a set called the exclusion set, that is used in Theorem 4.2 to enable the composition of descriptors under our triple symmetry.

Definition B.15 (The exclusion set)**.** *The set*

$$\begin{aligned}\mathcal{E} = &\bigcup_{1 \le f_1 < f_2 \le F} \left\{ \boldsymbol{X} \in \mathbb{R}^{N \times (F+C)} \,\middle|\, \sum_{n=1}^{N} \boldsymbol{X}_{n,f_1} = \sum_{n=1}^{N} \boldsymbol{X}_{n,f_2} \right\} \bigcup \\ &\bigcup_{1 \le c_1 < c_2 \le C} \left\{ \boldsymbol{X} \in \mathbb{R}^{N \times (F+C)} \,\middle|\, \sum_{n=1}^{N} \boldsymbol{X}_{n,F+c_1} = \sum_{n=1}^{N} \boldsymbol{X}_{n,F+c_2} \right\} \bigcup \\ &\bigcup_{1 \le n_1 < n_2 \le N} \left\{ \boldsymbol{X} \in \mathbb{R}^{N \times (F+C)} \,\middle|\, \sum_{j=1}^{F+C} \boldsymbol{X}_{n_1,j} = \sum_{j=1}^{F+C} \boldsymbol{X}_{n_2,j} \right\}\end{aligned}$$

is called the exclusion set corresponding to $\mathbb{R}^{N \times (F+C)}$.

The exclusion set $\mathcal{E}$ is a finite union of linear subspaces of co-dimension one. We consider input matrices $\boldsymbol{X} \in \mathbb{R}^{N \times (F+C)}$, where $\boldsymbol{X}^{(F)}$ correspond to continuous real-valued features and $\boldsymbol{X}^{(C)}$ encode one-hot vectors with entries in $\{0, 1\}$, assigning probability one to exactly one class and zero to the others in each row. The real-valued features $\boldsymbol{X}^{(F)}$ vary continuously in $\mathbb{R}^F$. The set of features that lie inside the exclusion set $\mathcal{E}$ is meagre with respect to the standard topology of the feature space (more specifically, it is nowhere dense). Hence, being outside $\mathcal{E}$ is the generic case in the topological sense. From a probabilistic perspective, if the features are sampled from a joint distribution that is continuous with respect to the Lebesgue measure (in the Radon–Nikodym sense), then the features belong to $\{\boldsymbol{X}^{(F)} \mid \boldsymbol{X} \in \mathcal{E}\}$ in probability zero. Hence, for the features to belong to $\mathcal{E}$ is event that almost surely does not occur. With regard to the labels, the exclusion set $\mathcal{E}$ rules out the special scenario in which multiple classes have exactly the same number of examples. If labels are samples from a reasonable joint probability space, where labels are not constrained to be balanced, such an even typically occurs in low probability. To summarize, from both the feature and label perspectives, the data typically lies outside the exclusion set $\mathcal{E}$, and hence considering a domain $\mathcal{K}$ disjoint from $\mathcal{E}$ is reasonable. A similar philosophy regarding an exclusion set is adopted by Maron et al. [2020].

Note. The following Lemma B.16 follows the same proof strategy as Theorem 3 of Maron et al. [2020] with added elements from Segol and Lipman [2019] to correct the issue detailed in Section 4.2 and extend the Theorem for our triple-symmetry.

We define *TSNet** as the class of architectures obtained by removing the final label projection layer from TSNets. Formally, a TSNet* architecture is composed solely of interleaved $(S_N \times S_F \times S_C)$-equivariant linear layers (as defined in Equation (3)) and activation functions.

To demonstrate that TSNets can approximate functions that are $(S_N \times S_C)$-equivariant and S_F-invariant to arbitrary accuracy, we prove that TSNet* can approximate any $(S_N \times S_F \times S_C)$-equivariant function arbitrarily well. This result is then used in the proof of Theorem 4.2, before the S_F-invariance is imposed by the final label projection layer.

Lemma B.16. *Let $\mathcal{K} \subset \mathbb{R}^{N \times (F+C)}$ be a compact domain such that $\mathcal{K} = \cup_{g \in S_N \times S_F \times S_C} g\mathcal{K}$ and $\mathcal{K} \cap \mathcal{E} = \emptyset$, where $\mathcal{E} \subset \mathbb{R}^{N \times (F+C)}$ is the exclusion set corresponding to $\mathbb{R}^{N \times (F+C)}$ (Definition B.15). Then, TSNet* are universal approximators in L_∞ of continuous $\mathcal{K} \to \mathbb{R}^{N \times (F+C)}$ functions that are $(S_N \times S_F \times S_C)$-equivariant.*

Proof. **Reduction to polynomials.** Let $q : \mathcal{K} \to \mathbb{R}^{N\times(F+C)}$ be a continuous $(S_N \times S_F \times S_C)$-equivariant function, and let $\epsilon > 0$. By Lemma B.8, there exists an $(S_N \times S_F \times S_C)$-equivariant polynomial $p : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times(F+C)}$ of some degree M, such that

$$\sup_{\boldsymbol{X}\in\mathcal{K}} \|p(\boldsymbol{X}) - q(\boldsymbol{X})\|_\infty < \frac{\epsilon}{2}. \tag{18}$$

Until the rest of the proof, we show how to approximate the $(S_N \times S_F \times S_C)$-equivariant polynomial $p : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times(F+C)}$ by a TSNet*.

Reduction to $p_{1,1}$ and $p_{1,F+1}$. Recall that we denote by $p_{i,j}(\boldsymbol{X}) \in \mathbb{R}$ the (i, j)-th entry of $p(\boldsymbol{X})$. The polynomial p is $(S_N \times S_F \times S_C)$-equivariant, meaning that by definition

$$(\sigma_N, \sigma_F, \sigma_C) \cdot p(\boldsymbol{X}) = p((\sigma_N, \sigma_F, \sigma_C) \cdot \boldsymbol{X}),$$

for all $(\sigma_N, \sigma_F, \sigma_C) \in S_N \times S_F \times S_C$. Applying, $\sigma_N = (1n)$ and $\sigma_F = (1f)$, we obtain

$$p_{n,f}(\boldsymbol{X}) = p_{1,1}((\sigma_N, \sigma_F, e) \cdot \boldsymbol{X}), \quad \forall n \in [N], f \in [F], \tag{19}$$

and by applying, $\sigma_N = (1n)$ and $\sigma_C = (1c)$, we obtain

$$p_{n,F+c}(\boldsymbol{X}) = p_{1,1}((\sigma_N, e, \sigma_C) \cdot \boldsymbol{X}), \quad \forall n \in [N], c \in [C]. \tag{20}$$

Thus, every entry in the output of p can be recovered from $p_{1,1}$ or $p_{1,F+1}$, reducing the problem to the approximation of $p_{1,1}$ and $p_{1,F+1}$, as we show next.

The symmetries of $p_{1,1}$ and $p_{1,F+1}$. The output element in the first row and column $p_{1,1}$ can be viewed as the composition of p with the projection map $\pi_{1,1} : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}$, defined by

$$\pi_{1,1}(\boldsymbol{X}) = \boldsymbol{X}_{1,1},$$

which extracts the $(1, 1)$-st element of $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$. Namely, $p_{1,1} = \pi_{1,1} \circ p$.

The projection map $\pi_{1,1}$ is invariant the action of $S_{N-1}\times S_{F-1}\times S_C$. This is because the permutations leave the $(1, 1)$-st element unaffected while affecting all other elements. This symmetry is captured by the group $S_{N-1} \times S_{F-1} \times S_C$. Now, since p is an $(S_N \times S_F \times S_C)$-equivariant map, it is also an $(S_{N-1} \times S_{F-1} \times S_C)$-equivariant map. Thus, $p_{1,1}$ is a composition of an $(S_{N-1} \times S_{F-1} \times S_C)$-equivariant map p and an $(S_{N-1} \times S_{F-1} \times S_C)$-invariant map $\pi_{1,1}$, making $p_{1,1}$ an $(S_{N-1} \times S_{F-1} \times S_C)$-invariant function.

By the same reasoning, $p_{1,F+1}$ is an $(S_{N-1} \times S_F \times S_{C-1})$-invariant function.

A $(S_{F-1} \times S_C)$-descriptor $Q^{(F)}$. By Lemma B.10, there exists an $S_{F-1} \times S_C$-descriptor (see Definition B.9) $q^{(F)} : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}^{L_F}$ with degree M_F, namely, a polynomial such that

$$\begin{aligned}&\forall \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{(F+C)\times 2} : \\ &\Big(q^{(F)}(\boldsymbol{Z}) = q^{(F)}(\boldsymbol{Y})\Big) \iff \Big(\exists (\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C : \; \boldsymbol{Z} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Y}\Big).\end{aligned} \tag{21}$$

We define a polynomial $Q^{(F)} : \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E} \to \mathbb{R}^{N\times L_F}$ as follows. Given $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$, the polynomial is defined by taking the row vector $\sum_{i=1}^{N} \boldsymbol{X}_{i,:} \in \mathbb{R}^{F+C}$ and appending it to each row of $\boldsymbol{X}$ forming a new dimension and creating the intermediate tensor

$$\hat{\boldsymbol{X}} = [\boldsymbol{X}, \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{X}_{i,:}] \in \mathbb{R}^{N\times(F+C)\times 2}, \tag{22}$$

To define $Q^{(F)}$, the function $q^{(F)}$ is then applied on $\hat{\boldsymbol{X}}$ via

$$\forall n \in [N], \quad Q^{(F)}(\boldsymbol{X})_{n,:} := q^{(F)}(\hat{\boldsymbol{X}}_{n,:,:}) = q^{(F)}([\boldsymbol{X}_{n,:}, \sum_{i=1}^{N} \boldsymbol{X}_{i,:}]). \tag{23}$$

By construction, this makes $Q^{(F)}$ an S_N-equivariant polynomial. Note that given a permutation $(\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C$, for every $\boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{N\times(F+C)}$, we have

$$\begin{aligned}&\Big(\forall n : \boldsymbol{Z}_{n,:} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Y}_{n,:}\Big) \\ \iff &\Big(\forall n : \Big[\boldsymbol{Z}, \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Z}_{i,:}\Big]_{n,:,:} = \Big[(\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Y}, (\sigma_{F-1}, \sigma_C) \cdot \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Y}_{i,:}\Big]_{n,:,:}\Big) \\ \iff &\Big(\forall n : \hat{\boldsymbol{Z}}_{n,:,:} = (\sigma_{F-1}, \sigma_C) \cdot \hat{\boldsymbol{Y}}_{n,:,:}\Big).\end{aligned}$$

Thus, we obtained that

$$\Big(\forall n : \boldsymbol{Z}_{n,:} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Y}_{n,:}\Big) \iff \Big(\forall n : \hat{\boldsymbol{Z}}_{n,:,:} = (\sigma_{F-1}, \sigma_C) \cdot \hat{\boldsymbol{Y}}_{n,:,:}\Big). \tag{24}$$

We now prove that $Q^{(F)}$ is an S_N-equivariant polynomial that satisfies

$$\begin{aligned} &\forall \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{N \times (F+C)} \setminus \mathcal{E} : \\ &\quad Q^{(F)}(\boldsymbol{Z}) = Q^{(F)}(\boldsymbol{Y}) \iff \exists (\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C : \boldsymbol{Y} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}. \end{aligned} \tag{25}$$

Direction $\Leftarrow$ of Equation (25). Assume there exists $(\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C$, such that $\boldsymbol{Y} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z} \in \mathbb{R}^{N \times (F+C)}$. By definition, the action of (σ_{F-1}, σ_C) permutes the columns of $\boldsymbol{Z}$ identically across all rows, namely,

$$\forall n \in [N] : \quad \boldsymbol{Y}_{n,:} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}_{n,:}$$

and by Equation (24)

$$\forall n \in [N] : \quad \hat{\boldsymbol{Y}}_{n,:,:} = (\sigma_{F-1}, \sigma_C) \cdot \hat{\boldsymbol{Z}}_{n,:,:}. \tag{26}$$

By Equations (21) and (26), we obtain

$$\forall n \in [N] : \quad q^{(F)}(\hat{\boldsymbol{Z}}_{n,:,:}) = q^{(F)}(\hat{\boldsymbol{Y}}_{n,:,:}),$$

and by the definition of $Q^{(F)}$ in Equation (23), we have that $Q^{(F)}(\boldsymbol{Y}) = Q^{(F)}(\boldsymbol{Z})$.

Direction $\Rightarrow$ of Equation (25). We now show the opposite direction. Assume that $Q^{(F)}(\boldsymbol{Y}) = Q^{(F)}(\boldsymbol{Z})$ for some $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N \times (F+C)} \setminus \mathcal{E}$. By the definition of $Q^{(F)}$ in Equation (23), we have that

$$\forall n \in [N] : \quad q^{(F)}(\hat{\boldsymbol{Z}}_{n,:,:}) = q^{(F)}(\hat{\boldsymbol{Y}}_{n,:,:}),$$

and by Equation (21), the equality implies the existence of a permutation $(\sigma_{F-1}^{(n)}, \sigma_C^{(n)}) \in S_{F-1} \times S_C$ for each $n \in [N]$, such that

$$\hat{\boldsymbol{Y}}_{n,:,:} = (\sigma_{F-1}^{(n)}, \sigma_C^{(n)}) \cdot \hat{\boldsymbol{Z}}_{n,:,:}. \tag{27}$$

Next, we would like to show that

$$\forall n_1, n_2 \in [N] : \quad (\sigma_{F-1}^{(n_1)}, \sigma_C^{(n_1)}) = (\sigma_{F-1}^{(n_2)}, \sigma_C^{(n_2)}).$$

For that, consider two indices $n_1 \neq n_2 \in [N]$. By Equation (27) and the definition of $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{Y}}$ (Equation (22)), we get

$$\mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Y}_{i,:} = (\sigma_{F-1}^{(n_1)}, \sigma_C^{(n_1)}) \cdot \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Z}_{i,:} \text{ and } \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Y}_{i,:} = (\sigma_{F-1}^{(n_2)}, \sigma_C^{(n_2)}) \cdot \mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Z}_{i,:}, \tag{28}$$

Since $\mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Z}_{i,:}$ has equal rows (and so does $\mathbf{1}_N \sum_{i=1}^{N} \boldsymbol{Y}_{i,:}$) and $S_{F-1} \times S_C$ permutes all columns simultaneously, Equation (28) can be written as

$$(\sigma_{F-1}^{(n_1)}, \sigma_C^{(n_1)}) \cdot \sum_{i=1}^{N} \boldsymbol{Z}_{i,:} = (\sigma_{F-1}^{(n_2)}, \sigma_C^{(n_2)}) \cdot \sum_{i=1}^{N} \boldsymbol{Z}_{i,:}. \tag{29}$$

Recall that $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N \times (F+C)} \setminus \mathcal{E}$. Hence the first F entries of $\sum_{i=1}^{N} \boldsymbol{Y}_{i,:}$ and $\sum_{i=1}^{N} \boldsymbol{Z}_{i,:}$, each have F distinct elements. Moreover, the last C entries of $\sum_{i=1}^{N} \boldsymbol{Y}_{i,:}$ and $\sum_{i=1}^{N} \boldsymbol{Z}_{i,:}$, each have C distinct elements. Therefore, the only way that Equation (29) holds is when $\sigma_{F-1}^{(n_1)} = \sigma_{F-1}^{(n_2)}$ and $\sigma_C^{(n_1)} = \sigma_C^{(n_2)}$. Thus, all row–wise permutations coincide, i.e.

$$\exists (\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C,\ \forall i \in [N] : \quad (\sigma_{F-1}^{(i)}, \sigma_C^{(i)}) = (\sigma_{F-1}, \sigma_C),$$

and consequently by Equations (24) and (27),

$$\forall i \in [N] : \quad \boldsymbol{Y}_{i,:} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}_{i,:}.$$

Thus, $Q^{(F)} : \mathbb{R}^{N \times (F+C)} \to \mathbb{R}^{N \times L_F}$ is an S_N-equivariant polynomial that satisfies Equation (25).

A $(S_{N-1} \times S_{F-1} \times S_C)$-descriptor $Q^{(NF)}$. By Lemma B.10, there exists an S_{N-1}-invariant polynomial $q^{(N)} : \mathbb{R}^{N\times 3} \to \mathbb{R}^{L_N}$ with degree M_N, such that

$$\forall \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{N\times 3} : \quad \Big(q^{(N)}(\boldsymbol{Z}) = q^{(N)}(\boldsymbol{Y}) \Big) \iff \Big(\exists \sigma_{N-1} \in S_{N-1} : \; \boldsymbol{Z} = \sigma_{N-1} \cdot \boldsymbol{Y} \Big). \tag{30}$$

We define a polynomial $Q^{(NF)} : \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E} \to \mathbb{R}^{L_N \times L_F}$, where L_F is the output dimension of the polynomial descriptor $q^{(F)}$ given in Equation (21), as follows. Given an input $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$, the output $Q^{(NF)}(\boldsymbol{X})$ is defined via the following composition of computations.

(i) Compute $Q^{(F)}(\boldsymbol{X}) \in \mathbb{R}^{N \times L_F}$, as defined in Equation (23).

(ii) Consider the column vectors $\sum_{f=1}^{F} \boldsymbol{X}_{:,f} \in \mathbb{R}^N$ and $\sum_{c=1}^{C} \boldsymbol{X}_{:,F+c} \in \mathbb{R}^N$. Append these vectors to $Q_F(\boldsymbol{X})$ forming a new dimension. Namely, construct the intermediate tensor

$$\widetilde{\boldsymbol{X}} = [Q^{(F)}(\boldsymbol{X}), \sum_{f=1}^{F} \boldsymbol{X}_{:,f} \mathbf{1}_{L_F}^{\top}, \sum_{c=1}^{C} \boldsymbol{X}_{:,F+c} \mathbf{1}_{L_F}^{\top}] \in \mathbb{R}^{N \times L_F \times 3}, \tag{31}$$

(iii) Apply the function $q^{(N)}$ on $\widetilde{\boldsymbol{X}}$ to define $Q^{(NF)}(\boldsymbol{X})$ via

$$\begin{aligned} \forall l \in [L_F] : \quad Q^{(NF)}(\boldsymbol{X})_{:,l} &:= q^{(N)}(\widetilde{\boldsymbol{X}}_{:,l,:}) \\ &= q^{(N)}([Q^{(F)}(\boldsymbol{X})_{:,l}, \sum_{f=1}^{F} \boldsymbol{X}_{:,f}, \sum_{c=1}^{C} \boldsymbol{X}_{:,F+c}]). \end{aligned} \tag{32}$$

By construction, $Q^{(NF)}$ is a polynomial from $\mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ to $\mathbb{R}^{L_N \times L_F}$. We now prove that $Q^{(NF)}$ satisfies $\forall \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$:

$$\begin{gathered} Q^{(NF)}(\boldsymbol{Z}) = Q^{(NF)}(\boldsymbol{Y}) \\ \Updownarrow \\ \exists (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \in S_{N-1} \times S_{F-1} \times S_C : \quad \boldsymbol{Y} = (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}. \end{gathered} \tag{33}$$

Direction $\Uparrow$ of Equation (33). Let $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$. Suppose there exists $(\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \in S_{N-1} \times S_{F-1} \times S_C$ such that $\boldsymbol{Y} = (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}$. Recall that $Q^{(F)}$ is S_N-equivariant. This, together with Equation (25), gives

$$\forall l \in [L_F] : \quad Q^{(F)}(\boldsymbol{Y})_{:,l} = \sigma_{N-1} \cdot Q^{(F)}(\boldsymbol{Z})_{:,l}. \tag{34}$$

Substituting Equation (34) into the definition of $\widetilde{\boldsymbol{Y}}$ in Equation (31), we get

$$\widetilde{\boldsymbol{Y}} = [\sigma_{N-1} \cdot Q^{(F)}(\boldsymbol{Z}), \sum_{f=1}^{F} \boldsymbol{Y}_{:,f} \mathbf{1}_{L_F}^{\top}, \sum_{c=1}^{C} \boldsymbol{Y}_{:,F+c} \mathbf{1}_{L_F}^{\top}].$$

Since $\boldsymbol{Y} = (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}$, we obtain, by the fact that summation is invariant to permutations,

$$\widetilde{\boldsymbol{Y}} = \sigma_{N-1} \cdot [Q^{(F)}(\boldsymbol{Z}), \sum_{f=1}^{F} \boldsymbol{Z}_{:,f} \mathbf{1}_{L_F}^{\top}, \sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} \mathbf{1}_{L_F}^{\top}] = \sigma_{N-1} \cdot \widetilde{\boldsymbol{Z}}.$$

Hence, by Equation (30)

$$\forall l \in [L_F] : \quad q^{(N)}(\widetilde{\boldsymbol{Z}}_{:,l,:}) = q^{(N)}(\widetilde{\boldsymbol{Y}}_{:,l,:}),$$

and by the definition of $Q^{(F)}$ in Equation (32), we have that $Q^{(NF)}(\boldsymbol{Y}) = Q^{(NF)}(\boldsymbol{Z})$.

Direction $\Downarrow$ of Equation (33). We now show the opposite direction. Let $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ and assume $Q^{(NF)}(\boldsymbol{Z}) = Q^{(NF)}(\boldsymbol{Y})$. By the definition of $Q^{(F)}$ in Equation (32), we have that

$$\forall l \in [L_F] : \quad q^{(N)}(\widetilde{\boldsymbol{Z}}_{:,l,:}) = q^{(N)}(\widetilde{\boldsymbol{Y}}_{:,l,:}),$$

and by Equation (30), the equality implies the existence of a permutation $\sigma_{N-1}^{(l)} \in S_{N-1}$ for each $l \in [L_F]$, such that

$$\widetilde{\boldsymbol{Y}}_{:,l,:} = \sigma_{N-1}^{(l)} \cdot \widetilde{\boldsymbol{Z}}_{:,l,:}. \tag{35}$$

Next, we would like to show that

$$\forall l_1, l_2 \in [L_F]: \quad \sigma_{N-1}^{(l_1)} = \sigma_{N-1}^{(l_2)}.$$

For that, consider two indices $l_1 \neq l_2 \in [L_F]$. By Equation (35) and the definition of $\widetilde{\boldsymbol{Z}}$ and $\widetilde{\boldsymbol{Y}}$ (Equation (31)), we get

$$\sum_{f=1}^{F} \boldsymbol{Y}_{:,f} \mathbf{1}_{L_F}^{\top} = \sigma_{N-1}^{(l_1)} \cdot \sum_{f=1}^{F} \boldsymbol{Z}_{:,f} \mathbf{1}_{L_F}^{\top} \text{ and } \sum_{c=1}^{C} \boldsymbol{Y}_{:,F+c} \mathbf{1}_{L_F}^{\top} = \sigma_{N-1}^{(l_1)} \cdot \sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} \mathbf{1}_{L_F}^{\top}, \tag{36}$$

and similarly for $\sigma_{N-1}^{(l_2)}$.

Since $\sum_{f=1}^{F} \boldsymbol{Z}_{:,f} \mathbf{1}_{L_F}^{\top}$ and $\sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} \mathbf{1}_{L_F}^{\top}$ has equal columns (and so does $\sum_{f=1}^{F} \boldsymbol{Z}_{:,f} \mathbf{1}_{L_F}^{\top}$ and $\sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} \mathbf{1}_{L_F}^{\top}$) and $S_{F-1} \times S_C$ permutes all rows simultaneously, Equation (36) can be written as

$$\sigma_{N-1}^{(l_1)} \cdot \sum_{f=1}^{F} \boldsymbol{Z}_{:,f} = \sigma_{N-1}^{(l_2)} \cdot \sum_{f=1}^{F} \boldsymbol{Z}_{:,f} \text{ and } \sigma_{N-1}^{(l_1)} \cdot \sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} = \sigma_{N-1}^{(l_2)} \cdot \sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c} \tag{37}$$

Recall that $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N \times (F+C)} \setminus \mathcal{E}$. Hence $\sum_{f=1}^{F} \boldsymbol{Y}_{:,f}, \sum_{f=1}^{F} \boldsymbol{Z}_{:,f}, \sum_{c=1}^{C} \boldsymbol{Y}_{:,F+c}$ and $\sum_{c=1}^{C} \boldsymbol{Z}_{:,F+c}$, each have N distinct elements. Therefore the only way that Equation (37) holds is when all col–wise permutations coincide, i.e.,

$$\exists \sigma_{N-1} \in S_{N-1}, \forall l \in [L_F]: \quad \sigma_{N-1}^{(l)} = \sigma_{N-1}.$$

and by Equation (35)

$$\forall l \in [L_F]: \quad Q^{(F)}(\boldsymbol{Y})_{:,l} = \sigma_{N-1} \cdot Q^{(F)}(\boldsymbol{Z})_{:,l}.$$

Recall that $Q^{(F)}$ is S_N-equivariant, meaning,

$$Q^{(F)}(\boldsymbol{Y}) = Q^{(F)}(\sigma_{N-1} \cdot \boldsymbol{Z}),$$

and by Equation (25)

$$\exists (\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C: \quad \boldsymbol{Y} = (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}.$$

Thus, $Q^{(NF)}: \mathbb{R}^{N \times (F+C)} \to \mathbb{R}^{L_N \times L_F}$ satisfies Equation (33). Note that we can also flatten it to be $Q^{(NF)}: \mathbb{R}^{N \times (F+C)} \to \mathbb{R}^{L_N L_F}$.

Expressing $q^{(F)}$ and $q^{(N)}$ using DMPs. Recall the property of $q^{(F)}: \mathbb{R}^{(F+C) \times 2} \to \mathbb{R}^{L_F}$ with degree M_F, in Equation (21):

$$\forall \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{(F+C) \times 2}: \quad q^{(F)}(\boldsymbol{Z}) = q^{(F)}(\boldsymbol{Y}) \iff \exists (\sigma_{F-1}, \sigma_C) \in S_{F-1} \times S_C : \boldsymbol{Z} = (\sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Y}.$$

Let us focus on the l-the output dimension $q_l^{(F)}: \mathbb{R}^{(F+C) \times 2} \to \mathbb{R}$ and denote it by $\hat{q}^{(F)}$ for brevity. By Lemma B.6, the $(S_{F-1} \times S_C)$-invariant polynomial $\hat{q}^{(F)}: \mathbb{R}^{(F+C) \times 2} \to \mathbb{R}$ can be expressed by

$$\hat{q}^{(F)}(\boldsymbol{V}) = \sum_{\|\boldsymbol{\beta}\|_1 \leq M_F} \boldsymbol{V}_{1,:}^{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}(\boldsymbol{V}), \tag{38}$$

where $w_{\boldsymbol{\beta}}: \mathbb{R}^{(F+C) \times 2} \to \mathbb{R}$ are $(S_F \times S_C)$-invariant polynomials with degree at most M_F.

The doubly multi-symmetric power-sum polynomials (DMPs) $s_t^{(C)}, s_t^{(F)}: \mathbb{R}^{(F+C) \times 2} \to \mathbb{R}$, as defined in Definition B.4 with $N = 2$ transposed to the second axis, map the variable $\boldsymbol{V} \in \mathbb{R}^{(F+C) \times 2}$ to

$$s_t^{(F)}(\boldsymbol{V}) = \sum_{f=1}^{F} \boldsymbol{V}_{:,f}^{\boldsymbol{\alpha}^{(t)}} \quad \text{and} \quad s_t^{(C)}(\boldsymbol{V}) = \sum_{c=1}^{C} \boldsymbol{V}_{:,F+c}^{\boldsymbol{\alpha}^{(t)}},$$

where these polynomials are enumerated/indexed by $t \in [T]$ for $T = \binom{M_F+2}{2}$, in order to parameterize all multi-indices $\boldsymbol{\alpha}^{(t)} \in \mathbb{N}_0^2$ such that $\left\|\boldsymbol{\alpha}^{(t)}\right\|_1 \leq M_F$.

Fix $\boldsymbol{\beta} \in \mathbb{R}^2$ and denote $w_{\boldsymbol{\beta}}$ by w for brevity. By Lemma B.5, the $(S_F \times S_C)$-invariant polynomial w can be expressed as a polynomial $u : \mathbb{R}^{2T} \to \mathbb{R}$ in the doubly multi-symmetric power-sum polynomials (DMPs) $\{s_t^{(F)}\}_{t=1}^T : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}$ and $\{s_t^{(C)}\}_{t=1}^T : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}$ by

$$w(\boldsymbol{V}) = u\left(s_1^{(F)}(\boldsymbol{V}), \ldots, s_T^{(F)}(\boldsymbol{V}), s_1^{(C)}(\boldsymbol{V}), \ldots, s_T^{(C)}(\boldsymbol{V})\right). \tag{39}$$

We can write w as composition of the three maps

$$w = u \circ t \circ b, \tag{40}$$

where $b : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}^{(F+C)\times T}$, $t : \mathbb{R}^{(F+C)\times T} \to \mathbb{R}^{2T}$ and $u : \mathbb{R}^{2T} \to \mathbb{R}$.

Here, the map $b : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}^{(F+C)\times T}$ calculates all monomials of the DMPs. It does so by applying the signle function $\hat{b} : \mathbb{R}^2 \to \mathbb{R}^T$ row-wise, which calculates all monomials of the DMPs per row, i.e.,

$$\forall j \in [F+C]: \quad b(\boldsymbol{V})_{j,:} = \hat{b}(\boldsymbol{V}_{j,:}) := \left(\boldsymbol{V}_{j,:}^{\boldsymbol{\alpha}^{(1)}}, \ldots, \boldsymbol{V}_{j,:}^{\boldsymbol{\alpha}^{(T)}}\right) \tag{41}$$

The map $t : \mathbb{R}^{(F+C)\times T} \to \mathbb{R}^{2T}$ performs the relevant sums over F and C to derive the DMPs, when applied to $\boldsymbol{H} = b(\boldsymbol{X}) \in \mathbb{R}^{(F+C)\times T}$ like so

$$t(\boldsymbol{H}) = (\sum_{f=1}^{F} \boldsymbol{H}_{f,:}, \sum_{c=1}^{C} \boldsymbol{H}_{F+c,:}). \tag{42}$$

Lastly, $u : \mathbb{R}^{2T} \to \mathbb{R}$ is defined by (39).

Approximating $Q^{(F)}$ using TSNet*. Note that $\hat{b}$ belongs to the space $\mathcal{B}_1$ of all polynomials from $\mathbb{R}^2$ to $\mathbb{R}^T$. By the Universal Approximation Theorem (Theorem B.12), the space $\mathcal{N}_1$ of multilayer perceptrons (MLPs) from $\mathbb{R}^2$ to $\mathbb{R}^T$ is a universal approximator of $\mathcal{B}_1$.

The polynomial b applies $\hat{b}$ row-wise (see Equation (41)) and belongs to the space $\mathcal{B}_2$ of all polynomials $z : \mathbb{R}^{(F+C)\times 2} \to \mathbb{R}^{(F+C)\times T}$ of the form $z(\boldsymbol{V}) = (y(\boldsymbol{V}_{1,:}), \ldots, y(\boldsymbol{V}_{F+C,:}))^\top$ for arbitrary polynomials $y \in \mathcal{B}_1$. Consider such a polynomial z and corresponding y. For any $\epsilon > 0$ and compact domain $C \subset \mathbb{R}^{(F+C)\times 2}$, the domain C in contained in a larger compact domain of the form $C' \times \ldots \times C'$ where $C' \subset \mathbb{R}^2$ is compact. Hence, by Theorem B.12 $q : C' \to \mathbb{R}^T$ can be approximated in infinity norm by an MLP $\theta \in \mathcal{N}_1$ up to error ϵ. Hence, we also have that $\eta(\boldsymbol{V}) := (\theta(\boldsymbol{V}_{1,:}), \ldots, \theta(\boldsymbol{V}_{F+C,:}))^\top$ approximates z up to error ϵ in infinity norm. Such an η belongs to the space $\mathcal{N}_2$ of networks from $\mathbb{R}^{(F+C)\times 2}$ to $\mathbb{R}^{(F+C)\times 2}$ of the form $\hat{\eta}(\boldsymbol{V}) := (\hat{\theta}(\boldsymbol{V}_{1,:}), \ldots, \hat{\theta}(\boldsymbol{V}_{F+C,:}))^\top$, where $\hat{\theta} \in \mathcal{N}_1$. This shows that $\mathcal{N}_2$ is a universal approximator of $\mathcal{B}_2$.

The transformation t belongs to the space $\mathcal{B}_3$ of linear $(S_F \times S_C)$-equivariant functions from $\mathbb{R}^{(F+C)\times T}$ to $\mathbb{R}^{2T}$. Specifically, it corresponds to the layers next to coefficient matrices $\boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_6, \boldsymbol{\Lambda}_8, \boldsymbol{\Lambda}_{12}$ in Equation (3). Consider the space $\mathcal{N}_3$ of linear MLPs consisting of the $(S_F \times S_C)$-equivariant linear layers from $\mathbb{R}^{(F+C)\times T}$ to $\mathbb{R}^{2T}$, Specifically, the layers which correspond to the coefficient matrices $\boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_6, \boldsymbol{\Lambda}_8, \boldsymbol{\Lambda}_{12}$ in Equation (3). By Proposition 4.1, any function from $\mathcal{B}_3$ can be implemented as a function from $\mathcal{N}_3$, so $\mathcal{N}_3$ is trivially a universal approximator of $\mathcal{B}_3$.

The polynomial u belongs to the space $\mathcal{B}_4$ of all polynomials from $\mathbb{R}^{2T}$ to $\mathbb{R}$. By the Universal Approximation Theorem (Theorem B.12) the space $\mathcal{N}_4$ of multilayer perceptrons (MLPs) from $\mathbb{R}^{2T}$ to $\mathbb{R}$ is a universal approximator of $\mathcal{B}_4$.

Denote

$$\mathcal{N}_w := \left\{\theta_4 \circ \theta_3 \circ \theta_2 \mid \theta_2 \in \mathcal{N}_2, \theta_3 \in \mathcal{N}_3, \theta_4 \in \mathcal{N}_4\right\},$$

which contains the polynomial $w_{\boldsymbol{\beta}}$ (see Equation (40)). By Theorem B.14, the function space $\mathcal{N}_w$ is a universal approximator for the class of

$$\mathcal{B}_w := \left\{f_4 \circ f_3 \circ f_2 \mid f_2 \in \mathcal{B}_2, f_3 \in \mathcal{B}_3, f_4 \in \mathcal{B}_4\right\}.$$

Next, we construct a space $\mathcal{B}_F$ of polynomials from $\mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ to $\mathbb{R}^{N\times L_F}$, and show that $Q^{(F)} \in \mathcal{B}_F$. A general function from the space $\mathcal{B}_F$ is defined constructively as follows. Given an input $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$, we first compute its row-wise sum $\sum_{i=1}^{N} \boldsymbol{X}_{i,:}$ and append it to a new dimension of $\boldsymbol{X}$, creating a tensor $\boldsymbol{V} \in \mathbb{R}^{N\times(F+C)\times 2}$ via

$$\forall n \in [N]: \quad \boldsymbol{V}_{n,:,:} = [\boldsymbol{X}_{n,:}, \sum_{i=1}^{N} \boldsymbol{X}_{i,:}].$$

Next, for some $Y \in \mathbb{N}$ (which can be any value from $\mathbb{N}$), we define for each multi-index $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_1 < Y$, a polynomial $R_{\boldsymbol{\beta}} \in \mathcal{B}_w$ ($R_{\boldsymbol{\beta}}$ can be any function from $\mathcal{B}_w$). We then apply $R_{\boldsymbol{\beta}}$ row-wise on each $\boldsymbol{V}_{n,:,:}$, and multiply it by $\boldsymbol{V}^{\boldsymbol{\beta}}_{n,j,:}$, to create the following array in $\mathbb{R}^{N\times L_F}$

$$\Big\{ \sum_{\|\boldsymbol{\beta}\|_1 \leq Y} \boldsymbol{V}^{\boldsymbol{\beta}}_{n,j,:} R_{\boldsymbol{\beta}}(\boldsymbol{V}_{n,:,:}) \Big\}_{i\in[N], j\in[F+C]}.$$

Note that by Equations (22) and (23) and 38)–(42), $Q^{(F)} \in \mathcal{B}_F$.

We now define a corresponding space of MLPs from $\mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ to $\mathbb{R}^{N\times L_F}$, that we denote by $\mathcal{N}_F$, as follows. Given an input $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$, first compute its row-wise sum $\sum_{i=1}^{N} \boldsymbol{X}_{i,:}$ and append it to a new dimension of $\boldsymbol{X}$, creating a tensor $\boldsymbol{V} \in \mathbb{R}^{N\times(F+C)\times 2}$ like so

$$\forall n \in [N]: \quad \boldsymbol{V}_{n,:,:} = [\boldsymbol{X}_{n,:}, \sum_{i=1}^{N} \boldsymbol{X}_{i,:}].$$

This operation can be can be implemented using the general form of $(S_N \times S_F \times S_C)$-equivariant linear maps. Specifically, it corresponds to a choice of coefficient matrices $\boldsymbol{\Lambda}_3, \boldsymbol{\Lambda}_4, \boldsymbol{\Lambda}_9, \boldsymbol{\Lambda}_{10}$ in Equation (3). Next, we apply on $\boldsymbol{V}_{n,j,:}$, for each $n \in [N]$ and $j \in [F+C]$, any sequence of general MLPs $\theta_l : \mathbb{R}^2 \to \mathbb{R}$, $l \in [L]$ for some $L \in \mathbb{N}$. Then, apply the same general MLP $\zeta \in \mathcal{N}_w$ on $\boldsymbol{V}_{n,:,:}$ for all $n \in [N]$. Then apply any general MLP $\kappa : \mathbb{R}^2 \to \mathbb{R}$ on each $(\theta_l(\boldsymbol{V}_{n,l,:}), \zeta(\boldsymbol{V}_{n,:,:}))$, and sum, to obtain

$$\Big\{ \sum_{l=1}^{L} \kappa(\theta_l(\boldsymbol{V}_{n,j,:}), \zeta(\boldsymbol{V}_{n,:,:})) \Big\}_{j\in[F+C]}.$$

Since we can approximate the products of two numbers by an MLP $\kappa : \mathbb{R}^2 \to \mathbb{R}$, since we can approximate monomials $\boldsymbol{V}^{\boldsymbol{\beta}}_{n,j,:}$ by MLPs $\theta_l : \mathbb{R}^2 \to \mathbb{R}$, and since we can approximate polynomials from $\mathcal{B}_w$ by MLPs from $\mathcal{N}_w$, the above space $\mathcal{N}_F$ is a universal approximator of $\mathcal{B}_F$. Moreover, by construction, the space $\mathcal{N}_F$ consists only of TSNets*.

Approximating $Q^{(NF)}$ using TSNet*. Similarly to the function space $\mathcal{B}_w$ of $(S_F \times S_C)$-invariant polynomials. We can construct a the function space $\hat{\mathcal{B}}_w$ of S_N-invariant polynomials from $\mathbb{R}^{N\times 3}$ to $\mathbb{R}$. By the same reasoning, similarly to the function space $\mathcal{N}_w$ of $(S_F \times S_C)$-invariant MLPs, we can also construct a the function space $\hat{\mathcal{N}}_w$ of S_N-invariant MLPS from $\mathbb{R}^{N\times 3}$ to $\mathbb{R}$.

Next, we construct a space $\mathcal{B}_N$ of polynomials from $\mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ to $\mathbb{R}^{L_n\times L_F}$, and show that $Q^{(NF)} \in \mathcal{B}_N$. A general function from the space $\mathcal{B}_N$ is defined constructively as follows. Given an input $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$, we first compute $T(\boldsymbol{X})$, where T can be any function from $\mathcal{B}_F$. We also compute its column-wise sums $\sum_{f=1}^{F} \boldsymbol{X}_{:,f}$ and $\sum_{c=1}^{C} \boldsymbol{X}_{:,F+c}$ and concatenate the sums to the result of $T(\boldsymbol{X})$ over a new tensors to create a tensor $\boldsymbol{Z} \in \mathbb{R}^{N\times L_F\times 3}$ via

$$\forall j \in [F+C]: \quad \boldsymbol{Z}_{:,j,:} = [T(\boldsymbol{X})_{:,j}, \sum_{f=1}^{F} \boldsymbol{X}_{:,f}, \sum_{c=1}^{C} \boldsymbol{X}_{:,F+c}].$$

Next, for some $Y \in \mathbb{N}$ (which can be any value from $\mathbb{N}$), we define for each multi-index $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_1 < Y$, a polynomial $\hat{R}_{\boldsymbol{\beta}} \in \hat{\mathcal{B}}_w$ ($\hat{R}_{\boldsymbol{\beta}}$ can be any function from $\hat{\mathcal{B}}_w$). We then apply $\hat{R}_{\boldsymbol{\beta}}$ column-wise on each $\boldsymbol{Z}_{:,j,:}$, and multiply it by $\boldsymbol{V}^{\boldsymbol{\beta}}_{n,j,:}$, to create the following array in $\mathbb{R}^{L_N\times L_F}$

$$\Big\{ \sum_{\|\boldsymbol{\beta}\|_1 \leq Y} \boldsymbol{Z}^{\boldsymbol{\beta}}_{n,j,:} \hat{R}_{\boldsymbol{\beta}}(\boldsymbol{Z}_{:,j,:}) \Big\}_{n\in[N], j\in[L_F]}.$$

Note that by Equations (31) and (32), $Q^{(NF)} \in \mathcal{B}_N$.

We now define a corresponding space of MLPs from $\mathbb{R}^{N\times(F+C)} \setminus \mathcal{E}$ to $\mathbb{R}^{L_N \times L_F}$, that we denote by $\mathcal{N}_N$, as follows. Given an input $\boldsymbol{X} \in \mathbb{R}^{N\times(F+C)}$, first compute $\hat{T}(\boldsymbol{X})$, where T can be any function from $\mathcal{N}_F$. We also compute its column-wise sums $\sum_{f=1}^{F} \boldsymbol{X}_{:,f}$ and $\sum_{c=1}^{C} \boldsymbol{X}_{:,F+c}$ and concatenate the sums to the result of $\hat{T}(\boldsymbol{X})$ over a new tensors to create a tensor $\boldsymbol{Z} \in \mathbb{R}^{N\times L_F \times 3}$ like so

$$\forall j \in [F+C]: \quad \boldsymbol{Z}_{:,j,:} = [T(\boldsymbol{X})_{:,j}, \sum_{f=1}^{F} \boldsymbol{X}_{:,f}, \sum_{c=1}^{C} \boldsymbol{X}_{:,F+c}].$$

This operation can be can be implemented using the general form of $(S_N \times S_F \times S_C)$-equivariant linear maps. Specifically, it corresponds to a choice of coefficient matrices $\boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_4, \boldsymbol{\Lambda}_6, \boldsymbol{\Lambda}_8, \boldsymbol{\Lambda}_{10}, \boldsymbol{\Lambda}_{12}$ in Equation (3). Next, we apply on $\boldsymbol{Z}_{n,j,:}$, for each $n \in [N]$ and $j \in [L_F]$, any sequence of general MLPs $\hat{\theta}_l : \mathbb{R}^3 \to \mathbb{R}$, $l \in [L]$ for some $L \in \mathbb{N}$. Then, apply the same general MLP $\hat{\zeta} \in \hat{\mathcal{N}_w}$ on $\boldsymbol{Z}_{:,j,:}$ for all $j \in [L_F]$. Then apply any general MLP $\hat{\kappa} : \mathbb{R}^2 \to \mathbb{R}$ on each $(\hat{\theta}_l(\boldsymbol{V}_{n,j,:}), \hat{\zeta}(\boldsymbol{V}_{:,j,:}))$, and sum, to obtain

$$\Big\{ \sum_{l=1}^{L} \hat{\kappa}(\hat{\theta}_l(\boldsymbol{Z}_{n,j,:}), \hat{\zeta}(\boldsymbol{Z}_{:,j,:})) \Big\}_{n\in[N], j\in[L_F]}.$$

Since we can approximate the products of two numbers by an MLP $\hat{\kappa} : \mathbb{R}^2 \to \mathbb{R}$, since we can approximate monomials $\boldsymbol{Z}_{n,j,:}^{\boldsymbol{\beta}}$ by MLPs $\hat{\theta}_l : \mathbb{R}^3 \to \mathbb{R}$, and since we can approximate polynomials from $\hat{\mathcal{B}}_w$ by MLPs from $\hat{\mathcal{N}_w}$, the above space $\mathcal{N}_N$ is a universal approximator of $\mathcal{B}_N$. Moreover, by construction, the space $\mathcal{N}_N$ consists only of TSNets*.

Approximating $p_{1,1}$ using our architecture. Let $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{N\times(F+C)}$. Define the equivalence relation $\boldsymbol{Y} \sim_{NF} \boldsymbol{Z}$ if there exists $(\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \in S_{N-1} \times S_{F-1} \times S_C$ such that $\boldsymbol{Y} = (\sigma_{N-1}, \sigma_{F-1}, \sigma_C) \cdot \boldsymbol{Z}$. The quotient space $\mathbb{R}^{N\times(F+C)}/(S_{N-1} \times S_{F-1} \times S_C)$ is the set of equivalence classes under this relation, equipped with the quotient topology, i.e., the finest topology that makes the canonical projection

$$\widetilde{\pi} : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times D}/(S_{N-1} \times S_{F-1} \times S_C), \quad \widetilde{\pi}(\boldsymbol{Y}) = [\boldsymbol{Y}]_{NF}$$

continuous, where $[\boldsymbol{Y}]_{NF}$ denotes the equivalence class of $\boldsymbol{Y}$.

Let $Q : \mathcal{K} \to \mathbb{R}^{L_F L_N}$ be the invariant map defined earlier, and for brevity write $Q = Q^{(NF)}$. Since both Q and $p_{1,1}\mathcal{K} \to \mathbb{R}$ are $(S_{N-1} \times S_{F-1} \times S_C)$-invariant, they factor uniquely through the quotient space. That is, there exist unique continuous maps

$$\widetilde{p_{1,1}} : \widetilde{\pi}(\mathcal{K}) \to p_{1,1}(\mathcal{K}) \quad \text{and} \quad \widetilde{Q} : \widetilde{\pi}(\mathcal{K}) \to Q(\mathcal{K}),$$

such that

$$p_{1,1} = \widetilde{p_{1,1}} \circ \widetilde{\pi} \text{ and } Q = \widetilde{Q} \circ \widetilde{\pi}.$$

Since the canonical projection $\widetilde{\pi}$ is continuous by definition, and p_{11} is continuous, by Lemma B.11 $\widetilde{p}_{1,1}$ must also be continuous. Now as $\mathcal{K}$ is compact, $\widetilde{\pi}(\mathcal{K})$ is also compact. Moreover, $\widetilde{Q}$ is a bijection on its domain and co-domain by Equation (33). Thus, we can now write $p_{1,1}$ as

$$p_{1,1} = (\widetilde{p_{1,1}} \circ \widetilde{Q}^{-1}) \circ \widetilde{Q} \circ \widetilde{\pi} = (\widetilde{p_{1,1}} \circ \widetilde{Q}^{-1}) \circ Q.$$

Denote $\widetilde{p_{1,1}} \circ \widetilde{Q}^{-1}$ by $r : Q(\mathcal{K}) \to p_{1,1}(\mathcal{K})$. Since $\mathcal{K}$ is compact, both $Q(\mathcal{K})$ and $p_{1,1}(\mathcal{K})$ are compact as well. By Lemma B.11, the function r is continuous.

Note that r belongs to the space $\mathcal{B}_r$ of all polynomials from $Q(\mathcal{K})$ to $p_{1,1}(\mathcal{K})$. By the Universal Approximation Theorem (Theorem B.12), the space $\mathcal{N}_r$ of multilayer perceptrons (MLPs) from $Q(\mathcal{K})$ to $p_{1,1}(\mathcal{K})$ is a universal approximator of $\mathcal{B}_r$.

Denote

$$\mathcal{N} := \{\theta_r \circ \theta_N \mid \theta_N \in \mathcal{N}_N, \theta_r \in \mathcal{N}_r\},$$

which contains the polynomial $p_{1,1}$. By Theorem B.14, the function space $\mathcal{N}$ is a universal approximator for the class of

$$\mathcal{N} := \{f_r \circ f_N \mid f_N \in \mathcal{B}_N, \theta_r \in \mathcal{B}_r\}.$$

Recall that $p_{1,F+1}$ is an $(S_{N-1} \times S_F \times S_{C-1})$-invariant function. By the same reasoning, we can define a function similar to $q^{(F)}$ which is not $(S_{F-1} \times S_C)$-invariant but $(S_F \times S_{C-1})$-invariant instead. We can then follow up by approximating it using Lemma B.7 instead of Lemma B.6, such that by the same reasoning we can also approximate $p_{1,F+1}$ using TSNet*.

Thus, there exists a TSNet* in $\mathcal{B}$, denoted by $f : \mathcal{K} \to \mathbb{R}^{N\times(F+C)}$, such that

$$\sup_{\boldsymbol{X}\in\mathcal{K}} \|p(\boldsymbol{X}) - f(\boldsymbol{X})\|_\infty < \frac{\epsilon}{2}.$$

Then, Equation (18) and the triangle inequality, we obtain

$$\sup_{\boldsymbol{X}\in\mathcal{K}} \|q(\boldsymbol{X}) - f(\boldsymbol{X})\|_\infty < \epsilon,$$

which completes the proof. □

Note. By applying same techniques in the proof of Lemma B.16, one can correctly prove Theorem 3 of Maron et al. [2020].

Theorem 4.2. *Let $K \subset \mathbb{R}^{N\times(F+C)}$ be a compact domain such that $\mathcal{K} = \cup_{g\in S_N\times S_F\times S_C} gK$ and $\mathcal{K}\cap\mathcal{E} = \emptyset$, where $\mathcal{E} \subset \mathbb{R}^{N\times(F+C)}$ is the exclusion set corresponding to $\mathbb{R}^{N\times(F+C)}$ (Definition B.15). Then, TSNets are universal approximators in L_∞ of continuous $\mathcal{K} \to \mathbb{R}^{N\times C}$ functions that are $(S_N \times S_C)$-equivariant and S_F-invariant.*

Proof. Let $q : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times C}$ be a continuous $S_N \times S_C$-equivariant and S_F-invariant function on K. We define the lifted function $q^* : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times(F+C)}$ by

$$q^*(\boldsymbol{X}) = [\mathbf{0}^{N\times F} \parallel q(X)] \in \mathbb{R}^{N\times(F+C)},$$

where we pad $q(\boldsymbol{X})$ with zeros in the first F columns. It is easy to verify that q^* is $(S_F \times S_C \times S_N)$-equivariant, as the S_F part of the group action acts only on the zero-padded section, and thus does not affect $q(\boldsymbol{X})$.

By Lemma B.16, for any $\varepsilon > 0$, there exists a network TSNet*, denoted by $f^*_\varepsilon : \mathbb{R}^{N\times(F+C)} \to \mathbb{R}^{N\times(F+C)}$ such that

$$\|q^*(\boldsymbol{X}) - f^*_\varepsilon(\boldsymbol{X})\|_\infty < \varepsilon \quad \text{for all} \quad \boldsymbol{X} \in \mathcal{K}.$$

Consider the label projection map π, which satisfies

$$\pi(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y}, \quad \forall(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{N\times F} \times \mathbb{R}^{N\times C}.$$

Notice that $q := \pi \circ q^*$ and that $\pi \circ f^*_\varepsilon$ is a TSNet architecture, denoted by $f_\varepsilon := \pi \circ f{*}_\varepsilon$.

Thus, we get

$$\|q(\boldsymbol{X}) - f_\varepsilon(\boldsymbol{X})\|_\infty = \|\pi\left(q^*(\boldsymbol{X}) - f^*_\varepsilon(\boldsymbol{X})\right)\|_\infty \le \|q^*(\boldsymbol{X}) - f^*_\varepsilon(\boldsymbol{X})\|_\infty < \varepsilon,$$

for all $\boldsymbol{X} \in \mathcal{K}$. Hence, there exists a TSNet that approximates q on K. □

C Further Details on the Experiments

C.1 Dataset Statistics

The statistics of the 28 node classification datasets used in Table 1 can be found in Table 2.

C.2 Hyperparameters for all Experiments

We adopt the hyperparameters used in Zhao et al. [2025] for the experiments in Table 1 over the architectures MeanGNN, GAT and GraphAny. The full set of hyperparameters is reported in Table 3.

Table 2: Statistics of the 28 node node classification datasets.

Dataset	#Nodes	#Edges	#Feature	#Classes	Train/Val/Test Ratios (%)
actor	7600	30019	932	5	48.0/32.0/20.0
amazon-ratings	24492	186100	300	5	50.0/25.0/25.0
Arxiv	169343	1166243	128	40	53.7/17.6/28.7
blogcatalog	5196	343486	8189	6	2.3/48.8/48.8
brazil	131	1074	131	4	61.1/19.1/19.8
chameleon	2277	36101	2325	5	48.0/32.0/20.0
citeseer	3327	9104	3703	6	3.6/15.0/30.1
co-cs	18333	163788	6805	15	1.6/49.2/49.2
co-physics	34493	495924	8415	5	0.3/49.9/49.9
computers	13752	491722	767	10	1.5/49.3/49.3
cora	2708	10556	1433	7	5.2/18.5/36.9
cornell	183	554	1703	5	47.5/32.2/20.2
deezer	28281	185504	128	2	0.1/49.9/49.9
europe	399	5995	399	4	20.1/39.8/40.1
full-DBLP	17716	105734	1639	4	0.5/49.8/49.8
full-cora	19793	126842	8710	70	7.1/46.5/46.5
last-fm-asia	7624	55612	128	18	4.7/47.6/47.6
minesweeper	10000	78804	7	2	50.0/25.0/25.0
photo	7650	238162	745	8	2.1/49.0/49.0
pubmed	19717	88648	500	3	0.3/2.5/5.1
questions	48921	307080	301	2	50.0/25.0/25.0
roman-empire	22662	65854	300	18	50.0/25.0/25.0
squirrel	5201	217073	2089	5	48.0/32.0/20.0
texas	183	558	1703	5	47.5/31.7/20.2
tolokers	11758	1038000	10	2	50.0/25.0/25.0
usa	1190	13599	1190	4	6.7/46.6/46.6
wiki	2405	17981	4973	17	14.1/42.9/43.0
wiki-cs	11701	431206	300	10	5.0/15.1/49.9

Table 3: Hyperparameters used in Table 1.

	MeanGNN	GAT	GraphAny	TS-Mean	TS-GAT
lr	$2 \cdot 10^{-4}$ $5 \cdot 10^{-4}$ 10^{-3}	$2 \cdot 10^{-4}$ $5 \cdot 10^{-4}$ 10^{-3}	$2 \cdot 10^{-4}$	0.01	0.01 0.03
hidden dimension	64, 128	-	64	16	16
# layers	1, 2, 3	-	-	2	2
# batch	-	-	500, 1000	5, 10	5, 10
visible train labels (%)	-	-	0.5	0.3, 0.4, 0.5	0.1, 0.2
# epochs	400	400	400	2000	2000
Entropy	-	-	1, 2	-	-
# MLP layers	-	-	1	-	-