# Improved Generalization Bounds of Group Invariant / Equivariant Deep Networks via Quotient Feature Spaces (Supplementary material)

**Akiyoshi Sannai**[1,2] **Masaaki Imaizumi**[3,1] **Makoto Kawano**[3]

[1]RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo, Japan
[2]Keio University, Minato, Tokyo, Japan
[3] The University of Tokyo, Bunkyo, Tokyo, Japan

## A SUPPLEMENT

We provide the deferred proofs of each section.

### A.1 PROOF FOR SECTION 3

*Proof of Lemma 1.* For any $\epsilon > 0$, there is a sequence of rational numbers $\{p_i/q_i\}$ such that $p_i/q_i < \epsilon$ and converges to $\epsilon$. Assume that Lemma 1 holds for rational numbers, then we have$\mathcal{N}_{\varepsilon,\infty}(\Delta_{S_n}) \leq \mathcal{N}_{p_i/q_i,\infty}(\Delta_{S_n}) \leq C/(n!\ (p_i/q_i)^n)$. Since $1/x^n$ is a continuous function and $\{p_i/q_i\}$ converges to $\epsilon$, we obtain $\mathcal{N}_{\varepsilon,\infty}(\Delta_{S_n}) \leq C/(n!\ (\varepsilon)^n)$. Hence it is enough to show the case of rational numbers.

We assume $\varepsilon = p/q$ for some integers $p, q > 0$. Let $\mathcal{C}(I)$ be the covering of $I$, which is a set of $\varepsilon$-cubes

$$c_{j_1,..,j_n} = \{x = (x_i) \in I \mid \varepsilon j_i \leq x_i \leq \varepsilon(j_i + 1)\},$$

for $j_i = 1, .., [q/p] + 1$. We can easily see that $\mathcal{C}(I)$ attains the minimum number of $\varepsilon$-cubes covering $I$ and the number is $(\varepsilon^{-1} + 1)^n = \frac{\varepsilon^{-n}}{n!} + \mathcal{O}(\varepsilon^{-(n-1)})$. We show that we can find a subset of $\mathcal{C}(I)$ which cover $\Delta_{S_n}$ and whose cardinality is $\frac{\varepsilon^{-n}}{n!} + \mathcal{O}(\varepsilon^{-(n-1)})$. The proof is as follows. At first, we calculate the number $A$ of cubes in $\mathcal{C}(I)$ which intersect with the boundary of $\sigma \cdot \Delta$. Then since the number of the orbit of the cubes which do not intersect with the boundary of $\sigma \cdot \Delta$ is $n!$, if $A$ is $\mathcal{O}(\varepsilon^{-(n-1)})$, we can find the covering whose cardinality is $\frac{\varepsilon^{-n}}{n!} + \mathcal{O}(\varepsilon^{-(n-1)})$. Since $\sigma \cdot \Delta$ is $\left\{x \in I \mid x_{\sigma^{-1}(1)} \geq x_{\sigma^{-1}(2)} \geq \cdots \geq x_{\sigma^{-1}(n)}\right\}$, any boundary of $\sigma \cdot \Delta$ is of the form $\left\{x \in I \mid x_{\sigma^{-1}(1)} \geq \cdots x_{\sigma^{-1}(i)} = x_{\sigma^{-1}(i+1)} \geq \cdots \geq x_{\sigma^{-1}(n)}\right\}$.

From here, we fix $\sigma$ and $i$. Consider the canonical projection $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ which sends $x_{\sigma^{-1}(i)}$-axis to zero. $\pi$ induces the map $\widetilde{\pi} : \mathcal{C}(I) \to \mathcal{C}(\pi(I))$, where $\mathcal{C}(\pi(I))$ is the covering of $\pi(I)$ . Let $\mathcal{C}(I)_B$ denote the subset of cubes in $\mathcal{C}(I)$ which intersect with the set $B = \left\{x \in I \mid x_{\sigma^{-1}(i)} = x_{\sigma^{-1}(i+1)}\right\}$ . Then we can see that $\widetilde{\pi}$ is injective on $\mathcal{C}(I)_B$ as follows. Assume that there are two cubes in $\mathcal{C}(I)_B$ whose images by $\widetilde{\pi}$ are equal. Let us denote the centers of two cubes by $c_{j_1,..,j_n}$ and $c_{k_1,..,k_n}$ .Then, since $\pi$ only kills $x_{\sigma^{-1}(i)}$, $j_p = k_p$ holds for $p \neq \sigma^{-1}(i)$. But since $c_{j_1,..,j_n}$ and $c_{k_1,..,k_n}$ are in $\mathcal{C}(I)_B$, we have $j_{\sigma^{-1}(i)} = j_{\sigma^{-1}(i+1)}$ and $k_{\sigma^{-1}(i)} = k_{\sigma^{-1}(i+1)}$. Hence $j_p = k_p$ for any $p$ and $\widetilde{\pi}$ is injective on $\mathcal{C}(I)_B$.

Next, let $\mathcal{C}(I)_{\widetilde{B}}$ be the subset of $\varepsilon$-cubes in $\mathcal{C}(I)$ which intersect a boundary of $\sigma \cdot \Delta$ for some $\sigma$. We see that the cardinality of $\mathcal{C}(I)_{\widetilde{B}}$ is bounded by $e\varepsilon^{-(n-1)}$ for some constant $e > 0$. Since the number of components of the boundaries is finite, we prove the claim for a component $B$. As we see before, $\widetilde{\pi}_{\restriction \mathcal{C}(I)_B}$ is injective. This result implies that the number of cubes that intersect $B$ is bounded by a number of $\varepsilon$-cubes in $\mathcal{C}(p(I)) = \varepsilon^{-(n-1)}$. Put $\mathcal{C}(I)_F = \mathcal{C}(I) - \mathcal{C}(I)_{\widetilde{B}}$. Then we note that the action of $S_n$ on$\mathcal{C}(I)_F$ is free, namely the number of the orbit of any cube in $\mathcal{C}(I)_F$ is $|S_n|$. Hence,

$$|\mathcal{C}(I)_F \cap \Delta| = |\{c \in \mathcal{C}(i) \mid c \subset \Delta\}| = 1/|S_n||\mathcal{C}(I)_F| \leq \varepsilon^{-n}/|S_n|.$$

Here, $(\mathcal{C}(I)_F \cap \Delta) \cup \mathcal{C}(I)_{\widetilde{B}}$ gives the covering of $\Delta$. This covering gives

$$\mathcal{N}_{\varepsilon,\infty}(\Delta) \leq \frac{\varepsilon^{-n}}{n!} + e\varepsilon^{-(n-1)}.$$

$\square$

*Proof of Lemma 2.* By Proposition 3, we have $\widetilde{\Delta}_G$ satisfying two conditions above. Since the covering of $\widetilde{\Delta}_G$ induces the covering of $\Delta_G$ by the condition 2, $\mathcal{N}_{\varepsilon,\infty}(\Delta_G) \leq \mathcal{N}_{\varepsilon,\infty}(\widetilde{\Delta}_G)$. On the other hand, by the condition 1, we have $\mathcal{N}_{\varepsilon,\infty}(\widetilde{\Delta}_G) \leq |S_n|/|G| \cdot \mathcal{N}_{\varepsilon,\infty}(\Delta_{S_n})$. Combining with Lemma 1, we have the desired result. $\square$

*Proof of Proposition 1.* For the claim, assume that an action of $G$ preserves distance, namely, $\|x - x'\|_2 = \|g(x) - g(x')\|_2$ holds. We show that $d_G(y, y') = \inf_{x,x' \in \mathbb{R}^n}\{\|x - x'\|_2 | \phi_G(x) = y, \phi_G(x') = y'\}$. Consider the sum $\|x - b_1\|_2 + \|a_2 - b_2\|_2 + ... + \|a_n - x'\|_2$ and take an element $g \in G$ such that $a_2 = g \cdot b_1$. Then, $\|x - b_1\|_2 + \|a_2 - b_2\|_2 = \|x - b_1\|_2 + \|g \cdot b_1 - b_2\|_2 = \|x - b_1\|_2 + \|b_1 - g^{-1} \cdot b_2\|_2 \geq \|x - g^{-1} \cdot b_2\|_2$. By repeating this process, we have $\|x - b_1\|_2 + \|a_2 - b_2\|_2 + ... + \|a_n - x'\|_2 \geq \|x - g \cdot x'\|_2$ for some $g \in G$. Hence, $d_G(y, y') = \inf_{x,x' \in I}\{\|x - x'\|_2 | \phi_G(x) = y, \phi_G(x') = y'\}$. This implies $d_G(y, y') = 0 \Rightarrow y = y'$. $\square$

*Proof of Proposition 3.* We confirm that $\widetilde{\Delta}_G$ satisfies both the conditions 1 and 2. As the action of $G$ preserves the distance, $g_k : \Delta_{S_n} \to g_k \cdot \Delta_{S_n}$ is an isomorphism on metric spaces. Hence, condition 1 is satisfied.

For condition 2, we consider $y \in \Delta_G = \phi_G(I)$. Then, there is an element $x \in I$ such that $y = \phi_G(x)$. As $I = \cup_{\sigma \in S_n} \sigma \cdot \Delta_{S_n}$, there exist $\sigma \in S_n$ and $z \in \Delta_{S_n}$ such that $x = \sigma \cdot z$.

In contrast, as $\{g_1, .., g_K | g_k \in G\}$ is a complete system of representatives of $G \backslash S_n$, there exist $\tau \in G$ and $g_k$ such as $\sigma = \tau \cdot g_i$. Then $\phi_G(g_k z) = \phi_G(\tau \cdot g_k z) = \phi_G(x) = y$ and $g_k \cdot z \in \widetilde{\Delta}_G$. Hence $\phi_G(\widetilde{\Delta}_G) = \Delta_G$. $\square$

### A.2 PROOF FOR SECTION 3.2

*Proof of Proposition 4.* We prove $\widehat{\phi}_G$ is injective and surjective. Assume $f \in C(\Delta_G)$ and put $\widehat{\phi}_G(f) = f \circ \phi_G$. Then since $\phi_G$ is $G$-invariant, so is $\widehat{\phi}_G(f)$. Also, since $\phi_G$ is surjective, $\widehat{\phi}_G$ is injective. Take $g \in C^G(I)$, then we define $f \in C(\Delta_G)$ as follows; for any $y \in \Delta_G$, take $x \in I$ such that $\phi_G(x) = y$ and define $f(y) = g(x)$. This map is well defined because $g$ is $G$-invariant. $\widehat{\phi}_G(f)(x) = f \circ \phi_G(x) = f(y) = g(x)$. Hence, we obtain the desired result.
Next, we prove the Lipschitz properties. Take $f \in C(\Delta_G)$ and assume $f$ is $K$-Lipschitz. Then for any $x, x' \in I$,

$$d_G(\phi_G(x), \phi_G(x')) \geq Kd(f(\phi_G(x)), f(\phi_G(x'))),$$

by $K$-Lipschitz property of $f$. By the definition of $d_G$, we have $d_G(\phi_G(x), \phi_G(x')) \leq d(x, x')$. Hence, $\widehat{\phi}_G(f)$ is $K$-Lipschitz continuous. Conversely, assume $\widehat{\phi}_G(f)$ is $K$-Lipschitz. Take any $y, y' \in I$, then for any $x, x' \in I$ satisfying $\phi_G(x) = y, \phi_G(x) = y'$,

$$d(x, x') \geq Kd(f(\phi_G(x)), f(\phi_G(x'))) = d(f(y), f(y')),$$

by $K$-Lipschitz property of $\widehat{\phi}_G(f)$. Hence by taking infimum of the left hand side, we have

$$d_G(y, y') = \inf d_G(\phi_G(x), \phi_G(x')) \geq Kd(f(y), f(y')).$$

Hence, $f$ is $K$-Lipschitz. $\square$

*Proof of Proposition 2.* We first note that $\phi$ is the identity map on $\Delta$, because elements in $\Delta$ are sorted. This implies $\Delta \cong \phi(\Delta)$. Therefore, it is sufficient to show $\Delta \cong \Delta_{S_n}$. As $\Delta$ is a subset of $I$, we have the distance preserving map $\phi_{S_n \restriction_\Delta} : \Delta \to \Delta_{S_n}$.

Then, we show that $\phi_{S_n \restriction_\Delta}$ is a bijection. *Injectivity*: Ley us take any $x, y \in \Delta$ such that $\phi_{S_n \restriction_\Delta}(x) = \phi_{S_n \restriction_\Delta}(y)$. Then $x = g \cdot y$ for some $g \in S_n$. However, as $y$ is in $\Delta$, $\{g \cdot y | g \in S_n\} \cap \Delta = \{y\}$. Hence, $x = y$. *Surjectivity*: Take any $z \in \Delta_{S_n}$, then there is $x \in I$ such that $z = \phi_{S_n}(x)$. By the construction of $\Delta$, there is $g \in S_n$ and $y \in \Delta$ that satisfies $x = g \cdot y$. Hence, $z = \phi_{S_n}(x) = \phi_{S_n}(g \cdot y) = \phi_{S_n}(y)$. $\square$

*Proof of Proposition 5.* Firstly, we show $\widehat{\phi}_G^{-1}(f) \in \mathcal{F}(I)$ with any $f \in \mathcal{F}^G(I)$. For $f \in \mathcal{F}^G(I)$, we consider $\widehat{\phi}_G^{-1}(f) \in C(\Delta_G)$ as Proposition 4. Suppose $f$ and $f'$ are $K$-Lipschitz continuous, then $\widehat{\phi}_G^{-1}(f)$ is also $K$-Lipschitz continuous by Proposition 4. Since Zhang et al. [2018] states that Lipschitz continuous functions are represented by DNNs, we have $\widehat{\phi}_G^{-1}(f) \in \mathcal{F}(\Delta_G)$.

Fix $f_1, f_2 \in \mathcal{F}^G(I)$. Then, there exist $f_1', f_2' \in \mathcal{F}(\Delta_G)$ such as $f_1 = \widehat{\phi}_G(f_1')$ and $f_2 = \widehat{\phi}_G(f_2')$. Then, we have

$$\|f_1 - f_2\|_{L^\infty(I)} = \|\widehat{\phi}_G(f_1') - \widehat{\phi}_G(f_1')\|_{L^\infty(I)} = \|f_1' \circ \phi_G - f_2' \circ \phi_G\|_{L^\infty(I)} \leq \|f_1' - f_2'\|_{L^\infty(\Delta_G)}.$$

Based on the result, we can bound $\mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^G(I))$ by $\mathcal{N}_{\varepsilon,\infty}(\mathcal{F}(\Delta_G))$. Let us define $N := \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}(\Delta_G))$. Then, there exist $f_1', ..., f_N'$ such that for any $f' \in \mathcal{F}(\Delta_G)$, there exists $j \in \{1, ..., N\}$ such as $\|f_j' - f'\|_{L^\infty(\Delta_G)} \leq \varepsilon$. Here, for any $f \in \mathcal{F}^G(I)$, there exists $f_j := \widehat{\phi}_G^{-1}(f_j') \in \mathcal{F}^G(I)$ and it satisfies $\|f - f_j\|_{L^\infty(I)} \leq \|\widehat{\phi}_G(f) - \widehat{\phi}_G(f_j)\|_{L^\infty(\Delta_G)} \leq \varepsilon$. Then, we obtain the statement. □

*Proof of Theorem 2.* Combining Proposition 5 and 6, we obtain a bound for $\log \mathcal{N}_{2C_\Delta\delta,\infty}(\mathcal{F}^G(I))$. Then, we substitute it into (3) and obtain the statement of Theorem 2. □

### A.3 PROOF FOR SECTION 4

*Proof of Proposition 6.* We bound a covering number of a set of $C_\Delta$-Lipschitz continuous functions on $\Delta$. Let $\{x_1, ..., x_K\} \subset \Delta$ by a set of centers of $\delta$-covering set for $\Delta$. By Lemma 1, we set $K = C/(|G|\ \delta^n)$ with $\delta$ with a parameter $\delta > 0$, where $C > 0$ is a constant.

We will define a set of vectors to bound the covering number. We define a discretization operator $A : \mathcal{F}(\Delta_G) \to \mathbb{R}^K$ as

$$Af = (f(x_1)/\delta, ..., f(x_K)/\delta)^\top.$$

Let $\mathcal{B}_\delta(x)$ be a ball with radius $\delta$ in terms of the $\|\cdot\|_\infty$-norm. For two functions $f, f' \in \mathcal{F}(\Delta_G)$ such as $Af = Af'$, we obtain

$$\begin{aligned}\|f - f'\|_{L^\infty(I)} &= \max_{k=1,\dots,K} \sup_{x \in \mathcal{B}_\delta(x_k)} |f(x) - f'(x)| \\ &\leq \max_{k=1,\dots,K} \sup_{x \in \mathcal{B}_\delta(x_k)} |f(x) - f(x_k)| + |f'(x_k) - f(x_k)| \\ &\leq 2C_\Delta\delta,\end{aligned}$$

where the second inequality follows $f(x_k) = f'(x_k)$ for all $k = 1, ..., K$ and the last inequality follows the $C_\Delta$-Lipschitz continuity of $f$ and $f'$. By the relation, we can claim that $\mathcal{F}(\Delta_G)$ is covered by $2C_\Delta\delta$ balls whose center is characterized by a vector $b \in \mathbb{R}^K$ such as $b = Af$ for $f \in \mathcal{F}(\Delta_G)$. Namely, $\mathcal{N}_{2C_\Delta\delta,\infty}(\mathcal{F}(\Delta_G))$ is bounded by a number of possible $b$.

Then, we construct a specific set of $b$ to cover $\mathcal{F}(\Delta_G)$. Without loss of generality, assume that $x_1, ..., x_K$ are ordered satisfies such as $\|x_k - x_{k+1}\|_\infty \leq 2\delta$ for $k = 1, ..., K-1$. By the definition, $f \in \mathcal{F}(\Delta_G)$ satisfies $\|f\|_{L^\infty(\Delta)} \leq B$. $b_1 = f(x_1)$ can take values in $[-B/\delta, B/\delta]$. For $b_2 = f(x_2)$, since $\|x_1 - x_2\|_\infty \leq 2\delta$ and hence $|f(x_1) - f(x_2)| \leq 2C_\Delta\delta$, a possible value for $b_2$ is included in $[(b_1 - 2\delta)/\delta, (b_1 + 2\delta)/\delta]$. Hence, $b_2$ can take a value from an interval with length $4$ given $b_1$. Recursively, given $b_k$ for $k = 1, ..., K-1$, $b_{k+1}$ can take a value in an interval with length $4$.

Then, we consider a combination of the possible $b$. Simply, we obtain the number of vectors is $(2cB/\delta) \cdot (4c)^{K-1} \leq (8c^2B/\delta)^{K-1}$ with a universal constant $c \geq 1$. Then, we obtain that

$$\log \mathcal{N}_{2C_\Delta\delta,\infty}(\mathcal{F}(\Delta_G)) \leq (K-1)\log(8c^2B/\delta).$$

Then, we specify $K$ which describe a size of $\Delta$ through the set of covering centers. □

### A.4 PROOF FOR SECTION 5

**Proposition 7.** *Suppose $G$ is transitive. Then, for any $\varepsilon > 0$, we have*

$$\mathcal{N}_{\varepsilon,\infty}(\widetilde{\mathcal{F}}^G(I)) \le \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^{\mathrm{St}(G)}(I)).$$

*Proof of Proposition7.* The first statement simply follows Proposition 11 with setting $J = 1$, since $g \in G$ is transitive. In the case of $S_n$, we have $J = 1$ and Stab$(1) \cong S_{n-1}$. This gives the second statement. □

*Proof of Theorem 3 and Corollary 2.* For Theorem 3, we combine the bound (3), Lemma 2 and Proposition 5. Thus, we obtain the statement.

For Corollary 2, since $S_n$ is transitive, the statement obviously holds with $|\mathrm{St}(G)| = |S_{n-1}| = (n-1)!$. □

### A.5 PROOF FOR SECTION 6

To prove Theorem 4, we consider a Sort map and show that DNNs can represent the map. Let $\max^{(k)}(x_1, ..., x_n)$ be a map which returns the $k$-th largest value of inputted elements $x_1, ..., x_n$ for $k = 1, .., n$. Then, we provide a form of Sort as

$$\text{Sort}(x_1, \ldots, x_n) = (\max^{(1)}(x_1, \ldots, x_n), \ldots, \max^{(n)}(x_1, \ldots, x_n)).$$

To represent it, we provide the following propositions.

**Proposition 8.** $\max^{(j)}(z_1, \ldots, z_N)$ *and* $\min^{(j)}(z_1, \ldots, z_N)$ *are represented by an existing deep neural networks with an ReLU activation for any $j = 1, ..., N$.*

*Proof of Proposition 8.* Firstly, since

$$\max(z_1, z_2) = \max(z_1 - z_2, 0) + z_2,$$

and

$$\min(z_1, z_2) = -\max(z_1 - z_2, 0) + z_1$$

hold, we see the case of $j = 1, N = 2$. By repeating $\max(z_1, z_2)$, we construct $\max^{(1)}(z_1, \ldots, z_N)$ and $\min^{(1)}(z_1, \ldots, z_N)$. Namely, we prove the claim in the case of $j = 1$ and arbitrary $N$. At first, we assume $N$ is even without loss of generality, then we divide the set $\{z_1, ...z_N\}$ into sets of pairs $\{(z_1, z_2), ...(z_{N-1}, z_N)\}$. Then, by taking a max operation for each of the pairs, we have $\{y_1 = \max(z_1, z_2), ..., y_{N/2} = \max(z_{N-1}, z_N)\}$ . We repeat this process to terminate. Then we have $\max^{(1)}(z_1, \ldots, z_N)$, which is represented by an existing deep neural network. Similarly, we have $\min^{(1)}(z_1, \ldots, z_N)$. Finally, we prove the claim on $j = 2, ..., N$ by induction. Assume that for any $N$ and $\ell < j$, $\max^{(\ell)}(z_1, \ldots, z_N)$ is represented by a deep neural network. We construct $\max^{(j)}(z_1, \ldots, z_N)$ as follows: since

$$\max^{(j-1)}(z_{-\ell}) = \begin{cases} \max^{(j-1)}(z_1, \ldots, z_N) & (\text{if } z_\ell \le \max^{(j)}(z_1, \ldots, z_N)) \\ \max^{(j)}(z_1, \ldots, z_N) & (\text{otherwise}) \end{cases}$$

holds, we have $\max^{(j)}(z_1, \ldots, z_N) = \min(\{\max^{(j-1)}(Z_\ell) \mid \ell = 1, ..., N\})$. By inductive hypothesis, the right hand side is represented by a deep neural network. □

Further, we provide the following result for a technical reason.

**Proposition 9.** *The restriction map*

$$\Lambda : \mathcal{F}^{S_n}(I) \to \mathcal{F}(\Delta_{S_n})$$

*is bijective, where $\Lambda(f) = f_{\restriction \Delta_{S_n}}$.*

*Proof of Proposition 4.* To show the Proposition, we firstly define sorting layers which is an $S_n$-invariant network map from $I$ to $\Delta$. Then by Proposition 8, $\mathrm{Sort}(x_1, \ldots, x_n)$ is also a function by an $S_n$-invariant deep neural network and $\mathrm{Sort}(x_1, \ldots, x_n)$ is the function from $I$ to $\Delta$.

By using this function, we define the inverse of $\Lambda$. For any function $f$ by a deep neural network on $\Delta$, we define $\Phi(f) = f \circ \mathrm{Sort}$. We confirm $\Lambda \circ \Phi = \mathrm{id}_{\mathcal{F}_\Delta}$ and $\Phi \circ \Lambda = \mathrm{id}_{\mathcal{F}^{S_n}}$. Since we have

$$\Lambda \circ \Phi(f) = \Lambda \circ f \circ \mathrm{Sort} = (f \circ \mathrm{Sort})_{\restriction_\Delta} = f,$$

$\Lambda \circ \Phi$ is equal to $\mathrm{id}_{\mathcal{F}_\Delta}$. Similarly,

$$\Phi \circ \Lambda(f) = \Phi \circ f_{\restriction_\Delta} = f_{\restriction_\Delta} \circ \mathrm{Sort} = f,$$

where the last equality follows from the $S_n$-invariance of $f$. Hence, we have the desired result. □

Now, we are ready to prove Theorem 4.

*Proof of Theorem 4.* Let $f^*$ be an $S_n$-invariant function on $I$. Then by Proposition 9, we have a function $f$ on $\Delta_{S_n}$ such that $f^* = f \circ \mathrm{Sort}$ holds. By Theorem 5 in Schmidt-Hieber [2017], for enough big $N$, there exists a constant $c > 0$ and a neural network $f'$ with at most $\mathcal{O}(\log(N))$ layers and at most $\mathcal{O}(N \log(N))$ nonzero weights such that $\|f - f'\|_{L^\infty(I)} \leq cN^{-\alpha/p}$. Then, we have

$$\|f^* - f' \circ \mathrm{Sort}\|_{L^\infty(I)} = \|f \circ \mathrm{Sort} - f' \circ \mathrm{Sort}\|_{L^\infty(I)} \leq \|f - f'\|_{L^\infty(\Delta)} \leq \|f - f'\|_{L^\infty(I)} \leq cN^{-\alpha/p},$$

where $f \circ \mathrm{Sort}$ is a neural network with at most $\mathcal{O}(\log(N)) + K_1$ layers and at most $\mathcal{O}(N \log(N)) + K_2$ nonzero weights, where $K_1$ and $K_2$ are the number of layers and the number of nonzero weights of the neural network expressing Sort respectively. By replacing $N^{-1}$ with $\varepsilon$, we have the desired inequality. □

# B GENERALIZATION BOUND FOR EQUIVALENT DNN WITHOUT TRANSITIVE ASSUMPTION

In this section, we provide a general version of the result in Section 5. Namely, we relax the transitive assumption in the section. To the goal, we newly define a general version of a stabilizer subgroup.

Let $[n] = \{1, 2, \ldots, n\}$ be an index set and $G$ be a finite group action on $[n]$. For $i \in [n]$, we define the stabilizer subgroup $\mathrm{Stab}_G(i)$ associated with $G$ as

$$\mathrm{Stab}_G(i) = \{\sigma \in G \mid \sigma \cdot i = i\}.$$

We also consider the following decomposition of $[n]$ as

$$[n] = \bigsqcup_{j \in \mathcal{J}} \mathcal{O}_j,$$

where $\mathcal{J} \subset I$ and $\mathcal{O}_j$ is a $G$-orbit of $j$, namely the set of the form $G \cdot j$. Any $G$-orbit $G \cdot j$ is isomorphic to the set $G/\mathrm{Stab}(j)$. We denote $|\mathcal{J}|$ by $J$ and $|\mathcal{O}_j|$ by $l_j$. For each $j \in \mathcal{J}$, let $G = \bigsqcup_{j \in \mathcal{J}} \bigsqcup_{k=1}^{l_j} \mathrm{Stab}_G(j)\tau_{j,k}$ be the coset decomposition by $\mathrm{Stab}_G(j)$. Then, we may assume that $\tau_{j,k} \in G$ satisfies $\tau_{j,k}^{-1}(j) = j + k$.

Then, we provide another representation for equivariant functions from the following study.:

**Proposition 10** (Representation for Equivariant Functions Sannai et al. [2019])**.** *A map $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is $G$-equivariant if and only if $F$ can be represented by $F = (f_1 \circ \tau_{1,1}, f_1 \circ \tau_{1,2}, \ldots, f_1 \circ \tau_{1,l_1}, f_2 \circ \tau_{2,1} \ldots, f_J \circ \tau_{J,l_J})^\top$ for some $\mathrm{Stab}_G(j)$-invariant functions $f_j\colon \mathbb{R}^n \to \mathbb{R}$. Here, $\tau_{j,k} \in G$ is regarded as a linear map $\mathbb{R}^n \to \mathbb{R}^n$.*

**Proposition 11.** *For any $\varepsilon > 0$, we have*

$$\widetilde{\mathcal{N}}_{\varepsilon,\infty}(\widetilde{\mathcal{F}}^G(I)) \leq \prod_{j \in \mathcal{J}} \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^{\mathit{Stab}_G(j)}(I_{l_j})),$$

*where $I_{l_j} = [0,1]^{l_j}$. Further, if $G = S_n$,*

$$\widetilde{\mathcal{N}}_{\varepsilon,\infty}(\widetilde{\mathcal{F}}^{S_n}(I)) \leq \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^{S_{n-1}}(I)).$$

*Proof of Proposition 11.* We put $N_j = \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^{\mathrm{Stab}_G(j)}(I))$. For each $j \in \mathcal{J}$, by the definition of covering numbers, there exist $f_j^{(1)}, .., f_j^{(N_j)} \in \mathcal{F}^{\mathrm{Stab}_G(j)}(I_{l_j})$ such that for any $f' \in \mathcal{F}^{\mathrm{Stab}_G(j)}(I_{l_j})$, there exists $f_j^{(p)}$ satisfying $\|f' - f_j^{(p)}\|_\infty < \varepsilon$.

With a tuple $(p_1, ..., p_J)$, we consider a map $F_{p_1,..,p_J} : I \to \mathbb{R}^n$ from $\widetilde{\mathcal{F}}^G(I)$ and claim that balls $\mathcal{B}_\varepsilon(F_{p_1,..,p_J})$ give a covering set of $\widetilde{\mathcal{F}}(I)$. Put $F_{p_1,..,p_J} = (f_1^{(p_1)} \circ \tau_{1,1}, f_1^{(p_1)} \circ \tau_{1,2}, \ldots, f_1^{p_1} \circ \tau_{1,l_1}, f_2^{(p_2)} \circ \tau_{2,1} \ldots, f_J^{(p_J)} \circ \tau_{J,l_J})^\top$. Then $F_{p_1,..,p_J}$ is a $G$-equivariant map. Also, since $\tau_{j,k}$ is a linear map by Proposition 10, we can represent $\tau_{j,k}$ by DNNs. Hence, $F_{p_1,..,p_J} \in \widetilde{\mathcal{F}}^G(I)$ holds.

Fix $F' \in \widetilde{\mathcal{F}}^G(I)$ arbitrary. We have the representation $F' = (f_1' \circ \tau_{1,1}, f_1' \circ \tau_{1,2}, \ldots, f_1' \circ \tau_{1,l_1}, f_2' \circ \tau_{2,1} \ldots, f_J' \circ \tau_{j,l_J})^\top$ by Proposition10. Then, we can find a corresponding $F_{p_1,..,p_J}$ such as

$$\begin{aligned}
\|\|F_{p_1,..,p_J} - F'\|\|_{L^\infty(I)} &= \max\{\|f_j^{(p_j)} \circ \tau_{j,k_j} - f_j' \circ \tau_{j,k_j}\|_\infty \mid 1 \le k_j \le |G/\mathrm{Stab}_G(j)|, 1 \le p_j \le N_j\} \\
&= \max\{\|f_j^{(p_j)} - f_j'\|_\infty \mid 1 \le p_j \le N_j\} \\
&\le \varepsilon.
\end{aligned}$$

Hence, we have the first statement.

In the case of $S_n$, we have $J = 1$ and $\mathrm{Stab}(1) \cong S_{n-1}$. This gives the second statement. □

Then, we obtain the following general bound:

**Theorem 5** (Generalization of Equivariant DNN). *Suppose $\widetilde{f}^G \in \widetilde{\mathcal{F}}^G(I)$ is uniformly bounded by 1. Then, for any $\varepsilon > 0$, the following inequality holds with probability at least $1 - 2\varepsilon$:*

$$R(\widetilde{f}^G) \le R_m(\widetilde{f}^G) + \sqrt{\sum_{j \in \mathcal{J}} \frac{\widetilde{c}}{|Stab_G(j)|\, m^{2/n}}} + \sqrt{\frac{2\log(2/\varepsilon)}{m}}.$$

*where $\widetilde{c} > 0$ is a constant which are independent of $n$ and $m$.*

We omit rigorous proof of Theorem (5), because it is almost same to that of Theorem 3.

## References


Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. Universal approximations of permutation invariant/equivariant functions by deep neural networks. *arXiv preprint arXiv:1903.01939*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. *arXiv preprint arXiv:1805.07091*, 2018.