

856

857

858

Contents

859 **A Notation****24**860 **B Related work****24**861 **C Consistent sequences and compatible, transferable maps: details and missing proofs from Section 2** **25**

862 C.1 Consistent sequences and limit space 25

863 C.2 Compatible maps 27

864 C.3 Metrics on consistent sequences 28

865 **D Transferability: details and missing proofs from Section 3** **30**

866 D.1 Transferable maps 30

867 D.2 Convergence, transferability and stability 31

868 D.3 Convergence rates under sampling 33

869 **E Generalization bounds: details and missing proofs from Section 4** **34**

870 E.1 Transferable neural networks 37

871 **F Example 1 on sets: details and missing proofs from Section 5.1** **38**

872 F.1 Consistent sequences on sets 38

873 F.2 Invariant networks on sets 41

874 **G Example 2 on graphs: details and missing proofs from Section 5.2** **46**

875 G.1 Duplication consistent sequence for graphs 46

876 G.2 Message Passing Neural Networks (MPNNs) 48

877 G.3 Constructing new transferable GNNs: GGNN and continuous GGNN 50

878 **H Example 3 on point clouds: details and missing proofs from Section 5.3** **55**

879 H.1 Duplication consistent sequence for point clouds 55

880 H.2 Invariant networks on point clouds 56

881 H.3 Constructing new transferable models: SVD-DS 58

882 H.4 Transferability plots 62

883 **I Size generalization experiments: details from Section 6** **62**

884 I.1 Size generalization on sets 62

885 I.2 Size generalization on graphs 64

886 I.3 Size generalization on 3D point clouds 65

887 I.4 Continuity of Gromov-Wasserstein distance and its third lower bound 67

888 **J Limitations of this work** **68**

889

890

A Notation

We use “ \circ ” to denote composition of functions, and id for the identity element in a group or the identity map, depending on the context. Binary operations in groups are denoted by “ $*$ ”, with subscripts such as “ $*_n$ ” used when clarity is needed, e.g., to indicate the operation in the group G_n . Group actions are denoted by “ \cdot ”, with subscripts such as “ \cdot_n ” when needed.

We use \mathbb{R} and \mathbb{N} to denote the sets of real numbers and natural numbers. We use the pair (\mathbb{N}, \preceq) to denote a directed poset indexing set. We refer to it as \mathbb{N} because we always index with the natural numbers in this work, though our theory generalizes to any directed poset. The symbols \mathbb{V} and \mathbb{U} are used for consistent sequences, and V_n, U_n for the corresponding vector spaces at finite levels. Groups are denoted by G_n , with standard notations S_n for the symmetric group and $O(k)$ for the orthogonal group. We denote embeddings of vector spaces from dimension n to N by $\varphi_{N,n}$ and $\psi_{N,n}$. Embeddings of groups are denoted by $\theta_{N,n}$. A bar $\bar{}$ on a set is used to indicate either completion or closure, depending on the context; a bar $\bar{}$ on a function is used to denote the normalized variant of that function.

Given two sequences $(a_n), (b_n) \subseteq \mathbb{R}$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all $n \in \mathbb{N}$. For a given function $R : \mathbb{N} \rightarrow \mathbb{R}_+$, we say a sequence (x_n) converges to x at rate $R(n)$ with respect to distance metric d if $d(x_n, x) \lesssim R(n)$ and $R(n) \rightarrow 0$.

We write \sim to denote equivalence relations, and use $[x]$ to denote the equivalence class of x , i.e., $[x] = \{y : y \sim x\}$. We write \vee and \wedge for maximum and minimum respectively. In a metric space, we use $B(x_0, r)$ to denote the ball centered at x_0 with radius r , i.e. $B(x_0, r) := \{x : d(x_0, x) < r\}$.

B Related work

GNN transferability. The work on GNN transferability under the graphon framework was pioneered by [56], focusing on a variant of graph convolutional network (GCN) for deterministic graphs obtained from the same graphon. In parallel, [38] explores transferability with respect to an alternative limit space to graphon in the form of a topological space, and [31] examines an arguably equivalent notion—convergence and stability—by analyzing random graphs sampled from a graphon. This line of research has since been further developed [57, 47], extending to more general message-passing networks (MPNNs) [14], more general notion of graph limits [36], and other models [59, 64, 7, 27]. Our framework unifies and recovers several of the above-mentioned results, which we briefly discuss in Appendix G.2. Furthermore, we develop new insights into the transferability of Invariant Graph Networks (IGNs) [45], a topic previously examined in [7, 27]. We leverage our framework to advance this line of inquiry and elaborate on the connections between our approach and prior work in Appendix G.3.

GNN generalization A related line of research concerns the generalization theory of GNNs. Generalization bounds have been derived based on various frameworks, including VC dimension [60], the PAC-Bayesian approach [42], the neural tangent kernel [17], and covering numbers [48, 46, 37, 62, 54]. Notions of size generalization based on counting substructures [10, 21] without an explicit notion of convergence have been studied in [71]. For a comprehensive overview, we refer the reader to the recent survey [63].

While most of these works consider graphs of fixed or bounded size, [37, 54] explore the “uniform regime,” addressing unbounded graph sizes. This perspective is closely aligned with the transferability theory, as they both consider continuous extensions over suitable limit spaces. Our work formalizes this connection, showing that transferability implies generalization (Section 4, Appendix E). We derive a generalization bound via covering numbers using the same proof strategy as prior works, but applies well beyond GNNs and offers a setting that directly connects to the notion of size generalization.

Equivariant machine learning Our theory naturally applies to equivariant machine learning models, i.e., neural networks with symmetries imposed. We particularly focus on models from [73, 45, 52, 58, 2]. There are many other equivariant machine learning models we haven’t discussed here, such as the ones expressed in terms of group convolutions [13, 35, 25], representation theory [34, 61, 23], canonicalization [30], and invariant theory [66, 3, 67], among others.

Representation stability and any-dimensional learning Notably, many equivariant models naturally handle inputs of varying sizes—either through mechanisms such as padding, truncation, and sliding windows (e.g., transformers), patches (e.g., visual transformers), or by design principles that inherently support variable-sized inputs (e.g., GNNs and DeepSets). As noted by [39, 40], the presence of symmetry allows functions operating on inputs of arbitrary dimension to be parameterized with a finite number of parameters, which may be partially explained by the theory of representation stability [11]. Leveraging techniques from this theory, [40] provides the first general theoretical framework for any-dimensional equivariant models. Recent work applies similar techniques to study the generalization properties across dimensions of any-dimensional regression models [15]. Our work builds on the theoretical foundation established by this line of research.

Any-dimensional expressivity and universality Our framework naturally prompts the question of expressivity for any-dimensional neural networks on their limit spaces. While we do not pursue this direction here, related questions have been independently studied. [5] shows that normalized DeepSets and PointNet are universal for uniformly continuous functions on suitable limit spaces, closely tied to our results (see Appendix F.2.1). In GNNs, the notion of “uniform expressivity”—expressing logical queries without parameter dependence on input size—has been explored in [26, 1, 32, 55]. Our framework offers complementary insights despite differing foundations.

C Consistent sequences and compatible, transferable maps: details and missing proofs from Section 2

C.1 Consistent sequences and limit space

The concept of consistent sequences originated in the theory of representation stability [12]. We generalize this notion, originally considering sequence of vector spaces, by allowing indexing over any directed poset.

Directed poset indexing. We require the indexing set (\mathbb{N}, \preceq) to be a *directed poset*, meaning that \mathbb{N} is a non-empty set equipped with a binary operation \preceq satisfying:

(*Partial order*) The binary operation \preceq is a partial order; that is, it satisfies: reflexivity ($a \preceq a$ for all $a \in \mathbb{N}$), transitivity (if $a \preceq b$ and $b \preceq c$, then $a \preceq c$), and antisymmetry (if $a \preceq b$ and $b \preceq a$, then $a = b$).

(*Upper bound condition*) For every pair of elements $a, b \in \mathbb{N}$, there exists $c \in \mathbb{N}$ such that $a \preceq c$ and $b \preceq c$.

The directed poset indexing generalizes the notion of sequences to allow a more complex and flexible way of defining how smaller problem instances are embedded into larger ones, permitting “branching” directions of growth. We will see that the upper bound condition is crucial, as it ensures that any two problem instances are comparable—meaning they can both be embedded into a third, larger problem dimension and compared there.

While our theory allows indexing over arbitrary directed posets, in this work we only consider two cases: the natural numbers with the standard ordering \leq , and with the divisibility ordering, where $a \preceq b$ if and only if $a \mid b$. Accordingly, we use (\mathbb{N}, \preceq) to denote the index set.

Consistent sequences and their limit space.

Definition C.1 (Consistent sequence: detailed version of Definition 2.1). *A consistent sequence of group representations over directed poset (\mathbb{N}, \preceq) is $\mathbb{V} = \{(V_n)_{n \in \mathbb{N}}, (\varphi_{N,n})_{n \preceq N}, (G_n)_{n \in \mathbb{N}}\}$, where*

(1) *(G_n) is a sequence of groups indexed by \mathbb{N} such that whenever $n \preceq N$, G_n is embedded into G_N via an injective group homomorphism $\theta_{N,n}: G_n \rightarrow G_N$, where*

$$\begin{aligned} \theta_{i,i} &= \text{id}_{G_i} \quad \text{for all } i \in \mathbb{N}, \\ \theta_{k,j} \circ \theta_{j,i} &= \theta_{k,i} \quad \text{whenever } i \preceq j \preceq k \text{ in } \mathbb{N}. \end{aligned}$$

(2) *(V_n) is a sequence of (finite-dimensional, real) vector spaces indexed by \mathbb{N} , such that each V_n is a G_n -representation, and whenever $n \preceq N$, V_n is embedded into V_N through a linear embedding $\varphi_{N,n}: V_n \hookrightarrow V_N$, where*

$$\varphi_{i,i} = \text{id}_{V_i} \quad \text{for all } i \in \mathbb{N},$$

989

$$\varphi_{k,j} \circ \varphi_{j,i} = \varphi_{k,i} \quad \text{whenever } i \preceq j \preceq k \text{ in } \mathbb{N}.$$

990 (3) Every $\varphi_{N,n}$ is G_n -equivariant, i.e.,

$$\varphi_{N,n}(g \cdot v) = \theta_{N,n}(g) \cdot \varphi_{N,n}(v) \text{ for all } g \in G_n, v \in V_n.$$

991 Given a consistent sequence, we can first “summarize” the sequence of groups (G_n) into a single
 992 limit group G_∞ , and likewise “summarize” the sequence of vector spaces (V_n) into a single limit
 993 vector space V_∞ . We can then consider the action of G_∞ on V_∞ .

994 **Definition C.2** (Limit group: detailed version of Definition 2.3). The **limit group** G_∞ is defined as
 995 the disjoint union $\bigsqcup_n G_n$ modulo the equivalence relation that identifies each element $g \in G_n$ with
 996 its images under the transition maps $\theta_{N,n}(g)$ for all $N \geq n$, i.e.,

$$G_\infty := \bigsqcup_n G_n / \sim,$$

997 where $g \sim \theta_{N,n}(g)$ whenever $n \preceq N$ and $g \in G_n$. This construction is also known as the **direct limit**
 998 of groups, and is denoted by $G_\infty = \varinjlim G_n$.

999 For each $g \in G_n$, we denote by $[g]$ its equivalence class in G_∞ , representing the corresponding
 1000 limiting object.

1001 The group structure on G_∞ is inherited from the groups (G_n) as follows. For $g_n \in G_n$ and $g_m \in G_m$,
 1002 define the binary operation on equivalence classes by

$$[g_n] * [g_m] := [\theta_{N,n}(g_n) *_N \theta_{N,m}(g_m)],$$

1003 where $N \in \mathbb{N}$ is a common upper bound of n and m in (\mathbb{N}, \preceq) , and $*_N$ denotes the group operation
 1004 in G_N . It is straightforward to check that this operation is well-defined.

1005 **Definition C.3** (Limit space of consistent sequence: detailed version of Definition 2.3). Define V_∞
 1006 as the disjoint union $\bigsqcup_n V_n$ modulo an equivalence relation identifying each element $v \in V_n$ with its
 1007 images under the transition map $\varphi_{N,n}(v)$ for all $n \preceq N$, i.e.,

$$V_\infty := \bigsqcup_n V_n / \sim,$$

1008 where $v \sim \varphi_{N,n}(v)$ whenever $n \preceq N$. This construction is also known as the **direct limit** of vector
 1009 spaces, and is denoted as $V_\infty = \varinjlim V_n$.

1010 The vector space structure on V_∞ is inherited from the vector spaces (V_n) as follows. For $v_n \in$
 1011 $V_n, v_m \in V_m$, the addition and scalar multiplication on the equivalent classes are defined by

$$\begin{aligned} [v_n] + [v_m] &:= [\varphi_{N,n}(v_n) + \varphi_{N,m}(v_m)] \quad \text{where } N \text{ is an upper bound for } n, m \text{ in } (\mathbb{N}, \preceq), \\ \lambda[v_n] &:= [\lambda v_n]. \end{aligned}$$

1013 It is straightforward to check that these operations are well-defined.

1014 The limit group G_∞ acts on V_∞ by

$$[g] \cdot [v] := [\theta_{N,n}(g) \cdot_N \varphi_{N,m}(v)] \text{ for } g \in G_n, v \in V_m,$$

1015 where N is an upper bound of n, m in (\mathbb{N}, \preceq) , and \cdot_N is the group action of G_N on V_N . It is also
 1016 easy to check that this group action is well-defined, and V_∞ is a G_∞ -representation.

1017 The orbit space of V_∞ under the action of G_∞ is

$$\{G_\infty \cdot x : x \in V_\infty\},$$

1018 where $G_\infty \cdot x := \{g \cdot x : g \in G_\infty\}$ is the orbit of point $x \in V_\infty$ under the G_∞ -action. The orbits
 1019 form a partition of V_∞ into disjoint subsets.

1020 **Consistent sequences without symmetries.** A special case of consistent sequences arises when
 1021 $G_n = \{\text{id}\}$, the trivial group, for all $n \in \mathbb{N}$. In this case, the structure reduces to a directed system of
 1022 vector spaces

$$\mathbb{V} = \{(V_n)_{n \in \mathbb{N}}, (\varphi_{N,n})_{n \preceq N}\}.$$

1023 Hence, our theory of size generalization applies in scenarios where no intrinsic symmetries are
 1024 present.

1025 **Trivial consistent sequence.** Another special case arises when $V_n = V$, a fixed vector space, for
 1026 all $n \in \mathbb{N}$, with embeddings given by $\varphi_{N,n} = \text{id}_V$ for all $n \preceq N$. This yields the *trivial consistent*
 1027 *sequence* associated with V , which we denote by \mathbb{V}_V . This construction is useful for modelling
 1028 non-size-dependent spaces, such as the output space in graph-level classification or regression tasks.

1029 **Direct sum and tensor product.** Given a consistent sequence $\mathbb{V} = \{(V_n), (\varphi_{N,n}), (G_n)\}$, we can
 1030 define its *direct sum* and *tensor product*. Both of them are also consistent sequences.

1031 **Definition C.4.** The d -th *direct sum* of \mathbb{V} is defined as

$$\mathbb{V}^{\oplus d} := \{(V_n^{\oplus d}), (\varphi_{N,n}^{\oplus d}), (G_n)\},$$

1032 where $V_n^{\oplus d}$ denotes the direct sum of d copies of V_n and $\varphi_{N,n}^{\oplus d} : V_n^{\oplus d} \rightarrow V_N^{\oplus d}$ is defined by applying
 1033 $\varphi_{N,n}$ to each component. The group G_n acts on $V_n^{\oplus d}$ by simultaneously acting on every copy of V_n ,
 1034 i.e. $g \cdot (v_1, \dots, v_d) := (g \cdot v_1, \dots, g \cdot v_d)$.

1035 **Definition C.5.** The d -th *tensor product* of \mathbb{V} is defined as

$$\mathbb{V}^{\otimes d} := \{(V_n^{\otimes d}), (\varphi_{N,n}^{\otimes d}), (G_n)\},$$

1036 where $V_n^{\otimes d}$ denotes the d -fold tensor product of V_n . $\varphi_{N,n}^{\otimes d} : V_n^{\otimes d} \rightarrow V_N^{\otimes d}$ is uniquely defined by
 1037 $\varphi_{N,n}^{\otimes d}(v_1 \otimes \dots \otimes v_d) := \varphi_{N,n}(v_1) \otimes \dots \otimes \varphi_{N,n}(v_d)$, and the group action of G_n on $V_n^{\otimes d}$ is uniquely
 1038 defined by $g \cdot (v_1 \otimes \dots \otimes v_d) := (g \cdot v_1) \otimes \dots \otimes (g \cdot v_d)$. The universal property of tensor product
 1039 guarantees that $\varphi_{N,n}^{\otimes d}$ and the group action mentioned are well-defined.

1040 Similarly, we also consider the d -th *symmetric tensors* of \mathbb{V} :

$$\text{Sym}^d(\mathbb{V}) := \{(\text{Sym}^d(V_n)), (\varphi_{N,n}^{\otimes d}), (G_n)\},$$

1041 where $\text{Sym}^d(V_n)$ denotes the space of symmetric tensors of order d defined on V_n , i.e. the subspace
 1042 of $V_n^{\otimes d}$ invariant under the action of the symmetric group S_d .

1043 The direct sum, $\mathbb{V}^{\oplus d}$, is particularly useful for incorporating hidden channels into our analysis, as
 1044 it effectively adds the extra channel dimensions to our data. In contrast, the tensor product, $\mathbb{V}^{\otimes d}$, is
 1045 helpful to extend a consistent sequence on vectors or sets to higher-order objects such as matrices
 1046 or graphs. For example, the duplication consistent sequences for graphs exactly arise as the 2nd
 1047 symmetric tensors of the duplication consistent sequences for sets.

1048 C.2 Compatible maps

1049 Recall from Definition 2.4 that a sequence of maps $(f_n : V_n \rightarrow U_n)$ is *compatible* with respect to the
 1050 consistent sequences \mathbb{V}, \mathbb{U} if, for all $n \preceq N$,

$$f_N \circ \varphi_{N,n} = \psi_{N,n} \circ f_n,$$

1051 and each f_n is G_n -equivariant. This condition is equivalent to the existence of an extension to the
 1052 limit map f_∞ .

1053 **Proposition C.6** (Compatible maps and extension to limit). *Let $\mathbb{V} = \{(V_n), (\varphi_{N,n}), (G_n)\}$ and*
 1054 *$\mathbb{U} = \{(U_n), (\psi_{N,n}), (G_n)\}$ be two consistent sequences. A sequence of maps $(f_n : V_n \rightarrow U_n)$ is*
 1055 *compatible if and only if it extends to the limit; that is, there exists a G_∞ -equivariant map*

$$f_\infty : V_\infty \rightarrow U_\infty$$

1056 *such that $f_n = f_\infty|_{V_n}$ for all n .*

1057 *Proof.* The “ \Leftarrow ” direction follows immediately from the definition by restricting to each V_n . The
 1058 converse holds because we can always map objects of different dimensions into a common larger
 1059 dimension and verify the properties there. More concretely:

1060 (\Rightarrow) Suppose there exists a G_∞ -equivariant map f_∞ such that $f_n = f_\infty|_{V_n}$ for all n . Then for all
 1061 $n \preceq N$ and $x \in V_n$,

$$[f_n(x)] = f_\infty([x]) = f_\infty([\varphi_{N,n}(x)]) = [f_N(\varphi_{N,n}(x))],$$

1062 which implies $f_N \circ \varphi_{N,n} = \psi_{N,n} \circ f_n$. Moreover, for all $n \in \mathbb{N}$, $x \in V_n$, and $g \in G_n$,

$$[f_n(g \cdot x)] = f_\infty([g \cdot x]) = f_\infty([g] \cdot [x]) = [g] \cdot f_\infty([x]) = [g] \cdot [f_n(x)] = [g \cdot f_n(x)],$$

1063 so each f_n is G_n -equivariant.

1064 (\Rightarrow) Conversely, suppose (f_n) are compatible. Define

$$f_\infty: V_\infty \rightarrow W_\infty, \quad f_\infty([x]) := [f_n(x)] \quad \text{if } x \in V_n.$$

1065 Compatibility ensures this is well-defined. To verify equivariance, let $g \in G_m$ and $x \in V_n$, and let N
1066 be a common upper bound of n and m in (\mathbb{N}, \preceq) . Then,

$$\begin{aligned} f_\infty([g] \cdot [x]) &= f_\infty([\theta_{N,m}(g) \cdot \varphi_{N,n}(x)]) \\ &= [f_N(\theta_{N,m}(g) \cdot \varphi_{N,n}(x))] \\ &= [\theta_{N,m}(g) \cdot f_N(\varphi_{N,n}(x))] \\ &= [\theta_{N,m}(g) \cdot \psi_{N,n}(f_n(x))] \\ &= [g] \cdot f_\infty([x]). \end{aligned}$$

1067 Therefore, f_∞ is G_∞ -equivariant. □

1068 This proposition implies that learning a function on the infinite-dimensional space V_∞ , a task that
1069 may appear difficult, reduces to learning a compatible sequence of functions on the finite-dimensional
1070 vector spaces along the sequence, which is a more tractable problem.

1071 C.3 Metrics on consistent sequences

1072 In Section 2.1, we introduced a norm on V_∞ so as to define distance between objects of different
1073 dimensions. In this appendix, we take a more general perspective and examine metric structures on
1074 consistent sequences, and present detailed proofs. The same proofs carry over to the norm setting
1075 with minimal modification.

1076 **Definition C.7** (Compatible metrics: generalized version of Definition 2.5). *Let $\mathbb{V} =$
1077 $\{(V_n), (\varphi_{N,n}), (G_n)\}$ be a consistent sequence. A sequence of metrics (d_n) on the vector spaces V_n
1078 is said to be **compatible** if all the embeddings $\varphi_{N,n}$ and the G_n -actions are isometries. That is, for
1079 all $n \preceq N$, $x, y \in V_n$, and $g \in G_n$, we have:*

$$d_N(\varphi_{N,n}(x), \varphi_{N,n}(y)) = d_n(x, y), \quad \text{and} \quad d_n(g \cdot x, g \cdot y) = d_n(x, y).$$

1080 Similar to compatible maps, this is equivalent to the existence of an extension to a metric d_∞ on V_∞ .

1081 **Proposition C.8** (Compatible metrics and extension to the limit). *A sequence of metrics (d_n) on the
1082 spaces V_n is compatible if and only if it extends to a metric on the limit space. That is, there exists a
1083 metric d_∞ on V_∞ such that*

$$d_n(x, y) = d_\infty([x], [y]) \quad \text{for all } n \in \mathbb{N}, x, y \in V_n,$$

1084 *and the G_∞ -action on V_∞ is an isometry with respect to d_∞ , i.e.,*

$$d_\infty(g \cdot x, g \cdot y) = d_\infty(x, y) \quad \text{for all } x, y \in V_\infty, g \in G_\infty.$$

1085 *Proof.* The proof is primarily a matter of bookkeeping, similar in spirit to Proposition C.6. For
1086 completeness, we present the full argument below.

1087 (\Leftarrow) Suppose (d_n) extends to a metric d_∞ on V_∞ . Then for all $n \preceq N$ and $x, y \in V_n$, we have

$$d_n(x, y) = d_\infty([x], [y]) = d_\infty([\varphi_{N,n}(x)], [\varphi_{N,n}(y)]) = d_N(\varphi_{N,n}(x), \varphi_{N,n}(y)).$$

1088 Thus, the embeddings $\varphi_{N,n}$ are isometries. Moreover, for any $g \in G_n$ and $x, y \in V_n$,

$$d_n(g \cdot x, g \cdot y) = d_\infty([g \cdot x], [g \cdot y]) = d_\infty([g] \cdot [x], [g] \cdot [y]) = d_\infty([x], [y]) = d_n(x, y),$$

1089 so the group actions are isometries as well.

1090 (\Rightarrow) Conversely, suppose the collection (d_n) is compatible. Define a metric $d_\infty: V_\infty \times V_\infty \rightarrow \mathbb{R}$ as
1091 follows: for $x \in V_n$ and $y \in V_m$, let N be any common upper bound of n and m in (\mathbb{N}, \preceq) , and set

$$d_\infty([x], [y]) := d_N(\varphi_{N,n}(x), \varphi_{N,m}(y)).$$

1092 Compatibility of the metrics ensures that d_∞ is well-defined. It is also easy to check that d_∞ is a
 1093 metric. Moreover, by construction, $d_n = d_\infty|_{V_n}$ for all n .

1094 To verify that the G_∞ -action is isometric, take $x \in V_{n_1}$, $y \in V_{n_2}$, and $g \in G_n$ for some $n \in \mathbb{N}$. Let
 1095 N be a common upper bound of n, n_1, n_2 in (\mathbb{N}, \preceq) . Then,

$$\begin{aligned} d_\infty([g] \cdot [x], [g] \cdot [y]) &= d_N(\theta_{N,n}(g) \cdot \varphi_{N,n_1}(x), \theta_{N,n}(g) \cdot \varphi_{N,n_2}(y)) \\ &= d_N(\varphi_{N,n_1}(x), \varphi_{N,n_2}(y)) \\ &= d_\infty([x], [y]). \end{aligned}$$

1096

□

1097 With the metric structure in place, we define the limit space via the completion of the metric space.
 1098 The *completion* of a metric space M is a complete metric space \overline{M} —that is, a space in which every
 1099 Cauchy sequence converges—that contains M as a dense subset (i.e., the smallest closed subset of
 1100 \overline{M} containing M is \overline{M} itself).

1101 **Definition C.9** (Limit space: detailed version of Definition 2.7). *Let \mathbb{V} be a consistent sequence, and*
 1102 *let V_∞ be equipped with the metric d_∞ . Denote by $\overline{V_\infty}$ the completion of V_∞ with respect to d_∞ .*

1103 *The G_∞ -action on V_∞ extends to a well-defined action on $\overline{V_\infty}$ as follows: for any $x \in \overline{V_\infty}$ and*
 1104 *$g \in G_\infty$, choose a sequence (x_n) in V_∞ such that $x_n \rightarrow x$ in $\overline{V_\infty}$, and define*

$$g \cdot x := \lim_{n \rightarrow \infty} g \cdot x_n.$$

1105 *This limit exists because $(g \cdot x_n)$ is a Cauchy sequence, as the G_∞ -action on V_∞ is isometric. The*
 1106 *resulting action on $\overline{V_\infty}$ is linear and isometric.*

1107 *We define the **limit space** of the consistent sequence \mathbb{V} to be the G_∞ -representation $\overline{V_\infty}$.*

1108 The set of orbit closures in $\overline{V_\infty}$ under the action of G_∞ is

$$\{\overline{G_\infty \cdot x} : x \in \overline{V_\infty}\}.$$

1109 where $\overline{G_\infty \cdot x}$ is the closure of the orbit $G_\infty \cdot x$.

1110 Intuitively, $\overline{V_\infty}$ includes not only elements from finite-dimensional objects (elements in V_∞), but also
 1111 additional points that are “reachable” as limits of finite-dimensional objects. We can further define a
 1112 symmetrized metric on the limit space.

1113 **Proposition C.10** (Symmetrized metric). *Let $x, y \in \overline{V_\infty}$. Define*

$$\overline{d}(x, y) := \inf_{g \in G_\infty} d_\infty(g \cdot x, y).$$

1114 *Then \overline{d} is a pseudometric on $\overline{V_\infty}$ and induces a metric on the space of orbit closures in $\overline{V_\infty}$ under the*
 1115 *G_∞ -action. We refer to \overline{d} as the symmetrized metric.*

1116 *Proof.* The non-negativity of \overline{d} follows directly from the non-negativity of d_∞ .

1117 *Symmetry:* Since $d_\infty(g \cdot x, y) = d_\infty(x, g^{-1} \cdot y)$ for any $g \in G_\infty$ by isometry of the G_∞ -action,
 1118 taking the infimum over all $g \in G_\infty$ (equivalently over g^{-1}) yields

$$\overline{d}(x, y) = \overline{d}(y, x).$$

1119 *Triangle inequality:* Let $\epsilon > 0$. Then there exist $g, h \in G_\infty$ such that

$$\overline{d}(x, y) > d_\infty(g \cdot x, y) - \epsilon, \quad \overline{d}(y, z) > d_\infty(h \cdot y, z) - \epsilon.$$

1120 Using the isometry of the group action:

$$\begin{aligned} \overline{d}(x, y) + \overline{d}(y, z) &> d_\infty(g \cdot x, y) + d_\infty(h \cdot y, z) - 2\epsilon \\ &= d_\infty(hg \cdot x, h \cdot y) + d_\infty(h \cdot y, z) - 2\epsilon \\ &\geq d_\infty(hg \cdot x, z) - 2\epsilon \\ &\geq \overline{d}(x, z) - 2\epsilon. \end{aligned}$$

1121 Since this holds for arbitrary $\epsilon > 0$, the triangle inequality follows.

1122 *Definiteness:* We have $\bar{d}(x, y) = 0$ if and only if there exists a sequence $(g_n) \subseteq G_\infty$ such that
 1123 $d_\infty(g_n \cdot x, y) \rightarrow 0$. This is precisely the condition that $y \in \overline{G_\infty \cdot x}$, i.e., x and y lie in the same orbit
 1124 closure.

1125 Therefore, \bar{d} is a pseudometric on $\overline{V_\infty}$, and descends to a true metric on the space of orbit closures. \square

1126 D Transferability: details and missing proofs from Section 3

1127 D.1 Transferable maps

1128 Following Definition 3.1 of continuously, L -Lipschitz, and $L(r)$ -locally Lipschitz transferable, we
 1129 further define the following notion: If (f_n) is a sequence of mappings extending to f_∞ which is
 1130 locally Lipschitz at x_0 ⁶, we say (f_n) is *locally Lipschitz transferable* at x_0 . Being locally Lipschitz
 1131 transferable at x_0 is a weaker condition than being $L(r)$ -locally Lipschitz transferable, which is
 1132 itself weaker than L -Lipschitz transferable. This definition is useful when studying models which
 1133 are discontinuous on negligible sets of inputs, and which are therefore not $L(r)$ -locally Lipschitz
 1134 transferable. These often come up when constructing architectures based on canonicalizations [18],
 1135 as commonly done for point clouds for instance (see Section 5.3).

1136 We first show a useful property that continuity/Lipschitz with respect to $\|\cdot\|_{V_\infty}, \|\cdot\|_{U_\infty}$ implies the
 1137 same property with respect to the symmetrized metrics, even though the converse does not hold. We
 1138 will again state and prove the results for metrics instead.

1139 **Proposition D.1** (Continuity in d_∞ implies continuity in \bar{d}). *Let \mathbb{V}, \mathbb{U} be consistent sequences. If
 1140 $f_\infty : \overline{V_\infty} \rightarrow \overline{U_\infty}$ is continuous (respectively, $L(r)$ -Lipschitz on $B(0, r)$ for all $r > 0$, L -Lipschitz,
 1141 locally Lipschitz at $x_0 \in \overline{V_\infty}$) with respect to d_∞^V and d_∞^U , then f_∞ satisfies the same property with
 1142 respect to the symmetrized metrics \bar{d}_V and \bar{d}_U .*

1143 *Proof.* (Continuity) Let $x_n \rightarrow x$ in $\overline{V_\infty}$ with respect to the symmetrized metric \bar{d}_V , and $\epsilon > 0$.
 1144 By continuity of f_∞ with respect to d_∞ , there exists $\delta > 0$ such that whenever $d_\infty(x, y) < \delta$,
 1145 $d_\infty(f_\infty(x), f_\infty(y)) < \epsilon$. Take N such that for all $n \geq N$, $\bar{d}_V(x_n, x) < \frac{\delta}{2}$. Moreover, for each n ,
 1146 choose $g_n \in G_\infty$ such that $d_\infty(g_n \cdot x_n, x) \leq \bar{d}_V(x_n, x) + \frac{\delta}{2}$. Then for all $n \geq N$, $d_\infty(g_n \cdot x_n, x) < \delta$
 1147 and hence

$$\bar{d}_U(f_\infty(x_n), f_\infty(x)) \leq d_\infty(g_n \cdot f_\infty(x_n), f_\infty(x)) = d_\infty(f_\infty(g_n \cdot x_n), f_\infty(x)) < \epsilon.$$

1148 Therefore $f_\infty(x_n) \rightarrow f_\infty(x)$ with respect to \bar{d}_U .

1149 ($L(r)$ -Lipschitz on $B(0, r)$) Suppose f_∞ is $L(r)$ -Lipschitz on $B(0, r)$ for all $r > 0$ with respect to
 1150 d_∞ . Consider x, y such that $\bar{d}_V(0, x) = d_\infty^V(0, x) < r$, and $\bar{d}_V(0, y) = d_\infty^V(0, y) < r$. Then for any
 1151 $g \in G_\infty$, we have $d_\infty^V(0, g \cdot x) = d_\infty^V(0, x) < r$. Hence,

$$\bar{d}_U(f_\infty(x), f_\infty(y)) \leq d_\infty(g \cdot f_\infty(x), f_\infty(y)) = d_\infty(f_\infty(g \cdot x), f_\infty(y)) \leq L(r)d_\infty(g \cdot x, y).$$

1152 Take infimum over $g \in G_\infty$, get $\bar{d}_U(f_\infty(x), f_\infty(y)) \leq L(r)\bar{d}_V(x, y)$.

1153 (Lipschitz) For any $x, y \in V_\infty, g \in G_\infty$,

$$\bar{d}_U(f_\infty(x), f_\infty(y)) \leq d_\infty(g \cdot f_\infty(x), f_\infty(y)) = d_\infty(f_\infty(g \cdot x), f_\infty(y)) \leq Ld_\infty(g \cdot x, y).$$

1154 Take infimum over $g \in G_\infty$, get $\bar{d}_U(f_\infty(x), f_\infty(y)) \leq L\bar{d}_V(x, y)$.

1155 (Locally Lipschitz at x_0) Suppose $d_\infty(f(x), f(x_0)) \leq Ld_\infty(x, x_0)$ whenever $d_\infty(x, x_0) < r$. Then
 1156 for any such x we have $\bar{d}_V(x, x_0) = \inf_{g \in G_\infty^{(r)}} d_\infty(x, x_0)$ where $G_\infty^{(r)} = \{g \in G_\infty : d_\infty(x, x_0) \leq r\}$.

1157 Moreover, for any $g \in G_\infty^{(r)}$ we have $\bar{d}_U(f(x), f(x_0)) \leq d_\infty(f(g \cdot x), f(x_0)) \leq Ld_\infty(g \cdot x, x)$,
 1158 so after taking infimum over such g we conclude that $\bar{d}_U(f(x), f(x_0)) \leq L\bar{d}_V(x, x_0)$ whenever
 1159 $\bar{d}_V(x, x_0) < r$, as desired. \square

⁶We say f_∞ is locally Lipschitz at x_0 if there exists $r > 0$ and $L > 0$ such that for all $x \in B(x_0, r)$,
 $d_\infty(f_\infty(x), f_\infty(x_0)) \leq Ld_\infty(x, x_0)$. Notice that this is slightly different from saying f_∞ is locally Lipschitz
 around x_0 , which means that there exists $r > 0$ and $L > 0$ such that f_∞ is L -Lipschitz on $B(x_0, r)$.

Finally, we state and prove a set of more concrete characterizations of Lipschitz transferable sequence of functions defined in Definition 3.1, which are straightforward to check.

Proposition D.2. *Let \mathbb{V}, \mathbb{U} be consistent sequences endowed with metrics.*

- (1) *A compatible sequence of functions $(f_n : V_n \rightarrow U_n)$ is L -Lipschitz (respectively, $L(r)$ -locally Lipschitz) transferable if and only if for all n , f_n is L -Lipschitz (respectively, $L(r)$ -Lipschitz on $B_n(0, r) := \{v \in V_n : d_n^V(0, v) < r\}$).*
- (2) *When the metrics are induced by norms, a compatible sequence of linear maps $(W_n : V_n \rightarrow U_n)$ is continuously (respectively, L -Lipschitz) transferable if and only if $\sup_n \|W_n\|_{\text{op}} < \infty$ (respectively, $\sup_n \|W_n\|_{\text{op}} \leq L$).*

Proof. (1) The “ \Rightarrow ” direction again follows immediately from $d_n = d_\infty|_{V_n}, f_n = f_\infty|_{V_n}$ for all n . We focus on proving “ \Leftarrow ”. First, by Proposition C.6, compatibility implies that the sequence (f_n) extends to a function $f_\infty : V_\infty \rightarrow U_\infty$.

(Lipschitz) Suppose f_n is L -Lipschitz for all n . For any $x \in V_n$ and $y \in V_m$, let N be a common upper bound of n and m in (\mathbb{N}, \preceq) . Then:

$$\begin{aligned} d_\infty^U(f_\infty([x]), f_\infty([y])) &= d_N^U(f_N(\varphi_{N,n}(x)), f_N(\varphi_{N,m}(y))) \\ &\leq L d_N^V(\varphi_{N,n}(x), \varphi_{N,m}(y)) \\ &= L d_\infty^V([x], [y]). \end{aligned}$$

Hence, f_∞ is L -Lipschitz on V_∞ .

Since Lipschitz continuity implies Cauchy continuity, f_∞ extends uniquely to $f_\infty : \overline{V_\infty} \rightarrow \overline{U_\infty}$.

For any Cauchy sequences (x_n) and (y_n) in V_∞ with limits $x, y \in \overline{V_\infty}$, we have:

$$\begin{aligned} d_\infty^U(f_\infty(x), f_\infty(y)) &= \lim_{n \rightarrow \infty} d_\infty^U(f_\infty(x_n), f_\infty(y_n)) \\ &\leq \lim_{n \rightarrow \infty} L d_\infty^V(x_n, y_n) = L d_\infty^V(x, y), \end{aligned}$$

which shows that f_∞ remains L -Lipschitz after extending to $\overline{V_\infty}$.

($L(r)$ -locally Lipschitz) Suppose each f_n is $L(r)$ -Lipschitz on $B_n(0, r)$ for all $r > 0$. As above, for any $x \in V_n$ and $y \in V_m$ with $d_n(0, x), d_m(0, y) < r$, let N be a common upper bound of n, m in (\mathbb{N}, \preceq) . Then $\varphi_{N,n}(x), \varphi_{N,m}(y) \in B_N(0, r)$, and by the $L(r)$ -Lipschitz property of f_N on $B_N(0, r)$, we get

$$d_\infty^U(f_\infty([x]), f_\infty([y])) \leq L(r) \cdot d_\infty^V([x], [y]).$$

Thus, $f_\infty : V_\infty \rightarrow U_\infty$ is $L(r)$ -Lipschitz on $\{v \in V_\infty : d_\infty^V(0, v) < r\}$.

This implies f_∞ is Cauchy continuous: any Cauchy sequence (x_n) lies within some ball of radius R , and since f_∞ is Lipschitz continuous there, $(f_\infty(x_n))$ is also Cauchy. Hence, f_∞ extends uniquely to $\overline{V_\infty}$, and it is easy to check that it is $L(r)$ -Lipschitz on $B(0, r)$ for all r .

(2) (Lipschitz) By (1), (W_n) is L -Lipschitz transferable if and only if for all n , W_n is L -Lipschitz. By linearity of each W_n , this is equivalent to $\|W_n\|_{\text{op}} \leq L$ for all n .

(Continuity) It is sufficient to prove that $\{W_n\}$ is continuously transferable if and only if it is L -Lipschitz transferable for some $L > 0$. The “ \Leftarrow ” direction is immediate. To prove “ \Rightarrow ”, suppose (W_n) extends to a continuous function $W_\infty : \overline{V_\infty} \rightarrow \overline{U_\infty}$. Then W_∞ is linear on V_∞ because for any $x \in V_n, y \in V_m$, and any common upper bound N of n, m in (\mathbb{N}, \preceq) , we have

$$W_\infty(a[x] + b[y]) = [W_N(a\varphi_{N,n}(x) + b\varphi_{N,m}(y))] = aW_\infty([x]) + bW_\infty([y]).$$

By continuity of W_∞ , it remains linear on $\overline{V_\infty}$. The result follows since for linear operators, Lipschitz continuity and continuity are equivalent. \square

D.2 Convergence, transferability and stability

Stability. The following stability result states that small perturbations of the input (e.g., adding a small number of nodes to a graph) lead to small changes in the output. It resembles the stability considered in [56].

1198 **Proposition D.3** (Stability: detailed version of Proposition 3.2). *If the sequence of maps $(f_n: V_n \rightarrow$
1199 $U_n)$ is $L(r)$ -locally Lipschitz transferable, then for any two inputs $x_n \in V_n$ and $x_m \in V_m$ of any
1200 two sizes n, m with $d_n^V(0, x_n), d_m^V(0, x_m) \leq r$, we have*

$$d_\infty^U([f_n(x_n)], [f_m(x_m)]) \leq L d_\infty^V([x_n], [x_m]).$$

1201 *Moreover, by Proposition D.1, the same holds when replacing every d_∞ with the symmetrized metric*
1202 \bar{d} .

1203 **Convergence and transferability: deterministic sampling.** For a sequence of inputs (x_n) sampled
1204 (deterministically) from the same underlying limiting object x , the outputs of a transferable function
1205 satisfy $f_n(x_n) \approx f_m(x_m)$ for big n, m , and converge as $f_n(x_n) \rightarrow f_\infty(x)$. We provide examples of
1206 sampling procedures later in Appendix D.3.

1207 **Proposition D.4** (Convergence and transferability: detailed version of Proposition 3.2). *Let $(x_n \in$
1208 $V_n)_{n \in \mathbb{N}}$ be a sequence of inputs sampled from a limiting object $x \in \overline{V_\infty}$, such that $[x_n] \rightarrow x$ at rate
1209 $R(n)$ with respect to d_∞^V .*

1210 (1) **(Asymptotic)** *If the sequence of maps $(f_n: V_n \rightarrow U_n)_{n \in \mathbb{N}}$ is continuously transferable, then*

1211 *(Convergence) $[f_n(x_n)] \rightarrow f_\infty(x)$ with respect to d_∞^U .*

1212 *(Transferability) $d_\infty^U([f_n(x_n)], [f_m(x_m)]) \rightarrow 0$ as $n, m \rightarrow \infty$.*

1213 (2) **(Nonasymptotic)** *Moreover, if $(f_n: V_n \rightarrow U_n)_{n \in \mathbb{N}}$ is locally Lipschitz transferable at x , then*

1214 *(Convergence) $[f_n(x_n)] \rightarrow f_\infty(x)$ at rate $R(n)$ with respect to d_∞^U .*

1215 *(Transferability) $d_\infty^U([f_n(x_n)], [f_m(x_m)]) \lesssim R(n) + R(m)$.*

1216 *That is, the Lipschitz case additionally provides quantitative guarantees for the convergence*
1217 *rate.*

1218 **Remark D.5.** *By Proposition D.1, the same holds when replacing every d_∞ with the symmetrized*
1219 *metric \bar{d} .*

1220 *Proof.* (Convergence \Rightarrow Transferability) Notice the transferability result directly follows from con-
1221 vergence by the triangular inequality

$$d_\infty([f_n(x_n)], [f_m(x_m)]) \leq d_\infty([f_n(x_n)], f_\infty(x)) + d_\infty([f_m(x_m)], f_\infty(x)).$$

1222 (Convergence) (1) If f_∞ is continuous, then $[f_n(x_n)] = f_\infty([x_n]) \rightarrow f_\infty(x)$ immediately follows.

1223 (2) Suppose f_∞ is locally Lipschitz at x . Then there exists $r > 0$ such that for all $y \in B(x, r)$, we
1224 have $d_\infty(f_\infty(x), f_\infty(y)) \leq L d_\infty(x, y)$. Let N be large enough so that $x_n \in B(x, r)$ for all $n \geq N$.
1225 Then for all $n \geq N$, we have

$$d_\infty(f_\infty(x), [f_n(x_n)]) = d_\infty(f_\infty(x), f_\infty([x_n])) \leq L \cdot d_\infty(x, [x_n]) \lesssim R(n),$$

1226 as claimed. □

1227 **Convergence and transferability: random sampling.** Under random sampling of inputs, we need
1228 to specify the mode of convergence. In the case where $x_n \rightarrow x$ almost surely at rate $R(n)$, the
1229 results are identical to the deterministic case: both convergence and transferability hold almost surely.
1230 We now consider a different mode of convergence—convergence in expectation. As we will see in
1231 Appendix D.3, many common sampling procedures satisfy this condition.

1232 **Proposition D.6** (Convergence and transferability: Random sampling). *Let $(x_n \in V_n)$ be a sequence*
1233 *of inputs randomly sampled from a limiting object $x \in \overline{V_\infty}$, such that $[x_n] \rightarrow x$ in expectation at rate*
1234 *$R(n)$ with respect to d_∞^V , i.e. $\mathbb{E}[d_\infty^V([x_n], x)] \lesssim R(n)$ and $R(n) \rightarrow 0$.*

1235 *If $(f_n: V_n \rightarrow U_n)$ is locally Lipschitz transferable at x , so f_∞ is L -Lipschitz on $B(x, r)$, and if there*
1236 *exists $M > 0$ such that $\mathbb{E}[d_\infty(f_\infty([x_n]), f_\infty(x)) \mathbb{1}([x_n] \notin B(x, r))] \leq M \mathbb{E}[d_\infty([x_n], x)]$, then*

1237 *(Convergence) $[f_n(x_n)] \rightarrow f_\infty(x)$ in expectation at rate $R(n)$ with respect to d_∞^U , i.e.*
1238 $\mathbb{E}[d_\infty^U([f_n(x_n)], f_\infty(x))] \lesssim R(n) \rightarrow 0$.

1239 (Transferability) $\mathbb{E}[d_\infty^U([f_n(x_n)], [f_m(x_m)])] \lesssim R(n) + R(m) \rightarrow 0$.

1240 **Remark D.7.** The assumptions are satisfied under any of the following conditions:

1241 (1) (f_n) is globally Lipschitz transferable; or

1242 (2) f_∞ is bounded; or

1243 (3) (x_n) are supported in $B(x, r)$.

1244 Furthermore, the same conclusion remains valid when replacing every d_∞ with the symmetrized
1245 metric \bar{d} .

1246 *Proof.* Suppose for all $y \in B(x, r)$, we have $d_\infty(f_\infty(x), f_\infty(y)) \leq Ld_\infty(x, y)$. Then,

$$\begin{aligned} \mathbb{E}[d_\infty(f_\infty(x), [f_n(x_n)])] &\leq \mathbb{E}[d_\infty(f_\infty(x), f_\infty([x_n])) \cdot \mathbb{1}\{[x_n] \in B(x, r)\}] + M\mathbb{E}[d_\infty(x, [x_n])] \\ &\leq L \cdot \mathbb{E}[d_\infty(x, [x_n]) \cdot \mathbb{1}\{[x_n] \in B(x, r)\}] + M\mathbb{E}[d_\infty(x, [x_n])] \\ &\lesssim R(n). \end{aligned}$$

1247 Transferability then follows by the triangle inequality. \square

1248 D.3 Convergence rates under sampling

1249 Propositions D.4 and D.6 show that for Lipschitz transferable models (f_n) , the convergence of $f_n(x_n)$
1250 to $f_\infty(x)$ is at least as fast as the convergence of x_n to x , characterized by the rate $R(n)$. This rate
1251 depends on the specific application and sampling scheme. Below, we review common sampling
1252 schemes and their associated convergence rates from the literature.

1253 Random sampling.

1254 (Empirical distributions) Suppose $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $X \in \mathbb{R}^{n \times d}$ has rows sampled i.i.d. from μ .
1255 Let $\mu_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ be the corresponding (empirical) distribution on \mathbb{R}^d . Then whenever
1256 $p \in [1, \infty]$ we have [20]:

$$\mathbb{E}[W_p(\mu, \mu_X)] \lesssim \begin{cases} n^{-1/2p} & \text{if } p > d/2, \\ n^{-1/2p} \log^{1/p}(1+n) & \text{if } p = d/2, \\ n^{-1/d} & \text{if } p < d/2. \end{cases} \quad (2)$$

1257 (Point clouds) Suppose $\mu \in \mathcal{P}_p(\mathbb{R}^k)$ and $X \in \mathbb{R}^{n \times k}$ has rows sampled i.i.d. from μ . Let $G \in \mathcal{O}(k)$
1258 be a random (or deterministic) rotation, sampled independently of X , and consider the rotated
1259 point cloud XG . Then the expected symmetrized metric between μ_{XG} and μ can be bounded
1260 by (2) since

$$\begin{aligned} \mathbb{E} \left[\inf_{g \in \mathcal{O}(k)} W_p(\mu, \mu_{XGg^{-1}}) \right] &= \mathbb{E} \left\{ \mathbb{E} \left[\inf_{g \in \mathcal{O}(k)} W_p(\mu, \mu_{XGg^{-1}}) \middle| G \right] \right\} \\ &\leq \mathbb{E}[W_p(\mu, \mu_X)]. \end{aligned} \quad (3)$$

1261 (Graphons) Let $W: [0, 1]^2 \rightarrow [0, 1]$ and $A_n \in \mathbb{R}_{\text{sym}}^{n \times n}$ be sampled as $(A_n)_{i,j} \sim \text{Ber}(W(x_i, x_j))$
1262 where x_1, \dots, x_n are i.i.d. uniform $[0, 1]$. Let W_{A_n} be the step graphon associated to A_n .
1263 Then [44, §10.4] implies⁷

$$\mathbb{E}[\delta_\square(W, W_{A_n})] \lesssim \frac{1}{\sqrt{\log(n)}}.$$

1264 where δ_\square is the cut distance of graphons (See [44, §8.2.2]). Moreover, we have $W_{A_n} \rightarrow W$ in cut
1265 metric almost surely [44, Cor. 11.15].

1266 Similarly, if $T_W: L^2([0, 1]) \rightarrow L^2([0, 1])$ is the integral operator associated with W , then

$$\mathbb{E}[\|T_W - T_{W_{A_n}}\|_{\text{op}}] \leq \sqrt{2}\mathbb{E}[\delta_\square(W, W_{A_n})]^{1/2} \lesssim (\log n)^{-1/4}.$$

⁷In fact, this bound holds with probability at least $1 - \exp(-\frac{n}{2 \log n})$.

1267 (*Signals*) Suppose we have a signal $f \in L^p([0, 1])$ which we sample by $x_i = f(t_i)$ where t_1, \dots, t_n
 1268 are i.i.d. uniform $[0, 1]$. Let f_n be the step function corresponding to x . Then [37, Thm. 3.4]
 1269 shows that

$$\mathbb{E}[\bar{d}(f, f_n)] = \mathbb{E} \left[\inf_{\sigma \in G_\infty} \|f - \sigma f_n\| \right] \lesssim \frac{1}{\sqrt{\log n}},$$

1270 where $\|\cdot\| = \|\cdot\|_\square$ or $\|\cdot\|_1$.

1271 **Deterministic sampling.**

1272 (*Uniform grid*) Suppose $f: [0, 1]^k \rightarrow \mathbb{R}$ is L -Lipschitz with respect to $\|\cdot\|_p$, and consider its values
 1273 on a uniform grid $X_{i_1, \dots, i_k} = f((i_1 - 1)/n, \dots, (i_k - 1)/n) \in (\mathbb{R}^n)^{\otimes k}$. If we extend X to a step
 1274 function as usual by $f_X(x_1, \dots, x_k) = X_{\lceil x_1 n \rceil, \dots, \lceil x_k n \rceil}$, then

$$\begin{aligned} \|f - f_X\|_q &\leq \|f - f_X\|_\infty \\ &\leq L \sup_{x_1, \dots, x_k \in [0, 1]} \|(x_1, \dots, x_k) - ((\lceil x_1 n \rceil - 1)/n, \dots, (\lceil x_k n \rceil - 1)/n)\|_p \\ &= L \|(1/n, \dots, 1/n)\|_p = \frac{Lk^{1/p}}{n}, \end{aligned}$$

1275 for all $q \in [1, \infty]$. Also note that if we evaluate an L -Lipschitz graphon $W: [0, 1]^2 \rightarrow [0, 1]$ on
 1276 such a uniform grid, we have

$$\|T_W - T_{W_n}\|_{\text{op}} \leq \|W - W_n\|_2 \leq \frac{L\sqrt{2}}{n}.$$

1277 (*Local averaging*) For any $f \in L^p([0, 1]^k)$, we can locally average it over hypercubes of side length
 1278 $1/n$ to produce values

$$X_{i_1, \dots, i_k} = n^k \int_{(i_1-1)/n}^{i_1/n} \cdots \int_{(i_k-1)/n}^{i_k/n} f(x_1, \dots, x_k) dx_1 \cdots dx_k,$$

1279 and again extend these values to a step function f_X . In this case,

$$\|f - f_X\|_p \leq 2 \text{dist}(f, V_n),$$

1280 so we get the optimal rate of convergence.

1281 **E Generalization bounds: details and missing proofs from Section 4**

1282 We apply the framework connecting robustness and generalization established by [70], which is
 1283 built on the idea that algorithmic robustness—that is, a model’s stability to input perturbations—is
 1284 fundamentally linked to its ability to generalize. We refer readers to [70] for the necessary background.
 1285 This framework has also been recently employed to derive generalization bounds for GNNs in [62],
 1286 though their analysis is restricted to graphs with bounded size. Similar techniques are used in [37, 54].

1287 We consider an any-dimensional supervised learning task where consistent sequences model the input
 1288 and output space

$$\mathbb{V} = \{(V_n), (\varphi_{N,n}), (G_n)\} \quad \text{and} \quad \mathbb{U} = \{(U_n), (\psi_{N,n}), (G_n)\},$$

1289 with associated symmetrized metrics \bar{d}_V and \bar{d}_U . The dataset s consists of N input-output pairs
 1290 $(x_i, y_i) \in X \times Y \subseteq V_\infty \times U_\infty$, where $X \times Y$ are subsets whose sequence of orbit closures are
 1291 compact in the symmetrized metrics. More precisely, (x_i, y_i) are finite-dimensional representatives
 1292 of equivalence classes in $V_\infty \times U_\infty$. The hypothesis class \mathcal{H} consists of functions $\bar{V}_\infty \rightarrow \bar{U}_\infty$
 1293 parametrized by neural networks. A learning algorithm \mathcal{A} is a mapping

$$\mathcal{A}: (V_\infty \times U_\infty)^N \rightarrow \mathcal{H}.$$

1294 We write \mathcal{A}_s for the hypothesis learned from the dataset s .

1295 Assume training is performed using a neural network model that is $L(r)$ -locally Lipschitz transferable.
 1296 (Recall from Proposition D.1 that this implies \mathcal{A}_s is $L(r)$ -locally Lipschitz on $B(0, r)$ for all $r > 0$)

with respect to the symmetrized metrics \bar{d}_V and \bar{d}_U .) Since $X \times Y$ is compact, and hence bounded, there exists a constant $c_s > 0$ such that $\mathcal{A}_s: \bar{V}_\infty \rightarrow \bar{U}_\infty$ is c_s -Lipschitz on $X \times Y$ with respect to the symmetrized metrics. Further, let the loss function $\ell: \bar{U}_\infty \times \bar{U}_\infty \rightarrow \mathbb{R}$ be bounded by M , and c_ℓ -Lipschitz with respect to the product metric $d((x, y), (x', y')) := \bar{d}_U(x, x') + \bar{d}_U(y, y')$. By applying the framework of [70] to the limit space, we immediately obtain a generalization bound for learning tasks where the data consists of inputs of varying dimensions. We note that this is not the result stated in Proposition 4.2 of the main paper; the version claimed there will be established later in Proposition E.3.

Proposition E.1 (Any-dimensional generalization bound). *Assume that the training data consists of N i.i.d. samples $s = (x_i, y_i) \sim \hat{\mu}$ from a measure $\hat{\mu}$ supported on $X \times Y \subseteq V_\infty \times U_\infty$, where X and Y have finite ϵ -covering numbers $C_X(\epsilon), C_Y(\epsilon)$ with respect to the symmetrized metrics for all $\epsilon > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the generalization error satisfies*

$$\left| \frac{1}{N} \sum_{i=1}^N l(\mathcal{A}_s(x_i), y_i) - \mathbb{E}_{(x,y) \sim \hat{\mu}} l(\mathcal{A}_s(x), y) \right| \leq \inf_{\gamma > 0} \left(c_\ell(c_s \vee 1)\gamma + M \sqrt{\frac{2C_X(\gamma/4)C_Y(\gamma/4) \log 2 + 2 \log(1/\delta)}{N}} \right) \quad (4)$$

$$\leq c_\ell(c_s \vee 1)\xi^{-1}(N) + M \sqrt{(2 \log 2)\xi^{-1}(N)^2 + \frac{2 \log(1/\delta)}{N}}, \quad (5)$$

where $\xi(r) := \frac{C_X(r/4)C_Y(r/4)}{r^2}$ and we set $\gamma = \xi^{-1}(N)$ in the second line to obtain the third.

Remark E.2. We make the following observations:

- (1) The bound (5) converges to 0 as $N \rightarrow \infty$. Indeed, ξ is strictly decreasing, and hence its inverse is well-defined and also strictly decreasing. Since $\xi(x) \rightarrow \infty$ as $x \rightarrow 0^+$, we get $\xi^{-1}(x) \rightarrow 0$ as $x \rightarrow \infty$.
- (2) The generalization bound reveals that the ability to generalize improves with greater model transferability/stability (i.e., smaller Lipschitz constants), and deteriorates with increasing geometric complexity of the data space (i.e., larger covering numbers).
- (3) We emphasize that $\hat{\mu}$ is a distribution on $V_\infty \times U_\infty$, ensuring that every sample drawn from $\hat{\mu}$ admits a finite-dimensional representative, i.e., $(x_i, y_i) \in V_n \times U_n$ for some n . This reflects the realistic setting in which data consists of finite-dimensional inputs. This stands in contrast to prior work on GNN generalization bounds [37, 54], which considers data distributions on \bar{V}_∞ —the space of graphon signals in [38], and the space of iterated degree measures in [54]. Such an assumption is somewhat unrealistic, as many elements in these spaces cannot be realized as finite-dimensional data.
- (4) Note that $\hat{\mu}$ induces a distribution $(\hat{\mu}(V_n \times U_n))_{n \in \mathbb{N}}$ over sample dimensions in \mathbb{N} , which inherently places less weight on larger sizes. Consequently, the generalization bound does not offer guarantees on the asymptotic performance of the model as the input dimension $n \rightarrow \infty$. Next, we will derive the second generalization bound that addresses this problem.

Proof. For all $(x_1, y_1), (x_2, y_2) \in X \times Y$,

$$\begin{aligned} |l(\mathcal{A}_s(x_1), y_1) - l(\mathcal{A}_s(x_2), y_2)| &\leq c_\ell(c_s \bar{d}_V(x_1, x_2) + \bar{d}_U(y_1, y_2)) \\ &\leq c_\ell(c_s \vee 1)(\bar{d}_V(x_1, x_2) + \bar{d}_U(y_1, y_2)). \end{aligned}$$

Apply [70, Theorem 14], the algorithm \mathcal{A} is $(C_X(\gamma/4)C_Y(\gamma/4), \gamma c_\ell(c_s \vee 1))$ -robust [70, Definition 2] for all $\gamma > 0$.

Apply [70, Theorem 3], the generalization bound in the form of (4) follows immediately. (5) follows by taking the γ as defined. \square

The previous generalization bound follows the classical statistical learning setup, where both training and test data are assumed to be sampled i.i.d. from the same distribution. However, in any-dimensional learning, we are typically concerned with a different scenario: training on data of smaller sizes and

1336 testing on data of larger sizes. This motivates the need for a new form of generalization bound that
 1337 accounts for such settings. We propose the following set-up (described in the main paper):

1338 Let μ be a probability distribution supported on $X \times Y \subseteq \overline{V_\infty} \times \overline{U_\infty}$, which are subsets whose
 1339 sequence of orbit closures is compact in the symmetrized metrics. Consider a random sampling
 1340 procedure

$$\mathcal{S}_n : \overline{V_\infty} \times \overline{U_\infty} \rightarrow V_n \times U_n,$$

1341 such that for all n and for all $(x, y) \in \text{supp}(\mu)$, we have $\text{supp}(\mathcal{S}_n(x, y)) \subseteq X \times Y$. This sampling
 1342 induces a distribution μ_n on $V_n \times U_n$ via the sampling procedure; that is,

$$\mu_n := \text{Law}(x_n, y_n), \quad \text{where } (x_n, y_n) \sim \mathcal{S}_n(x, y) \text{ and } (x, y) \sim \mu.$$

1343 **Proposition E.3** (Size-generalization bound). *Suppose the training data consists of N i.i.d. samples
 1344 $s = (x_i, y_i) \sim \mu_n$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the generalization error
 1345 satisfies*

$$\begin{aligned} & \left| \frac{1}{N} \sum_{i=1}^N \ell(\mathcal{A}_s(x_i), y_i) - \mathbb{E}_{(x,y) \sim \mu} \ell(\mathcal{A}_s(x), y) \right| \\ & \leq c_\ell(c_s \vee 1) (\xi^{-1}(N) + W_1(\mu, \mu_n)) + M \sqrt{(2 \log 2) \xi^{-1}(N)^2 + \frac{2 \log(1/\delta)}{N}}, \end{aligned} \quad (6)$$

1346 where W_1 denotes the Wasserstein-1 distance, and $\xi(r) := \frac{C_X(r/4)C_Y(r/4)}{r^2}$, with $C_X(\epsilon)$ and $C_Y(\epsilon)$
 1347 denoting the ϵ -covering numbers of X and Y , respectively, with respect to the symmetrized metrics.

1348 Moreover, if the sampling has convergence rate $R(n)$ in expectation, i.e.,

$$\mathbb{E}_{(x_n, y_n) \sim \mathcal{S}_n(x, y)} [\bar{d}_V(x, [x_n]) + \bar{d}_U(y, [y_n])] \lesssim R(n) \rightarrow 0,$$

1349 then the bound satisfies

$$\begin{aligned} & \left| \frac{1}{N} \sum_{i=1}^N \ell(\mathcal{A}_s(x_i), y_i) - \mathbb{E}_{(x,y) \sim \mu} \ell(\mathcal{A}_s(x), y) \right| \\ & \lesssim c_\ell(c_s \vee 1) (\xi^{-1}(N) + R(n)) + M \sqrt{(2 \log 2) \xi^{-1}(N)^2 + \frac{2 \log(1/\delta)}{N}}. \end{aligned} \quad (7)$$

1350 **Remark E.4.** We make the following observations:

1351 (1) The bound (7) converges to 0 if both the training input dimension n and the amount of data N
 1352 goes to ∞ .

1353 (2) This new generalization bound aligns with the setup where training is performed on inputs of
 1354 fixed size n (also naturally extends to inputs of varying finite sizes), and testing evaluates the
 1355 asymptotic performance as $n \rightarrow \infty$. It can therefore be interpreted as a size generalization
 1356 bound, accounts for distributional shifts induced by size variation. As a consequence, an
 1357 additional term appears in the bound, reflecting the convergence rate of the sampling procedure.

1358 *Proof.* By triangle inequality,

$$\begin{aligned} \left| \frac{1}{N} \sum_{i=1}^N \ell(\mathcal{A}_s(x_i), y_i) - \mathbb{E}_{(x,y) \sim \mu} \ell(\mathcal{A}_s(x), y) \right| & \leq \underbrace{\left| \frac{1}{N} \sum_{i=1}^N \ell(\mathcal{A}_s(x_i), y_i) - \mathbb{E}_{(x,y) \sim \mu_n} \ell(\mathcal{A}_s(x), y) \right|}_{(A)} \\ & \quad + \underbrace{\left| \mathbb{E}_{(x,y) \sim \mu} \ell(\mathcal{A}_s(x), y) - \mathbb{E}_{(x,y) \sim \mu_n} \ell(\mathcal{A}_s(x), y) \right|}_{(B)}. \end{aligned}$$

1359 By the Kantorovich-Rubinstein duality, we have almost surely that

$$(B) \leq c_\ell(c_s \vee 1) \cdot W_1(\mu, \mu_n).$$

For (A), recall that $\text{supp}(\mathcal{S}_n(x, y)) \subseteq X \times Y$ for all $(x, y) \in \text{supp}(\mu)$. It follows that $\text{supp}(\mu_n) \subseteq X \times Y$, which is therefore totally bounded. Its covering number is upper bounded by that of $X \times Y$. Applying Proposition E.1 with $\hat{\mu} = \mu_n$ yields (6).

To bound $W_1(\mu, \mu_n)$, we note that the sampling procedure induces a natural coupling between μ and μ_n . Using this coupling,

$$W_1(\mu, \mu_n) \leq \mathbb{E}_{(x,y) \sim \mu, (x_n, y_n) \sim \mathcal{S}_n(x,y)} [\bar{d}_V(x, [x_n]) + \bar{d}_U(y, [y_n])] \lesssim R(n),$$

which yields (7). \square

This bound captures more realistic learning settings, as illustrated in the following examples.

Example E.5. Consider $\mathbb{V}_{\text{dup}}^{\oplus d}$, the duplication consistent sequence for sets endowed with the normalized ℓ_p metric. See Appendix F.1 for the precise definitions. The limit space is $\bar{V}_\infty = L^p([0, 1], \mathbb{R}^d)$, and the space of orbit closures of \mathbb{V} is $\mathcal{P}_p(\mathbb{R}^d)$, the space of probability measures on \mathbb{R}^d with finite p -th moment, endowed with the Wasserstein- p distance.

Fix a compact set $\Omega \subseteq \mathbb{R}^d$, and let

$$X := \{f \in L^p([0, 1], \mathbb{R}^d) : \mu_f \text{ is supported on } \Omega\}.$$

Note that the sequence of orbit closures in X , namely $\{\mu_f : f \in X\}$, is compact with respect to W_p . A bound on the covering number C_X is given by [51, Theorem 2.2.11].

Consider the sampling procedure $\mathcal{S}_n : L^p([0, 1], \mathbb{R}^d) \rightarrow \mathbb{R}^{n \times d}$ defined by drawing $z_i \stackrel{i.i.d.}{\sim} \text{Unif}([0, 1])$, and setting $\mathcal{S}_n(f)_{i:} = f(z_i)$ for $i = 1, \dots, n$. Then, for all $f \in X$, each entry of $\mathcal{S}_n(f)$ lie in Ω . Hence, we have $\mathcal{S}_n(f) \in X$. Our generalization bound therefore applies to this setting.

Example E.6. Consider $\mathbb{V}_{\text{dup}}^G$, the duplication-consistent sequence for graph signals endowed with the cut metrics. See Appendix G.1 for the precise definitions. Define the space

$$X = \{W : [0, 1]^2 \rightarrow [0, 1] \text{ measurable} : W(x, y) = W(y, x)\} \times \{f \in L^\infty([0, 1], \mathbb{R}) : \|f\|_\infty \leq r\}.$$

By [37, Theorem 3], the sequence of orbit closures in X is compact with respect to the cut metric on graphon signals. Moreover, a bound on the covering number is also provided in the same result.

Consider the sampling procedure $\mathcal{S}_n : X \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^n$ defined as follows: draw $z_i \stackrel{i.i.d.}{\sim} \text{Unif}([0, 1])$, and set $\mathcal{S}_n(W, f) = (A, X)$, where $A_{ij} \sim \text{Ber}(W(z_i, z_j))$ and $X_i = f(z_i)$ for $i = 1, \dots, n$. Then, for all $(W, f) \in X$, the sampled pair $\mathcal{S}_n(W, f)$ belongs to X . This is the standard sampling procedure for graphon signals, and once again our generalization bound applies.

E.1 Transferable neural networks

To prove the transferability of a neural network, we first observe that compatibility and transferability are preserved under composition. Therefore, it suffices to verify these properties for each individual layer. Moreover, by Propositions C.6 and D.2, it is enough to prove the compatibility and Lipschitz continuity of each f_n on the finite space, rather than analyzing the limiting function f_∞ directly.

This idea is formalized in the following proposition, which serves as a key tool in our transferability analysis of neural networks in the later sections. Importantly, this provides a general and easy-to-apply proof strategy. In contrast, previous works often begin by characterizing a natural limiting function f_∞ (e.g., a graphon neural network), and then directly prove its Lipschitz continuity in the limit space—a process that typically requires case-specific proof techniques.

Proposition E.7 (Transferable networks: detailed version of Proposition 5.1). *Let $(V_n^{(i)})_n, (U_n^{(i)})_n$ be consistent sequences for $i = 1, \dots, D$. For each i , let $(W_n^{(i)} : V_n^{(i)} \rightarrow U_n^{(i)})$ be linear maps and $(\rho_n^{(i)} : U_n^{(i)} \rightarrow V_n^{(i+1)})$ be nonlinearities. Assume the following properties hold:*

- (1) *The maps $(W_n^{(i)}), (\rho_n^{(i)})$ are compatible.*
- (2) *The linear maps are uniformly bounded $\sup_{n,i} \|W_n^{(i)}\|_{\text{op}} = L_W < \infty$.*
- (3) *The map $\rho_n^{(i)}$ is $L_i(r)$ -Lipschitz on $\{u \in U_n^{(i)} : \|u\| < r\}$ for all n .*

1401 Then the composition $(W_n^{(D)} \circ \rho_n^{(D-1)} \circ \dots \circ \rho_n^{(1)} \circ W_n^{(1)})$ is locally Lipschitz transferable, ex-
 1402 tending to a function on $\overline{V_\infty} \rightarrow \overline{U_\infty}$ that is $L_{\text{NN}}(r)$ -Lipschitz on $\{v \in \overline{V_\infty^{(1)}} : \|v\| < r\}$, where we
 1403 inductively define

$$\ell_1 = L_1(L_W r), \quad \ell_{i+1} = L_{i+1}(L_W^{i+1} \ell_i), \quad L_{\text{NN}}(r) = L_W^D \prod_{i=1}^{D-1} \ell_i. \quad (8)$$

1404 In particular, if $\rho_n^{(i)}$ is L_ρ -Lipschitz for all i, n then the composition is Lipschitz transferable, extending
 1405 to a function on $\overline{V_\infty} \rightarrow \overline{U_\infty}$ that is L_{NN} -Lipschitz where $L_{\text{NN}} = L_W^D L_\rho^{D-1}$.

1406 **Remark E.8.** By Proposition D.1, the composition is also $L_{\text{NN}}(r)$ -Lipschitz with respect to the
 1407 symmetrized metrics on the same r -ball.

1408 *Proof.* Note that if $f_1: \overline{V_\infty^{(1)}} \rightarrow \overline{V_\infty^{(2)}}$ and $f_2: \overline{V_\infty^{(2)}} \rightarrow \overline{V_\infty^{(3)}}$ are $L_1(r)$ - and $L_2(r)$ -locally Lipschitz,
 1409 respectively, then $f_2 \circ f_1$ is $L_2(L_1(r))L_1(r)$ -locally Lipschitz. Our claim follows by an inductive
 1410 application of this fact and Proposition D.2. \square

1411 F Example 1 on sets: details and missing proofs from Section 5.1

1412 F.1 Consistent sequences on sets

1413 Zero-padding consistent sequence \mathbb{V}_{zero} with ℓ_p norm

1414 The zero-padding consistent sequence $\mathbb{V}_{\text{zero}} = \{(V_n), (\varphi_{N,n}), (G_n)\}$ is defined as follows: The
 1415 index set $\mathbb{N} = (\mathbb{N}, \leq)$ is the poset of natural numbers with the standard ordering. Let $V_n = \mathbb{R}^n$ for
 1416 every $n \in \mathbb{N}$, and the zero-padding embedding is given by, for $n \leq N$,

$$\begin{aligned} \varphi_{N,n}: \mathbb{R}^n &\hookrightarrow \mathbb{R}^N \\ (x_1, \dots, x_n) &\mapsto (x_1, \dots, x_n, \underbrace{0, \dots, 0}_{(N-n) \text{ 0's}}). \end{aligned}$$

1417 The group of permutations S_n on n letters acts on \mathbb{R}^n by permuting coordinates: $(g \cdot x)_i := x_{g^{-1}(i)}$
 1418 for $g \in S_n$. The embedding of groups is given by, for $n \leq N$,

$$\begin{aligned} \theta_{N,n}: S_n &\hookrightarrow S_N \\ g &\mapsto \begin{bmatrix} g & 0 \\ 0 & I_{N-n} \end{bmatrix}. \end{aligned}$$

1419 That is, view S_n as the subgroup of S_N which acts trivially on $n+1, \dots, N$.

1420 In this case, V_∞ can be identified as \mathbb{R}^∞ , the space of infinite sequences with finitely many nonzero
 1421 entries. The limit group G_∞ is the group of permutations of \mathbb{N} fixing all but finitely many indices. The
 1422 orbit space of V_∞ under the action of G_∞ can be identified with infinite sequences with finitely-many
 1423 nonzero entries, where such entries are ordered.

1424 ℓ_p norm on \mathbb{V}_{zero} . We can endow each V_n with the ℓ_p norms

$$\|x\|_p = \begin{cases} (\sum_{i=1}^n |x_i|^p)^{1/p} & \text{if } p \in [1, \infty), \\ \max_{i=1}^n |x_i| & \text{if } p = \infty. \end{cases}$$

1425 It is easy to check by Proposition C.8 that this induces a norm on $V_\infty = \mathbb{R}^\infty$, which coincides with
 1426 the ℓ_p norm on infinite sequences. The limit space is then

$$\overline{V_\infty} = \begin{cases} \ell_p = \{(x_i)_{i=1}^\infty : \sum_{i=1}^\infty |x_i|^p < \infty\} & \text{if } p \in [1, \infty), \\ c_0 = \{(x_i)_{i=1}^\infty : \lim_{i \rightarrow \infty} |x_i| = 0\} & \text{if } p = \infty. \end{cases}$$

1427 Hence, the space of orbit closures can be identified with sorted sequences in ℓ_p (for $p \in [1, \infty)$) or c_0
 1428 (for $p = \infty$). This space is endowed with the symmetrized metric

$$\bar{d}(x, y) = \begin{cases} (\sum_{i=1}^\infty |x_{(i)} - y_{(i)}|^p)^{1/p} & \text{if } p \in [1, \infty), \\ \sup_i |x_{(i)} - y_{(i)}| & \text{if } p = \infty, \end{cases}$$

1429 where $x_{(i)}$ is the i -th largest entry in the sequence (x_i) .

1430 ℓ_p norm on the direct sum $\mathbb{V}_{\text{zero}}^{\oplus d}$. The direct sum $\mathbb{V}_{\text{zero}}^{\oplus d} = \{(\mathbb{R}^{n \times d}), (\varphi_{N,n}^{\oplus d}), (S_n)\}$ defined in
 1431 Definition C.4 extends the above to the case of a set of vectors in \mathbb{R}^d . To endow it with a norm, we
 1432 first fix an arbitrary norm $\|\cdot\|_{\mathbb{R}^d}$ on \mathbb{R}^d . Then the ℓ_p -norm on $\mathbb{R}^{n \times d}$ is defined analogously with
 1433 respect to $\|\cdot\|_{\mathbb{R}^d}$, i.e., for $X \in \mathbb{R}^{n \times d}$,

$$\|X\|_p = \begin{cases} (\sum_{i=1}^n \|X_{i,:}\|_{\mathbb{R}^d}^p)^{1/p} & \text{if } p \in [1, \infty), \\ \max_{i=1}^n \|X_{i,:}\|_{\mathbb{R}^d} & \text{if } p = \infty. \end{cases}$$

1434 The definitions of the limit space and symmetrized metric are analogous to the above, replacing $|\cdot|$
 1435 with $\|\cdot\|_{\mathbb{R}^d}$.

1436 Duplication consistent sequence \mathbb{V}_{dup} with normalized ℓ_p norm

1437 The duplication consistent sequence $\mathbb{V}_{\text{dup}} = \{(V_n), (\varphi_{N,n}), (G_n)\}$ is defined as follows. The index
 1438 set $\mathbb{N} = (\mathbb{N}, \cdot | \cdot)$ is the set of natural numbers with divisibility partial order, where $n \preceq N$ if and
 1439 only if $n | N$. Let $V_n = \mathbb{R}^n$ for all $n \in \mathbb{N}$, and the duplication embeddings is given by

$$\begin{aligned} \varphi_{N,n}: \quad \mathbb{R}^n &\hookrightarrow \mathbb{R}^N \\ (x_1, \dots, x_n) &\mapsto x \otimes \mathbb{1}_{N/n} = \underbrace{(x_1, \dots, x_1, \dots, x_n, \dots, x_n)}_{N/n \text{ times}}, \end{aligned}$$

1440 for $n \preceq N$. The group embeddings are given by

$$\begin{aligned} \theta_{N,n}: \quad S_n &\hookrightarrow S_N \\ g &\mapsto g \otimes I_{N/n}, \end{aligned}$$

1441 for $n \preceq N$. That is, $g \in S_n$ acts on $[N]$ by sending $(i-1)N/n + j$ to $(g(i)-1)N/n + j$ for
 1442 $i = 1, \dots, n$ and $j = 1, \dots, N/n$.

1443 In this case, V_∞ can be identified with step functions on $[0, 1]$ whose discontinuity points are in \mathbb{Q} :
 1444 each $x \in \mathbb{R}^n$ corresponds to $f_x: [0, 1] \rightarrow \mathbb{R}$ where $f_x(t) = x_{\lceil tn \rceil}$ for $t > 0$ and $f(0) = x_1$. In other
 1445 words, f_x is the step function which takes value x_i on $I_{i,n} = (\frac{i-1}{n}, \frac{i}{n}]$ for $i = 1, \dots, n$. Indeed, all
 1446 equivalent objects x and $\varphi_{N,n}x$ correspond to the same function in this way. Therefore, V_∞ can be
 1447 seen as the union of step functions in this forms for $n \in \mathbb{N}$.

1448 Under this identification, permutations S_n permute the n intervals $I_{i,n}$ and act on functions by
 1449 $g \cdot f = f \circ g^{-1}$. The limit group G_∞ is the union of such interval permutations. The orbit space of
 1450 the G_∞ -action on V_∞ can be identified with monotonically increasing step functions on $[0, 1]$ whose
 1451 discontinuity points are in \mathbb{Q} .

1452 Alternatively, the orbit space can be identified with the space of empirical measures on \mathbb{R} : each
 1453 $x \in \mathbb{R}^n$ corresponds to the empirical measure $\mu_x = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Indeed, any equivalent objects x
 1454 and $\varphi_{N,n}x$ are identified with the same measure; furthermore, this resulting measure is constant on
 1455 orbits under the G_∞ -action.

1456 Under the identification of V_∞ with step functions, a step function $f \in V_\infty$ is identified with the
 1457 probability measure μ_f , defined as the distribution of $f(T)$ for T uniformly sampled from $[0, 1]$.
 1458 Indeed, all elements in the orbit of f under the G_∞ -action correspond to this same measure μ_f . Note
 1459 that the generalized inverse CDF of $f(T)$ is precisely the ‘sorted’ version of f (called its increasing
 1460 rearrangement), relating the above two perspectives on the orbit space. The latter view of the orbit
 1461 space as a sequence of measures generalizes readily to other consistent sequences obtained from
 1462 \mathbb{V}_{dup} , such as $\mathbb{V}_{\text{dup}}^{\oplus d}$ which we consider below, so we shall take this view from now.

1463 *Normalized ℓ_p norm on \mathbb{V}_{dup} .* We can endow each space V_n with the normalized ℓ_p norm

$$\|x\|_{\bar{p}} = \begin{cases} (\frac{1}{n} \sum_{i=1}^n |x_i|^p)^{1/p} & \text{if } p \in [1, \infty), \\ \max_{i=1}^n |x_i| & \text{if } p = \infty. \end{cases}$$

1464 Using Proposition C.8, it is straightforward to verify that this defines a norm on V_∞ . Under the
 1465 identification of V_∞ with step functions, the induced norm on V_∞ coincides with the conventional

1466 L^p norm on measurable functions, given by

$$\|f\|_p = \begin{cases} \left(\int_0^1 |f(t)|^p dt \right)^{1/p} & \text{if } p \in [1, \infty), \\ \sup_{t \in [0,1]} |f(t)| & \text{if } p = \infty. \end{cases}$$

1467 That is, for any $x \in \mathbb{R}^n$, we have $\|x\|_{\bar{p}} = \|f_x\|_p$, where f_x is the corresponding step function. The
1468 limit space is then

$$\overline{V_\infty} = \begin{cases} L^p([0, 1]) = \left\{ f: [0, 1] \rightarrow \mathbb{R} \text{ measurable} : \int_0^1 |f(t)|^p dt < \infty \right\} & \text{if } p \in [1, \infty), \\ \left\{ f: [0, 1] \rightarrow \mathbb{R} : \begin{array}{l} f \text{ is bounded and continuous on } [0, 1] \setminus \mathbb{Q}, \\ \text{with left and right limits at every } x \in [0, 1] \cap \mathbb{Q} \end{array} \right\} & \text{if } p = \infty, \end{cases}$$

1469 where the result for $p = \infty$ follows from [16, Chap. VII.6]. These are a subspace of so-called
1470 *regulated functions*, which have left and right limits at each $x \in [0, 1]$.

1471 When $p \in [1, \infty)$, the space of orbit closures, equipped with the symmetric metric, can be identified
1472 with $\mathcal{P}_p(\mathbb{R})$, the space of probability measures on \mathbb{R} with finite p -th moment, endowed with the
1473 Wasserstein p -distance. In the case $p = \infty$, the space of orbit closures corresponds to a subset of
1474 $\mathcal{P}_\infty(\mathbb{R})$, the space of probability measures on \mathbb{R} with bounded support, equipped with the Wasserstein
1475 ∞ -distance. This is formalized and proved by the following propositions:

1476 **Proposition F.1.** *For any $p \in [1, \infty]$ and all $f, g \in \overline{V_\infty}$, the symmetrized metric $\bar{d}_p(f, g) :=$
1477 $\inf_{\sigma \in G_\infty} \|\sigma \cdot f - g\|_p$ equals the Wasserstein p -distance between the associated measures:*

$$\bar{d}_p(f, g) = W_p(\mu_f, \mu_g).$$

1478 *Proof.* We first prove they match on V_∞ . Consider vectors $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ under the action of $S_n,$
1479 S_m respectively, and let $N = \text{lcm}(n, m)$. Then, by the standard results on the Wasserstein distance
1480 of empirical measures,

$$\bar{d}_p(x, y) = \begin{cases} \min_{\sigma \in S_N} \left(\frac{1}{N} \sum_{i=1}^N |\varphi_{N,n}(x)_{\sigma(i)} - \varphi_{N,m}(y)_i|^p \right)^{1/p} = W_p(\mu_x, \mu_y) & \text{if } p \in [1, \infty), \\ \min_{\sigma \in S_N} \max_{i=1}^N |\varphi_{N,n}(x)_{\sigma(i)} - \varphi_{N,m}(y)_i| = W_\infty(\mu_x, \mu_y) & \text{if } p = \infty. \end{cases}$$

1481 where μ_x was defined above, and W_p is the Wasserstein p -distance. Hence under the identification
1482 with step functions, for any $f, g \in V_\infty$ we have $\bar{d}(f, g) = W_p(\mu_f, \mu_g)$.

1483 Now consider the limit points. Let $(f_n), (g_n)$ be two Cauchy sequences in V_∞ with $f_n \rightarrow f, g_n \rightarrow g$
1484 in $\overline{V_\infty}$ with respect to the L^p norm. Then

$$|\bar{d}_p(f_n, g_n) - \bar{d}_p(f, g)| \leq \bar{d}_p(f, f_n) + \bar{d}_p(g_n, g) \leq \|f - f_n\|_p + \|g_n - g\|_p \rightarrow 0.$$

1485 Similarly, since for any $\tilde{f}, \tilde{g} \in \overline{V_\infty}$,

$$W_p(\mu_{\tilde{f}}, \mu_{\tilde{g}}) \leq \left(\mathbb{E}_{T \sim \text{Unif}[0,1]} |\tilde{f}(T) - \tilde{g}(T)|^p \right)^{1/p} = \|\tilde{f} - \tilde{g}\|_p \rightarrow 0,$$

1486 we also get

$$|W_p(f_n, g_n) - W_p(f, g)| \rightarrow 0.$$

1487 But for all n , $\bar{d}_p(f_n, g_n) = W_p(\mu_{f_n}, \mu_{g_n})$, so by the uniqueness of the limit, $\bar{d}_p(f, g) = W_p(\mu_f, \mu_g)$.
1488 \square

1489 **Proposition F.2.** *For $p \in [1, \infty)$, the space of orbit closures $\{\mu_f : f \in \overline{V_\infty}\}$ coincides with $\mathcal{P}_p(\mathbb{R})$.
1490 For $p = \infty$, this set is the subset of measures in $\mathcal{P}_\infty(\mathbb{R})$ with compact.*

1491 *Proof.* For $p \in [1, \infty)$, by definition, the space of orbit closures is the set of probability measures
1492 $\{\mu_f : f \in L^p([0, 1])\}$. We claim that this set is equal to $\mathcal{P}_p(\mathbb{R})$.

1493 Observe that $((\mathbb{E}_{X \sim \mu_f} |X|^p)^{1/p} = \|f\|_p$. On the one hand, this implies that if $f \in L^p([0, 1])$, then
1494 $\mu_f \in \mathcal{P}_p(\mathbb{R})$. Conversely, given any $\mu \in \mathcal{P}_p(\mathbb{R})$, let f be the generalized inverse of the CDF of μ ,
1495 then $\mu_f = \mu$ and $f \in L^p([0, 1])$. Hence $\mu \in \mathcal{P}_p(\mathbb{R})$ implies that $\mu = \mu_f$ for $f \in L^p([0, 1])$.

1496 For $p = \infty$, note that any $f \in \overline{V_\infty}$ is bounded, so the support of $\mu_f \in \mathcal{P}_\infty(\mathbb{R})$ is compact. Conversely,
1497 if $\mu \in \mathcal{P}_\infty(\mathbb{R})$ has compact support then its generalized inverse CDF f is bounded and satisfies
1498 $\mu = \mu_f$. \square

1499 *Norms on $V_{\text{dup}}^{\oplus d}$.* Similarly, we fix an arbitrary norm $\|\cdot\|_{\mathbb{R}^d}$ on \mathbb{R}^d and define the norms on $\mathbb{R}^{n \times d}$ with
 1500 respect to $\|\cdot\|_{\mathbb{R}^d}$:

$$\|X\|_{\bar{p}} := \begin{cases} \left(\frac{1}{n} \sum_{i=1}^n \|X_{i:}\|_{\mathbb{R}^d}^p\right)^{1/p} & \text{if } p \in [1, \infty), \\ \max_{i=1}^n \|X_{i:}\|_{\mathbb{R}^d} & \text{if } p = \infty. \end{cases} \quad (9)$$

1501 For $p \in [1, \infty)$, the space of orbit closures can be identified with $\mathcal{P}_p(\mathbb{R}^d)$ endowed with the Wasser-
 1502 stein p -distance with respect to $\|\cdot\|_{\mathbb{R}^d}$. For $p = \infty$, it can be seen as the subset of compactly-supported
 1503 measures in $\mathcal{P}_{\infty}(\mathbb{R}^d)$ with the Wasserstein- ∞ distance with respect to $\|\cdot\|_{\mathbb{R}^d}$.

1504 F.2 Invariant networks on sets

1505 We consider three prominent permutation-invariant neural network architectures for set-structured
 1506 data: DeepSets [73], normalized DeepSets [5], and PointNet [52]. These models are defined as
 1507 follows:

$$\text{DeepSet}_n(X) = \sigma \left(\sum_{i=1}^n \rho(X_{i:}) \right),$$

$$\overline{\text{DeepSet}}_n(X) = \sigma \left(\frac{1}{n} \sum_{i=1}^n \rho(X_{i:}) \right),$$

$$\text{PointNet}_n(X) = \sigma \left(\max_{i=1}^n \rho(X_{i:}) \right),$$

1508 where $\rho: \mathbb{R}^d \rightarrow \mathbb{R}^h$ and $\sigma: \mathbb{R}^h \rightarrow \mathbb{R}$ are multilayer perceptrons (MLPs). In the case of PointNet, the
 1509 maximum is taken entrywise over vectors in \mathbb{R}^h .

1510 They follow the same paradigm $f_n(X) = \sigma(\text{Agg}_{i=1}^n \rho(X_{i:}))$ where the three models use different
 1511 permutation-invariant aggregations Agg.

1512 We refer the reader to [5] for a comprehensive study of the expressive power of these models
 1513 in the any-dimensional setting. In particular, they show that normalized DeepSets (respectively,
 1514 PointNet) can uniformly approximate all set functions that are uniformly continuous with respect to
 1515 the Wasserstein-1 distance (respectively, the Hausdorff distance). In contrast, our work focuses on
 1516 transferability and size generalization, rather than expressive power.

1519 F.2.1 Transferability analysis: Proof of Corollary 5.2

1520 We prove the Corollary by instantiating Proposition E.7. The invariant networks is given by the
 1521 following composition, so it is sufficient to check each of them individually:

$$\mathbb{R}^{n \times d} \xrightarrow[\text{row-wise}]{\rho^{\oplus n}} \mathbb{R}^{n \times h} \xrightarrow{\text{Agg}_{i=1}^n} \mathbb{R}^h \xrightarrow{\sigma} \mathbb{R},$$

1522 where we use $\rho^{\oplus n}$ to denote the row-wise application of the same $\rho: \mathbb{R}^d \rightarrow \mathbb{R}^h$.

1523 **DeepSet.** Notice that the sum aggregation is not compatible with the duplication embedding. Indeed,
 1524 for any $x \in \mathbb{R}^n$ such that $\sum x_i \neq 0$, and for $n \mid N, n \neq N$,

$$\sum_{i=1}^N (x \otimes \mathbb{1}_{N/n})_i = (N/n) \sum_{i=1}^n x_i \neq \sum_{i=1}^n x_i.$$

1525 Therefore, DeepSet is not compatible with respect to the duplication consistent sequence in general.

1526 We now prove its compatibility and transferability with respect to the zero-padding.

1527 **Corollary F.3.** Fix arbitrary norms $\|\cdot\|_{\mathbb{R}^d}$ on \mathbb{R}^d and $\|\cdot\|_{\mathbb{R}^h}$ on \mathbb{R}^h . Let $\rho: \mathbb{R}^d \rightarrow \mathbb{R}^h$ be L_{ρ} -Lipschitz
 1528 with $\rho(0) = 0$, and $\sigma: \mathbb{R}^h \rightarrow \mathbb{R}$ be L_{σ} -Lipschitz, with respect to the norms $\|\cdot\|_{\mathbb{R}^d}$, $\|\cdot\|_{\mathbb{R}^h}$, and $|\cdot|$.

1529 Then the sequence of maps (DeepSet_n) is $(L_{\rho}L_{\sigma})$ -Lipschitz transferable with respect to the zero-
 1530 padding consistent sequence $\mathbb{V}_{\text{zero}}^{\oplus d}$ (equipped with the ℓ_1 -norm induced by $\|\cdot\|_{\mathbb{R}^d}$) and the trivial
 1531 consistent sequence $\mathbb{V}_{\mathbb{R}}$ (with absolute value norm).

1532 Therefore, (DeepSet_n) extends to

$$\text{DeepSet}_\infty : \ell_1(\mathbb{R}^d) \rightarrow \mathbb{R}, \quad \text{DeepSet}_\infty((x_i)_{i=1}^\infty) = \sigma \left(\sum_{i=1}^\infty \rho(x_i) \right),$$

1533 which is $(L_\rho L_\sigma)$ -Lipschitz with respect to the ℓ_1 -norm on the infinite sequences.

1534 *Proof.* We model each intermediate space with consistent sequences:

$$\mathbb{V}_{\text{zero}}^{\oplus d} = (\mathbb{R}^{n \times d}) \xrightarrow[\text{(row-wise)}]{\rho^{\oplus n}} \mathbb{V}_{\text{zero}}^{\oplus h} = (\mathbb{R}^{n \times h}) \xrightarrow{\sum_{i=1}^n} \mathbb{V}_{\mathbb{R}^h} = (\mathbb{R}^h) \xrightarrow{\sigma} \mathbb{V}_{\mathbb{R}} = (\mathbb{R}).$$

1535 We first check the compatibility of each map:

1536 • As long as $\rho(0) = 0$, the ρ -map is compatible because

$$\rho^{\oplus N} \left(\begin{bmatrix} X \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \rho^{\oplus n}(X) \\ 0 \end{bmatrix} \text{ for all } X \in \mathbb{R}^{n \times d}, 0 \in \mathbb{R}^{(N-n) \times d}, n \leq N,$$

1537 and the row-wise application makes sure ρ is S_n -equivariant.

1538 • The sum aggregation $\text{Agg}_{i=1}^n = \sum_{i=1}^n$ is compatible because adding zeros does not change the
1539 sum, and the summation operation is S_n -invariant.

1540 • The map σ is between two trivial consistent sequences. Hence it is automatically compatible.

1541 Endow $\mathbb{V}_{\text{zero}}^{\oplus d}$ with the ℓ_1 norm induced by $\|\cdot\|_{\mathbb{R}^d}$, and $\mathbb{V}_{\text{zero}}^{\oplus h}$ with the ℓ_1 norm with induced by $\|\cdot\|_{\mathbb{R}^h}$.
1542 $\mathbb{V}_{\mathbb{R}^h}, \mathbb{V}_{\mathbb{R}}$ are endowed with $\|\cdot\|_{\mathbb{R}^h}$ and $|\cdot|$ respectively. We check the Lipschitz transferability of
1543 each map:

1544 • The ρ -map is L_ρ -Lipschitz transferable map because for all n , we can prove $\rho^{\oplus n} : \mathbb{R}^{n \times d} \rightarrow$
1545 $\mathbb{R}^{n \times h}$ (applying the same ρ row-wise) is L_ρ Lipschitz with respect to the ℓ_1 norms:

$$\|\rho^{\oplus n}(X) - \rho^{\oplus n}(Y)\|_1 = \sum_{i=1}^n \|\rho(X_{i:}) - \rho(Y_{i:})\|_{\mathbb{R}^h} \leq \sum_{i=1}^n L_\rho \|X_{i:} - Y_{i:}\|_{\mathbb{R}^d} = L_\rho \|X - Y\|_1.$$

1546 • $\text{Agg}_{i=1}^n = \sum_{i=1}^n$ is 1-Lipschitz transferable because for all n ,

$$\left\| \sum_{i=1}^n X_{i:} - \sum_{i=1}^n Y_{i:} \right\|_{\mathbb{R}^h} \leq \sum_{i=1}^n \|X_{i:} - Y_{i:}\|_{\mathbb{R}^h} = \|X - Y\|_1.$$

1547 We highlight that this does not necessarily hold for other ℓ_p norms for $p \neq 1$.

1548 • The map σ is L_σ -Lipschitz transferable.

1549 The result follows from Proposition 5.1. □

1550 **Normalized DeepSet.** The mean aggregation is not compatible with the zero-padding embedding.
1551 Consider a vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ such that $\sum x_i \neq 0$, and suppose $n < N$. When
1552 zero-padded to length N , we obtain

$$\tilde{x} = (x_1, \dots, x_n, 0, \dots, 0) \in \mathbb{R}^N.$$

1553 Then

$$\frac{1}{N} \sum_{i=1}^N \tilde{x}_i = \frac{1}{N} \sum_{i=1}^n x_i \neq \frac{1}{n} \sum_{i=1}^n x_i.$$

1554 Therefore, normalized DeepSet is not compatible with respect to the zero-padding consistent sequence
1555 in general.

1556 We now prove its compatibility and transferability with respect to the duplication consistent sequence
1557 with normalized ℓ_p norm.

1558 **Corollary F.4.** Fix arbitrary norms $\|\cdot\|_{\mathbb{R}^d}$ on \mathbb{R}^d and $\|\cdot\|_{\mathbb{R}^h}$ on \mathbb{R}^h . Let $\rho : \mathbb{R}^d \rightarrow \mathbb{R}^h$ be L_ρ -Lipschitz,
 1559 and $\sigma : \mathbb{R}^h \rightarrow \mathbb{R}$ be L_σ -Lipschitz, with respect to the norms $\|\cdot\|_{\mathbb{R}^d}$, $\|\cdot\|_{\mathbb{R}^h}$, and $|\cdot|$.

1560 Then for all $p \in [1, \infty]$, the sequence of maps $(\overline{\text{DeepSet}}_n)$ is $(L_\rho L_\sigma)$ -Lipschitz transferable with
 1561 respect to the duplication consistent sequence $\mathbb{V}_{\text{dup}}^{\oplus d}$ (equipped with the normalized ℓ_p norm induced
 1562 by $\|\cdot\|_{\mathbb{R}^d}$) and the trivial consistent sequence $\mathbb{V}_{\mathbb{R}}$ (with absolute value norm).

1563 Therefore, $(\overline{\text{DeepSet}}_n)$ extends to

$$\overline{\text{DeepSet}}_\infty : \mathcal{P}_p(\mathbb{R}^d) \rightarrow \mathbb{R}, \quad \overline{\text{DeepSet}}_\infty(\mu) = \sigma \left(\int \rho d\mu \right),$$

1564 which is $(L_\rho L_\sigma)$ -Lipschitz with respect to the Wasserstein- p distance on $\|\cdot\|_{\mathbb{R}^d}$.

1565 *Proof.* We model each intermediate space with consistent sequences:

$$\mathbb{V}_{\text{dup}}^{\oplus d} = (\mathbb{R}^{n \times d}) \xrightarrow[\text{(row-wise)}]{\rho^{\oplus n}} \mathbb{V}_{\text{dup}}^{\oplus h} = (\mathbb{R}^{n \times h}) \xrightarrow{\frac{1}{n} \sum_{i=1}^n} \mathbb{V}_{\mathbb{R}^h} = (\mathbb{R}^h) \xrightarrow{\sigma} \mathbb{V}_{\mathbb{R}} = (\mathbb{R}).$$

1566 We first consider the compatibility of each map:

- 1567 • The ρ -map is compatible because $\rho^{\oplus N}(X \otimes \mathbb{1}_{N/n}) = \rho^{\oplus n}(X) \otimes \mathbb{1}_{N/n}$ for all $n \mid N$, and the
 1568 row-wise application makes sure ρ is S_n -equivariant.
- 1569 • The mean aggregation $\text{Agg}_{i=1}^n = \frac{1}{n} \sum_{i=1}^n$ is compatible because for all $n \mid N$, $X \in \mathbb{R}^{n \times d}$,

$$\frac{1}{N} \sum_{i=1}^N (X \otimes \mathbb{1}_{N/n})_{i:} = \frac{1}{N} \sum_{i=1}^n (N/n) X_{i:} = \frac{1}{n} \sum_{i=1}^n X_{i:},$$

1570 and the mean operation is S_n -invariant.

- 1571 • The map σ is again automatically compatible.

1572 Endow $\mathbb{V}_{\text{dup}}^{\oplus d}$ with the normalized ℓ_p norm with respect to $\|\cdot\|_{\mathbb{R}^d}$, and $\mathbb{V}_{\text{dup}}^{\oplus h}$ with the normalized ℓ_p
 1573 norm with respect to $\|\cdot\|_{\mathbb{R}^h}$. The trivial consistent sequences $\mathbb{V}_{\mathbb{R}^h}, \mathbb{V}_{\mathbb{R}}$ are endowed with $\|\cdot\|_{\mathbb{R}^h}$ and
 1574 $|\cdot|$ respectively. We check the Lipschitz transferability of each map:

- 1575 • The ρ -map is L_ρ -Lipschitz because for all n , we can prove $\rho^{\oplus n} : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times h}$ is L_ρ
 1576 Lipschitz with respect to the normalized ℓ_p norm:

$$\begin{aligned} \|\rho^{\oplus n}(X) - \rho^{\oplus n}(Y)\|_{\bar{p}} &= \left(\frac{1}{n} \sum_{i=1}^n \|\rho(X_{i:}) - \rho(Y_{i:})\|_{\mathbb{R}^h}^p \right)^{1/p} \\ &\leq \left(\frac{1}{n} \sum_{i=1}^n L_\rho^p \|X_{i:} - Y_{i:}\|_{\mathbb{R}^d}^p \right)^{1/p} \\ &= L_\rho \|X - Y\|_{\bar{p}}. \end{aligned}$$

- 1577 • The mean aggregation $\text{Agg}_{i=1}^n = \frac{1}{n} \sum_{i=1}^n$ is 1-Lipschitz transferable because

$$\left\| \frac{1}{n} \sum_{i=1}^n X_{i:} - \frac{1}{n} \sum_{i=1}^n Y_{i:} \right\|_{\mathbb{R}^h} \leq \frac{1}{n} \sum_{i=1}^n \|X_{i:} - Y_{i:}\|_{\mathbb{R}^h} = \|X - Y\|_{\bar{1}} \leq \|X - Y\|_{\bar{p}},$$

1578 where the last inequality follows from Hölder's inequality.

- 1579 • The map σ is L_σ -Lipschitz transferable.

1580 The result follows from Proposition 5.1. □

1581 **Remark F.5.** The same result was also proved in [5, Theorem 3.7] by directly verifying the Lipschitz
 1582 property of DeepSet_∞ : for all $p \geq 1$,

$$\begin{aligned} |\overline{\text{DeepSet}}_\infty(\mu_X) - \overline{\text{DeepSet}}_\infty(\mu_Y)| &\leq L_\sigma \left| \int \rho d(\mu_X - \mu_Y) \right| \\ &\leq L_\sigma L_\rho W_1(\mu_X, \mu_Y) \leq L_\sigma L_\rho W_p(\mu_X, \mu_Y), \end{aligned}$$

1583 where the second inequality follows from the Kantorovich-Rubinstein duality. Our methods provide
 1584 an alternative proof, using a proof technique (Proposition 5.1) that applies more generally.

1585 **PointNet.** The max aggregation is not compatible with zero-padding. Consider a vector $x =$
 1586 $(x_1, \dots, x_n) \in \mathbb{R}^n$ where all entries $x_i < 0$, and suppose $n < N$. When zero-padded to length N ,
 1587 we obtain

$$\tilde{x} = (x_1, \dots, x_n, 0, \dots, 0) \in \mathbb{R}^N.$$

1588 Then,

$$\max_{1 \leq i \leq N} \tilde{x}_i = 0 \neq \max_{1 \leq i \leq n} x_i.$$

1589 Hence, unless we restrict the model to avoid all-negative entries, PointNet is not compatible with the
 1590 zero-padding consistent sequence.

1591 We now prove its compatibility and transferability with respect to the duplication sequence with the
 1592 ℓ_∞ norm.

1593 **Corollary F.6.** Fix arbitrary norms $\|\cdot\|_{\mathbb{R}^d}$ on \mathbb{R}^d and $\|\cdot\|_\infty$ on \mathbb{R}^h . Let $\rho : \mathbb{R}^d \rightarrow \mathbb{R}^h$ be L_ρ -Lipschitz,
 1594 and $\sigma : \mathbb{R}^h \rightarrow \mathbb{R}$ be L_σ -Lipschitz, with respect to the norms $\|\cdot\|_{\mathbb{R}^d}$, $\|\cdot\|_\infty$ on \mathbb{R}^h , and $|\cdot|$.

1595 Then the sequence of maps (PointNet_n) is $(L_\rho L_\sigma)$ -Lipschitz transferable with respect to the dupli-
 1596 cation consistent sequence $\mathbb{V}_{\text{dup}}^{\oplus d}$ (equipped with the ℓ_∞ -norm induced by $\|\cdot\|_{\mathbb{R}^d}$) and the trivial
 1597 consistent sequence $\mathbb{V}_{\mathbb{R}}$ (with absolute value norm).

1598 Therefore, (PointNet_n) extends to

$$\text{PointNet}_\infty : \mathcal{P}_\infty(\mathbb{R}^d) \rightarrow \mathbb{R}, \quad \text{PointNet}_\infty(\mu) = \sigma \left(\sup_{x \in \text{supp}(\mu)} \rho(x) \right),$$

1599 which is $(L_\rho L_\sigma)$ -Lipschitz with respect to the Wasserstein- ∞ distance on $\|\cdot\|_{\mathbb{R}^d}$.

1600 *Proof.* We again consider consistent sequences

$$\mathbb{V}_{\text{dup}}^{\oplus d} = (\mathbb{R}^{n \times d}) \xrightarrow[\text{(row-wise)}]{\rho^{\oplus n}} \mathbb{V}_{\text{dup}}^{\oplus h} = (\mathbb{R}^{n \times h}) \xrightarrow{\max_{i=1}^n} \mathbb{V}_{\mathbb{R}^h} = (\mathbb{R}^h) \xrightarrow{\sigma} \mathbb{V}_{\mathbb{R}} = (\mathbb{R}).$$

1601 For compatibility, it is left to check that $\text{Agg}_{i=1}^n = \max_{i=1}^n$ is compatible. Indeed, for any $X \in$
 1602 $\mathbb{R}^{n \times d}$, $n \mid N$, we have $\max_{i=1}^N (X \otimes \mathbb{1}_{N/n})_{i:} = \max_{i=1}^n X_{i:}$, and the max operation is S_n -invariant.

1603 Endow $\mathbb{V}_{\text{dup}}^{\oplus d}$ with the ℓ_∞ norm with respect to $\|\cdot\|_{\mathbb{R}^d}$, and $\mathbb{V}_{\text{dup}}^{\oplus h}$ with the ℓ_∞ norm with respect
 1604 to $\|\cdot\|_\infty$ on \mathbb{R}^h . $\mathbb{V}_{\mathbb{R}^h}$, $\mathbb{V}_{\mathbb{R}}$ are endowed with $\|\cdot\|_\infty$ and $|\cdot|$ respectively. We check the Lipschitz
 1605 transferability of each map:

- 1606 • We have proved in the proof for normalized DeepSet that ρ, σ are L_ρ, L_σ Lipschitz transferable
 1607 respectively.
- 1608 • For any $j \in [d]$, $|\max_{i=1}^n X_{ij} - \max_{i=1}^n Y_{ij}| \leq \max_{i=1}^n |X_{ij} - Y_{ij}|$. Take max over $j \in [d]$,
 1609 we conclude

$$\left\| \max_{i=1}^n X_{i:} - \max_{i=1}^n Y_{i:} \right\|_\infty = \max_{j=1}^d \left| \max_{i=1}^n X_{ij} - \max_{i=1}^n Y_{ij} \right| \leq \max_{i=1}^n \|X_{i:} - Y_{i:}\|_\infty.$$

1610 Hence $\text{Agg}_{i=1}^n = \max_{i=1}^n$ is 1-Lipschitz transferable.

1611 The result follows from Proposition 5.1. □

1612 **Remark F.7.** PointNet_∞ produces identical outputs for probability measures with the same support.
 1613 Thus, it can be viewed as a function

$$\text{PointNet}_\infty : \mathcal{K}(\mathbb{R}^d) \rightarrow \mathbb{R}, \quad \text{PointNet}_\infty(X) = \sigma \left(\sup_{x \in X} \rho(x) \right),$$

1614 where $\mathcal{K}(\mathbb{R}^d)$ denotes the set of non-empty compact subsets of \mathbb{R}^d . The W_∞ distance on $\mathcal{P}_\infty(\mathbb{R}^d)$
 1615 induces the quotient metric d_K on $\mathcal{K}(\mathbb{R}^d)$ via the equivalence relation $\mu \sim \nu$ if $\text{supp}(\mu) = \text{supp}(\nu)$.
 1616 Our results imply that PointNet_∞ is $(L_\rho L_\sigma)$ -Lipschitz with respect to d_K .

1617 A more commonly used metric on $\mathcal{K}(\mathbb{R}^d)$ is the Hausdorff distance, defined by

$$d_H(X, Y) := \max \left\{ \sup_{x \in X} \inf_{y \in Y} \|x - y\|, \sup_{y \in Y} \inf_{x \in X} \|x - y\| \right\}.$$

1618 [5, Theorem 3.7] shows that PointNet_∞ is $(2L_\rho L_\sigma)$ -Lipschitz with respect to d_H . It is easy to see
 1619 that $d_H \leq d_K$, but we leave the exploration of further relations between these two metrics to future
 1620 work.

1621 Finally, we show that the sequence of maps (PointNet_n) is, in general, not transferable with respect
 1622 to the duplication-based consistent sequence $\mathbb{V}_{\text{dup}}^{\oplus d}$ when equipped with the normalized ℓ_p norm for
 1623 any $p \in [1, \infty)$. Consider the sequence of matrices $(X^{(n)} \in \mathbb{R}^{n \times h})_n$ where the first row is the
 1624 all-one vector $\mathbb{1}_h^\top$, and the remaining $n - 1$ rows are zero vectors. Then,

$$\|X^{(n)} - 0\|_{\bar{p}} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

1625 which implies that $[X^{(n)}] \rightarrow [0]$ in V_∞ . However,

$$\max_{i=1}^n X_{i:}^{(n)} = \mathbb{1}_h \quad \text{for all } n.$$

1626 This demonstrates that the max aggregation $\text{Agg}_{i=1}^n = \max_{i=1}^n$ is not continuously transferable under
 1627 the normalized ℓ_p norm, and so neither is the sequence (PointNet_n) .

1628 F.2.2 Explanation of transferability plots (Figure 1)

1629 Our numerical experiment in Figure 1 illustrates the second column of Table 2 for $p = 1$: the
 1630 Lipschitz transferability of normalized DeepSet, and non-transferability of DeepSet and PointNet,
 1631 with respect to $(\mathbb{V}_{\text{dup}}, \|\cdot\|_1)$.

1632 First, applying Proposition D.6 to normalized DeepSet, and recalling the convergence rate of empirical
 1633 distributions given in (2) for $d = 1, p = 1$, we get the following Corollary:

1634 **Corollary F.8** (Convergence and transferability of normalized DeepSet). *Let $(x_n \in \mathbb{R}^n)_{n \in \mathbb{N}}$ be
 1635 a sequence of inputs with entries $(x_n)_i \stackrel{i.i.d.}{\sim} \mu$ for $i = 1, \dots, n$, where μ is a probability measure
 1636 on \mathbb{R} with finite expectation. Define the empirical measure $\mu_x := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ for $x \in \mathbb{R}^n$. Then
 1637 $\mathbb{E}[W_1(\mu, \mu_{x_n})] \lesssim n^{-1/2}$, and hence*

$$\mathbb{E} [|\overline{\text{DeepSet}}_n(x_n) - \overline{\text{DeepSet}}_\infty(\mu)|] \lesssim n^{-1/2},$$

1638

$$\mathbb{E} [|\overline{\text{DeepSet}}_n(x_n) - \overline{\text{DeepSet}}_m(x_m)|] \lesssim n^{-1/2} + m^{-1/2}.$$

1639 Indeed, Figure 1(b) shows convergence of the model outputs, and Figure 1(d) confirms that the
 1640 convergence rate is $O(n^{-1/2})$, as predicted.

1641 Figure 1(a) illustrates the divergence of outputs from DeepSet. This occurs because the sum
 1642 $\sum_{i=1}^n \rho(x_i) = O(n)$. If the function σ in DeepSet is unbounded, this leads to unbounded (blow-up)
 1643 outputs as n increases.

1644 Figure 1(c) shows divergent outputs from PointNet. When the input distribution μ has compact
 1645 support, the output of PointNet will converge, although without guarantees on the rate. However, in
 1646 our experiment where $\mu = \mathcal{N}(0, 1)$ has non-compact support. If ρ in the PointNet is unbounded,
 1647 the maximum value $\max_{i=1}^n \rho(x_i)$ diverges almost surely as $n \rightarrow \infty$. This again results in blow-up
 1648 outputs.

1649 G Example 2 on graphs: details and missing proofs from Section 5.2

1650 G.1 Duplication consistent sequence for graphs

1651 Start with the duplication consistent sequence for sets \mathbb{V}_{dup} defined in F.1, we define

$$\mathbb{V}_{\text{dup}}^G := \text{Sym}^2(\mathbb{V}_{\text{dup}}) \oplus \mathbb{V}_{\text{dup}}^{\oplus d},$$

1652 following the definition of direct sum and tensor product in Definition C.4, C.5. This gives the
 1653 duplication consistent sequence for graphs. Specifically, $\mathbb{V}_{\text{dup}}^G = \{(V_n), (\varphi_{N,n}), (G_n)\}$ where
 1654 $V_n = \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d}$ for each n and the embedding for $n \mid N$ is given by,

$$\begin{aligned} \varphi_{N,n} : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d} &\hookrightarrow \mathbb{R}_{\text{sym}}^{N \times N} \times \mathbb{R}^{N \times d} \\ (A, X) &\mapsto (A \otimes (\mathbb{1}_{N/n} \mathbb{1}_{N/n}^\top), X \otimes \mathbb{1}_{N/n}), \end{aligned}$$

1655 which can be interpreted as replacing each node in the graph with N/n duplicated copies. The
 1656 symmetric group S_n acts on V_n by $g \cdot (A, X) = (gAg^\top, gX)$.

1657 The space V_∞ can be identified with the space with all step graphons (and signals) in a similar
 1658 way: Given $(A, X) \in \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d}$, define $W_A : [0, 1]^2 \rightarrow \mathbb{R}$, $f_X : [0, 1] \rightarrow \mathbb{R}^d$ such that
 1659 W_A takes value $A_{i,j}$ on the interval $I_{i,n} \times I_{j,n} \subset [0, 1]^2$ for $i, j = 1, \dots, n \in [n]$, and f_X takes
 1660 value X_i on the interval $I_{i,n} \subset [0, 1]$ for $i = 1, \dots, n$. We call (W_A, f_X) the *induced step*
 1661 *graphon* from (A, X) . Under this identification, permutations S_n permute the n intervals $I_{i,n}$ and
 1662 act on (W, f) by $\sigma \cdot (W, f) = (\sigma \cdot W, \sigma \cdot f) = (W^{\sigma^{-1}}, f \circ \sigma^{-1})$, where $W^{\sigma^{-1}}$ is defined by
 1663 $W^{\sigma^{-1}}(x, y) := W(\sigma^{-1}(x), \sigma^{-1}(y))$. The limit group G_∞ is the union of such interval permutations.

1664 **p -norm on $\mathbb{V}_{\text{dup}}^G$.** Fix $\|\cdot\|_{\mathbb{R}^d}$ a norm on \mathbb{R}^d . We equip V_n with a p -norm given by

$$\begin{aligned} \|(A, X)\|_{\bar{p}} &:= \max(\|A\|_{\bar{p}}, \|X\|_{\bar{p}}) \\ &= \begin{cases} \max\left(\left(\frac{1}{n^2} \sum_{i,j \in [n]} |A_{ij}|^p\right)^{1/p}, \left(\frac{1}{n} \sum_{i=1}^n \|X_i\|_{\mathbb{R}^d}^p\right)^{1/p}\right) & p \in [1, \infty), \\ \max(\max_{i,j \in [n]} |A_{ij}|, \max_{i=1}^n \|X_i\|_{\mathbb{R}^d}) & p = \infty. \end{cases} \end{aligned} \quad (10)$$

1665 It is easy to check by Proposition C.8 that this extends to a norm on V_∞ . Under the identification
 1666 with step graphons, this norm on V_∞ coincides with the standard L^p -norm given by

$$\begin{aligned} \|(W, f)\|_p &:= \max(\|W\|_p, \|f\|_p) \\ &= \begin{cases} \max\left(\left(\iint |W(x, y)|^p dx dy\right)^{1/p}, \left(\int |f(x)|^p dx\right)^{1/p}\right) & p \in [1, \infty), \\ \max(\sup_{x,y} |W(x, y)|, \sup_x |f(x)|) & p = \infty. \end{cases} \end{aligned}$$

1667 That is, for any $(A, X) \in \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d}$, $\|(A, X)\|_{\bar{p}} = \|(W_A, f_X)\|_p$.

1668 The symmetrized metric is

$$\bar{d}_p((W, f), (W', f')) = \inf_{\sigma \in G_\infty} \{\max(\|W - \sigma \cdot W'\|_p, \|f - \sigma \cdot f'\|_p)\}. \quad (11)$$

1669 **Operator p -norm on $\mathbb{V}_{\text{dup}}^G$.** Fix $\|\cdot\|_{\mathbb{R}^d}$ a norm on \mathbb{R}^d , $p \in [1, \infty]$. We equip V_n with a norm given
 1670 by

$$\|(A, X)\|_{\text{op}, p} := \max\left(\frac{1}{n} \|A\|_{\text{op}, p}, \|X\|_{\bar{p}}\right), \quad (12)$$

1671 where $\|A\|_{\text{op}, p}$ is the operator norm of A with respect to the ℓ_p norm, i.e.

$$\|A\|_{\text{op}, p} = \max_{\|x\|_p=1} \|Ax\|_p,$$

1672 and $\|X\|_{\bar{p}}$ is the normalized ℓ_p norm with respect to $\|\cdot\|_{\mathbb{R}^d}$ defined in (9). It is easy to check by
 1673 Proposition C.8 that this extends to a norm on V_∞ . Under the identification with step graphons, this
 1674 norm on V_∞ coincides with

$$\|(W, f)\|_{\text{op}, p} := \max(\|T_W\|_{\text{op}, p}, \|f\|_p),$$

1675 where T_W is the shift operator of graphon W defined by $T_W(f)(u) := \int_0^1 W(u, v)f(v)dv$, and its
 1676 norm $\|T_W\|_{\text{op}, p} := \sup_{\|f\|_p=1} \|T_W(f)\|_p$. The norm on $\|f\|_p$ is the L^p norm as before.

1677 That is, for any $(A, X) \in \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d}$, $\|(A, X)\|_{\text{op}, p} = \|(W_A, f_X)\|_{\text{op}, p}$.

1678 For $p \in [1, \infty)$, the completion with respect to this metric is

$$\overline{V_\infty} = \{W \in L^p([0, 1]^2) : W(x, y) = W(y, x)\} \times \{f \in L^p([0, 1], \mathbb{R}^d)\},$$

1679 the space of L^p -graphon signals.

1680 For $p = \infty$, we won't exactly characterize $\overline{V_\infty}$ (it is again a space of certain regulated functions), but
 1681 it contains all bounded and continuous graphon signals.

1682 The symmetrized metric is

$$\bar{d}_p((W, f), (W', f')) = \inf_{\sigma \in \mathbb{G}_\infty} \{\max(\|T_W - T_{\sigma \cdot W'}\|_{\text{op}, p}, \|f - \sigma \cdot f'\|_p)\}. \quad (13)$$

1683 **Cut norm on $\mathbb{V}_{\text{dup}}^G$.** We can also equip V_n with the *cut norm* (on matrices and vectors),

$$\|(A, X)\|_\square := \max(\|A\|_\square, \|X\|_\square) = \max\left(\frac{1}{n^2} \max_{S, T \subseteq [n]} \left| \sum_{i \in S, j \in T} A_{ij} \right|, \frac{1}{n} \max_{S \subseteq [n]} \left\| \sum_{i \in S} X_i \right\|_{\mathbb{R}^d} \right). \quad (14)$$

1684 It is easy to check by Proposition C.8 that this extends to a norm on V_∞ . Under the identification
 1685 with step graphons, this norm on V_∞ coincides with the cut norm on graphons and graphon signals

$$\begin{aligned} \|(W, f)\|_\square &:= \max(\|W\|_\square, \|f\|_\square) \\ &= \max\left(\sup_{S, T \subseteq [0, 1]} \left| \int_{S \times T} W(x, y) dx dy \right|, \sup_{S \subseteq [0, 1]} \left\| \int_S f(x) dx \right\|_{\mathbb{R}^d} \right), \end{aligned}$$

1686 where the supremum is taken over all measurable sets S, T .

1687 The cut norm on graphon is studied in-depth in [44]. Though hard to compute, it has strong combina-
 1688 torial interpretations. Hence it has played an important role in the work of GNN transferability, and
 1689 has been extended to the graphon signals in [37] which we have adopted here.

1690 The symmetrized metric is

$$\bar{d}((W, f), (W', f')) := \inf_{\sigma \in \mathbb{G}_\infty} \{\max(\|W - \sigma \cdot W'\|_\square, \|f - \sigma \cdot f'\|_\square)\}.$$

1691 It can be proved that this exactly coincides with the *cut distance* below, defined on graphon signals
 1692 (extending the original definition of graphon *cut distance* from [44], similarly to the version in [37]):

$$\bar{d}((W, f), (W', f')) = \inf_{\sigma \in S_{[0, 1]}} \{\max(\|W - \sigma \cdot W'\|_\square, \|f - \sigma \cdot f'\|_\square)\}, \quad (15)$$

1693 where $S_{[0, 1]}$ is the group of measure-preserving bijections $\sigma : [0, 1] \rightarrow [0, 1]$ with measurable inverse.

1694 The proof follows analogously to [4, Lemma 3.5]. Specifically, we first verify that the definitions
 1695 agree on step graphons. Then, since both definitions are continuous with respect to the cut norm
 1696 $\|\cdot\|_\square$, they must also agree on the limit points.

1697 While the cut norm is hard to work with directly, it is topologically equivalent to the operator 2-norms
 1698 considered previously on a bounded domain. This means that any function continuous with respect to
 1699 one of these norms is also continuous with respect to the other.

1700 **Proposition G.1.** *If $\|W\|_\infty < r$ and $\|f\|_\infty < r$, then*

$$\|(W, f)\|_\square \leq \|(W, f)\|_{\text{op}, 2} \lesssim \|(W, f)\|_\square^{1/2}.$$

1701 *Consequently, for $p \in (1, \infty)$, $\|\cdot\|_\square$ and $\|\cdot\|_p$ are topologically equivalent on the space*

$$\{W : [0, 1]^2 \rightarrow [-r, r] \text{ measurable}, W(x, y) = W(y, x)\} \times \{f : [0, 1] \rightarrow [-r, r] \text{ measurable}\}.$$

1702 *Proof.* Without loss of generality, let $r = 1$. Consider the norm

$$\|T_W\|_{\text{op},\infty,1} := \sup_{\|f\|_\infty, \|g\|_\infty \leq 1} \left| \int_0^1 \int_0^1 W(u,v) f(u) g(v) du dv \right|.$$

1703 By [28, Equation 4.4], $\|W\|_\square \leq \|T_W\|_{\text{op},\infty,1} \leq 4\|W\|_\square$. By [28, Lemma E.6], if $\|W\|_\infty \leq 1$,

$$\|T_W\|_{\text{op},\infty,1} \leq \|T_W\|_{\text{op},2} \leq \sqrt{2}\|T_W\|_{\text{op},\infty,1}^{1/2}.$$

1704 Combining the inequalities,

$$\|W\|_\square \leq \|T_W\|_{\text{op},2} \leq 2^{3/2}\|W\|_\square^{1/2}.$$

1705 Moreover, by [37, Appendix A.2], $\|f\|_\square \leq \|f\|_1 \leq 2\|f\|_\square$. If $\|f\|_\infty \leq 1$, then $\|f\|_2^2 \leq \|f\|_1 \leq$
 1706 $\|f\|_2$. Combining the inequalities,

$$\|f\|_\square \leq \|f\|_2 \leq 2^{1/2}\|f\|_\square^{1/2}.$$

1707 Therefore, we conclude that

$$\|(W, f)\|_\square \leq \|(W, f)\|_{\text{op},p} \leq 2^{3/2}\|(W, f)\|_\square^{1/2},$$

1708 as claimed. \square

1709 G.2 Message Passing Neural Networks (MPNNs)

1710 **Background.** MPNN parametrizes a sequence of functions $(\text{MPNN}_n : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_1} \rightarrow \mathbb{R}^{n \times d_L})$
 1711 by composition of message passing layers. The l -th message passing layer

$$\text{MP}_n^{(l)} : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_{l+1}}, \quad (A, X^{(l)}) \mapsto (A, X^{(l+1)})$$

1712 is given by

$$X_{i:}^{(l+1)} = \phi^{(l)} \left(X_{i:}^{(l)}, \text{Agg}_{j \in \mathcal{N}_i} \psi^{(l)} \left(X_{i:}^{(l)}, X_{j:}^{(l)}, A_{ij} \right) \right), \quad i = 1, \dots, n, \quad (16)$$

1713 where Agg is a permutation-invariant aggregation function such as sum, mean, or max; $\mathcal{N}_i :=$
 1714 $\{j : A_{ij} \neq 0\}$ denotes the neighborhood of node i in the input graph; the message function
 1715 $\psi^{(l)} : \mathbb{R}^{d_l} \times \mathbb{R}^{d_l} \times \mathbb{R} \rightarrow \mathbb{R}^{h_l}$ and the update function $\phi^{(l)} : \mathbb{R}^{d_l} \times \mathbb{R}^{h_l} \rightarrow \mathbb{R}^{d_{l+1}}$ are independent of the
 1716 graph size n . Composing L message-passing layers defines an MPNN, mapping $(A, X^{(1)}) \mapsto X^{(L)}$.

1717 Observe that MPNN is permutation-equivariant: $\text{MPNN}_n(gAg^\top, gX) = g\text{MPNN}_n(A, X)$ for all
 1718 $g \in \mathcal{S}_n$. If we want a permutation-invariant function, this is followed by a read-out operation taking
 1719 the form of DeepSet. In this work, we focus on the equivariant case.

1720 MPNN is a general framework for GNNs based on local message passing: [24] formulates multiple
 1721 GNNs as MPNNs with specific choices of ϕ , ψ , Agg ; other state-of-the-art GNNs can be simulated
 1722 by MPNN on a transformed graph [29]. Moreover, ϕ and ψ can also be parameterized with MLPs to
 1723 provide good flexibility.

1724 Transferability analysis of MPNNs.

1725 **Corollary G.2.** Consider one message passing layer, $(\text{MP}_n^{(l)})$, as defined in (16), of the following
 1726 special form:

- 1727 • The message function $\psi^{(l)}$ takes the form $\psi^{(l)}(x_1, x_2, w) := w\xi(x_2)$, where $\xi : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{h_l}$ is
 1728 L_ξ Lipschitz with respect to $\|\cdot\|_{\mathbb{R}^{d_l}}, \|\cdot\|_{\mathbb{R}^{h_l}}$.
- 1729 • The aggregation used is the normalized sum aggregation $\text{Agg}_{j \in \mathcal{N}_i} := \frac{1}{n} \sum_{j \in \mathcal{N}_i}$.
- 1730 • The update function $\phi^{(l)}$ is L_ϕ Lipschitz, i.e. for all $(x, y), (x', y') \in \mathbb{R}^{d_l} \times \mathbb{R}^{h_l}$,

$$\|\phi^{(l)}(x, y) - \phi^{(l)}(x', y')\|_{\mathbb{R}^{d_{l+1}}} \leq L_\phi \max(\|x - x'\|_{\mathbb{R}^{d_l}}, \|y - y'\|_{\mathbb{R}^{h_l}})$$

1731 *Endow the space of duplication-consistent sequences with the operator p -norm as defined in (12),*
 1732 *where $p \in [1, \infty)$. Then the sequence of maps $(\text{MP}_n^{(l)})$ is locally Lipschitz transferable.*

1733 *Therefore, (MPNN_n) , which is a composition of message-passing layers, is locally Lipschitz trans-*
 1734 *ferable. It extends to a function MPNN_∞ on the space of graphon signals, which is $L(r)$ -Lipschitz*
 1735 *on $B(0, r)$ for all $r > 0$ with respect to the symmetrized operator p -metric defined in (13)).*

1736 **Remark G.3.** *By Proposition G.1, the sequence of maps (MPNN_n) is continuously transferable*
 1737 *with respect to the cut norm (14) on $B(0, r)$.*

1738 *The GNN studied in [56] is a special case of our MPNN considered here; meanwhile, ours is a special*
 1739 *case of [37], which directly establishes Lipschitzness with respect to the cut distance by analysis on*
 1740 *the graphon space. While our results are not new, our proof technique—following Proposition 5.1—is*
 1741 *new and generally applicable to various models.*

1742 *Proof.* We decompose $\text{MP}_n^{(l)}$ as a composition of the following maps, modelling each of the interme-
 1743 *diates spaces using the duplication consistent sequences endowed with compatible norms. For the*
 1744 *metric on the product spaces, we always use the L^∞ product metric, i.e., taking the maximum over*
 1745 *the individual components.*

$$\begin{aligned}
 1746 \quad & \bullet f_n^{(1)} : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \times \mathbb{R}^{n \times h_l}, \quad (A, X) \mapsto \left(A, X, \begin{bmatrix} \xi(X_{1:}) \\ \vdots \\ \xi(X_{n:}) \end{bmatrix} \right). \\
 1747 \quad & \bullet f_n^{(2)} : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \times \mathbb{R}^{n \times h_l} \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \times \mathbb{R}^{n \times h_l}, \quad (A, X, Y_0) \mapsto (A, X, \frac{1}{n} A Y_0). \\
 1748 \quad & \bullet f_n^{(3)} : \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_l} \times \mathbb{R}^{n \times h_l} \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d_{l+1}}, \quad (A, X, Y) \mapsto (A, \tilde{X}) \text{ where } \tilde{X}_{i:} = \\
 1749 \quad & \phi^{(l)}(X_{i:}, Y_{i:}).
 \end{aligned}$$

1750 It is easy to check that each of them is compatible with respect to the duplication embedding. We
 1751 now check the Lipschitz transferability.

1752

- $(f_n^{(1)})$ is Lipschitz transferable because

$$\left\| f_n^{(1)}(A, X) - f_n^{(1)}(A', X') \right\|_{\text{op}, p} \leq (1 \vee L_\xi) \|(A, X) - (A', X')\|_{\text{op}, p}.$$

1753

- $(f_n^{(2)})$ is locally Lipschitz transferable because of the following:

1754 Compute the Fréchet derivative:

$$Df_n^{(2)}(A, X, Y_0)[H_A, H_X, H_Y] = \left(H_A, H_X, \frac{1}{n} (A H_Y + H_A Y) \right).$$

1755 Hence, on $\{(A, X, Y_0) : \|(A, X, Y_0)\|_{\text{op}, p} < r\}$,

$$\begin{aligned}
 & \left\| Df_n^{(2)}(A, X, Y_0)[H_A, H_X, H_Y] \right\|_{\text{op}, p} \\
 & \leq \max \left(\frac{1}{n} \|H_A\|_{\text{op}, p}, \|H_X\|_{\bar{p}}, \frac{1}{n} \|A\|_{\text{op}, p} \|H_Y\|_{\bar{p}} + \frac{1}{n} \|H_A\|_{\text{op}, p} \|Y\|_{\bar{p}} \right) \\
 & \leq (1 \vee 2r) \|(H_A, H_X, H_Y)\|_{\text{op}, p},
 \end{aligned}$$

1756 i.e., $f_n^{(2)}$ is $(1 \vee 2r)$ Lipschitz on this space.

1757

- $(f_n^{(3)})$ is Lipschitz transferable because

$$\|f_n^{(3)}(A, X, Y) - f_n^{(3)}(A', X', Y')\|_{\bar{p}} = \max \left(\frac{1}{n} \|A - A'\|_{\text{op}, p}, \|\tilde{X} - \tilde{X}'\|_{\bar{p}} \right)$$

1758

where

$$\begin{aligned}
\|\tilde{X} - \tilde{X}'\|_p^p &\leq \frac{1}{n} \sum_{i=1}^n \|\phi^{(l)}(X_{i:}, Y_{i:}) - \phi^{(l)}(X'_{i:}, Y'_{i:})\|_{\mathbb{R}^{d_{l+1}}}^p \\
&\leq L_\phi^p \frac{1}{n} \sum_{i=1}^n (\|X_{i:} - X'_{i:}\|_{\mathbb{R}^{d_l}} + \|Y_{i:} - Y'_{i:}\|_{\mathbb{R}^{h_l}})^p \\
&\leq L_\phi^p 2^{p-1} \frac{1}{n} \sum_{i=1}^n (\|X_{i:} - X'_{i:}\|_{\mathbb{R}^{d_l}}^p + \|Y_{i:} - Y'_{i:}\|_{\mathbb{R}^{h_l}}^p) \\
&\quad \text{(by Jensen's inequality } (\frac{a+b}{2})^p \leq \frac{a^p+b^p}{2} \text{ for all } a, b) \\
&\leq L_\phi^p 2^p \|(X, Y) - (X', Y')\|_p^p.
\end{aligned}$$

1759

So $f_n^{(3)}$ is $(1 \vee 2L_\phi)$ -Lipschitz.

1760

Finally, apply Proposition 5.1, $(\text{MP}_n^{(l)})$ is locally Lipschitz transferable. \square

1761

1762

1763

1764

1765

1766

1767

Following this result, we can directly apply Propositions D.4 and D.6, along with the convergence rates described in Appendix D.3. The transferability results in [56] consider deterministic sampling, which corresponds to the "uniform grid" sampling scheme discussed in Appendix D.3. Our results yield an improved convergence rate of $O(n^{-1})$, enhancing the previous bounds. The results in [31] address transferability under random sampling from a graphon-based random graph model—similar to our "graphon" sampling scheme. Our framework recovers the same convergence rates established in their work.

1768

G.3 Constructing new transferable GNNs: GGNN and continuous GGNN

1769

1770

1771

1772

Background: Invariant Graph Networks (IGN). Invariant Graph Networks (IGN) [45] are a class of GNN architectures that alternate between linear S_n -equivariant layers and nonlinearities. They follow a design paradigm that differs fundamentally from MPNNs. Specifically, a D -layer 2-IGN parameterizes an S_n -equivariant function $(\mathbb{R}^n)^{\otimes 2} \rightarrow (\mathbb{R}^n)^{\otimes 2}$ as a composition:

$$W_n^{(D)} \circ \rho_n^{(D-1)} \circ \dots \circ \rho_n^{(1)} \circ W_n^{(1)},$$

1773

where for each i :

1774

1775

1776

1777

1778

- $W_n^{(i)}: ((\mathbb{R}^n)^{\otimes 2})^{\oplus d_i} \cong \mathbb{R}^{n^2 \times d_i} \rightarrow ((\mathbb{R}^n)^{\otimes 2})^{\oplus d_{i+1}} \cong \mathbb{R}^{n^2 \times d_{i+1}}$ is a linear S_n -equivariant map. Here, d_i denotes the number of feature channels. [45] provides a parameterization of $W_n^{(i)}$ as a linear combination of basis maps: In the special case where $d_i = d_{i+1} = 1$, the linear layer $W_n^{(i)}$ can be written as a linear combination of 17 basis functions (two of them are biases), where the coefficients α, β are the learnable parameters:

$$\begin{aligned}
W_n^{(i)}(A) &= \alpha_1 A + \alpha_2 A^\top + \alpha_3 \text{diag}(\text{diag}^*(A)) + \alpha_4 A \mathbb{1} \mathbb{1}^\top + \alpha_5 \mathbb{1} \mathbb{1}^\top A^\top + \alpha_6 \text{diag}(A \mathbb{1}) \\
&\quad + \alpha_7 A^\top \mathbb{1} \mathbb{1}^\top + \alpha_8 \mathbb{1} \mathbb{1}^\top A^\top + \alpha_9 \text{diag}(A^\top \mathbb{1}) + \alpha_{10} (\mathbb{1}^\top A \mathbb{1}) \mathbb{1} \mathbb{1}^\top \\
&\quad + \alpha_{11} (\mathbb{1}^\top A \mathbb{1}) \text{diag}(\mathbb{1}) + \alpha_{12} (\mathbb{1}^\top \text{diag}^*(A)) \mathbb{1} \mathbb{1}^\top + \alpha_{13} (\mathbb{1}^\top \text{diag}^*(A)) \text{diag}(\mathbb{1}) \\
&\quad + \alpha_{14} \text{diag}^*(A) \mathbb{1} \mathbb{1}^\top + \alpha_{15} \mathbb{1} \text{diag}^*(A)^\top + \beta_1 \mathbb{1} \mathbb{1}^\top + \beta_2 \text{diag}(\mathbb{1}).
\end{aligned} \tag{17}$$

1779

For general d_i, d_{i+1} , the number of basis terms becomes $17d_i d_{i+1}$.

1780

1781

- $\rho_n^{(i)}: ((\mathbb{R}^n)^{\otimes 2})^{\oplus d_{i+1}} \cong \mathbb{R}^{n^2 \times d_{i+1}} \rightarrow ((\mathbb{R}^n)^{\otimes 2})^{\oplus d_{i+1}} \cong \mathbb{R}^{n^2 \times d_{i+1}}$ applies a nonlinearity (e.g., ReLU) entry-wise.

1782

1783

1784

1785

1786

To improve expressivity, [45] proposed extending the architecture to use higher-order tensors in the intermediate layers. When the maximum tensor order is k , the architecture is referred to as a k -IGN. While this is theoretically tractable, due to the high memory cost and implementation challenges associated with higher-order tensors, in practice, only k -IGNs for $k \leq 2$ have been implemented to the best of our knowledge. In this work, we focus exclusively on 2-IGNs.

The basis in (17) is inherently dimension-agnostic, allowing IGN to serve as an any-dimensional neural network that parameterizes functions on inputs of arbitrary size n using a fixed set of parameters. This capability fundamentally relies on the phenomenon of representation stability, which is discussed in greater detail in [40].

Incompatibility of IGN. 2-IGN is incompatible with the subspace $\mathbb{V}_{\text{dup}}^G$. First, its basis functions are not properly normalized, and therefore cannot be extended to functions on graphons. For instance, the fourth basis function $\ell_4(A) = A\mathbb{1}\mathbb{1}^\top$ yields output entries of order $O(n)$, and should thus be normalized by a factor of n^{-1} . To address this issue, [7] introduces a normalized version of 2-IGN, defined by

$$\begin{aligned} W_n^{(i)}(A) = & \alpha_1 A + \alpha_2 A^\top + \alpha_3 \text{diag}(\text{diag}^*(A)) + \alpha_4 \frac{1}{n} A \mathbb{1} \mathbb{1}^\top + \alpha_5 \frac{1}{n} \mathbb{1} \mathbb{1}^\top A^\top + \alpha_6 \frac{1}{n} \text{diag}(A \mathbb{1}) \\ & + \alpha_7 \frac{1}{n} A^\top \mathbb{1} \mathbb{1}^\top + \alpha_8 \frac{1}{n} \mathbb{1} \mathbb{1}^\top A^\top + \alpha_9 \frac{1}{n} \text{diag}(A^\top \mathbb{1}) + \alpha_{10} \frac{1}{n^2} (\mathbb{1}^\top A \mathbb{1}) \mathbb{1} \mathbb{1}^\top \\ & + \alpha_{11} \frac{1}{n^2} (\mathbb{1}^\top A \mathbb{1}) \text{diag}(\mathbb{1}) + \alpha_{12} (\mathbb{1}^\top \text{diag}^*(A)) \mathbb{1} \mathbb{1}^\top + \alpha_{13} (\mathbb{1}^\top \text{diag}^*(A)) \text{diag}(\mathbb{1}) \\ & + \alpha_{14} \text{diag}^*(A) \mathbb{1}^\top + \alpha_{15} \mathbb{1} \text{diag}^*(A)^\top + \beta_1 \mathbb{1} \mathbb{1}^\top + \beta_2 \text{diag}(\mathbb{1}). \end{aligned} \quad (18)$$

However, the normalized 2-IGN is still not compatible. Consider the third basis function $\ell_3(A) := \text{diag}(\text{diag}^*(A))$. It fails to satisfy the compatibility condition:

$$\ell_3(A \otimes \mathbb{1}_m) \neq \ell_3(A) \otimes \mathbb{1}_m, m \geq 2,$$

as the left-hand side yields a diagonal matrix, while the right-hand side generally does not. In fact, all basis maps that output diagonal matrices share this incompatibility.

Nonetheless, our Proposition 5.1 immediately provides a constructive recipe for making 2-IGN transferable: we start from a basis for linear equivariant layers $W_n^{(i)}$ which is compatible under duplication, and then select only those basis elements which have a finite operator norm as n grows. Furthermore, we use nonlinearities $\rho_n^{(i)}$ which are compatible and Lipschitz continuous. Following this recipe, we introduce two modified versions of 2-IGN:

Generalizable Graph Neural Network (GGNN): Compatible with respect to $\mathbb{V}_{\text{dup}}^G$, locally Lipschitz transferable under the ∞ -norm.

Continuous GGNN: Compatible with respect to $\mathbb{V}_{\text{dup}}^G$, locally Lipschitz transferable under the operator 2-norm, and continuously transferable under the cut-norm.

We highlight that this is a general methodology for constructing transferable equivariant networks: the framework established in [40] yields bases for compatible equivariant linear layers. We can then select only those basis elements whose operator norms do not grow with dimension, which we have shown yields a transferable neural network.

GGNN Architecture. A D -layer GGNN parameterizes an S_n -equivariant function

$$\mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d'_1} \rightarrow \mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^{n \times d'_D}$$

as a composition:

$$W_n^{(D)} \circ \rho_n^{(D-1)} \circ \dots \circ \rho_n^{(1)} \circ W_n^{(1)},$$

where for each i :

- $W_n^{(i)} : (\mathbb{R}_{\text{sym}}^{n \times n})^{\oplus d_i} \oplus (\mathbb{R}^n)^{\oplus d'_i} \rightarrow (\mathbb{R}_{\text{sym}}^{n \times n})^{\oplus d_i} \oplus ((\mathbb{R}^n)^{\oplus d'_i})^{\oplus S}$ is a linear S_n -equivariant map that is compatible with the duplication embedding.
- $\rho_n^{(i)} : (\mathbb{R}_{\text{sym}}^{n \times n})^{\oplus d_i} \oplus ((\mathbb{R}^n)^{\oplus d'_i})^{\oplus S} \rightarrow (\mathbb{R}_{\text{sym}}^{n \times n})^{\oplus d_i} \oplus (\mathbb{R}^n)^{\oplus d'_i}$ is the nonlinearity,

d, d' are feature channels of A, X respectively, where we fix $d_1 = d_D = 1$.

1820 For the ease of notation, we assume $d_i = d_{i+1} = 1$ (The general case follows analogously). The
 1821 maps $W_n^{(i)}, \rho_n^{(i)}$ are given by

$$\begin{aligned}
 W_n^{(i)}(A, X) &= (A', (X'_s)_{s=0}^S) \\
 &= \left(\alpha_1 A + \alpha_2 \frac{\mathbb{1}^\top A \mathbb{1}}{n^2} \mathbb{1} \mathbb{1}^\top + \alpha_3 \frac{\text{Tr}(A)}{n} \mathbb{1} \mathbb{1}^\top + \alpha_4 \frac{1}{n} (A \mathbb{1} \mathbb{1}^\top + \mathbb{1} \mathbb{1}^\top A) \right. \\
 &\quad + \alpha_5 (\text{diag}(A) \mathbb{1} \mathbb{1}^\top + \mathbb{1} \text{diag}(A_1)^\top) + \sum_{j=1}^{d'_i} \left[\alpha_{6,j} (X_{:,j} \mathbb{1}^\top + \mathbb{1} X_{:,j}^\top) + \alpha_{7,j} \frac{1}{n} (\mathbb{1}^\top X_{:,j}) \mathbb{1} \mathbb{1}^\top \right] \\
 &\quad + \beta_1 \mathbb{1} \mathbb{1}^\top, \\
 &\quad \left. X \Theta_{1,s} + \frac{1}{n} \mathbb{1} \mathbb{1}^\top X \Theta_{2,s} + \frac{1}{n} A \mathbb{1} \theta_{1,s}^\top + \text{diag}(A) \theta_{2,s}^\top + \frac{\text{Tr}(A)}{n} \mathbb{1} \theta_{3,s}^\top + \frac{\mathbb{1}^\top A \mathbb{1}}{n^2} \mathbb{1} \theta_{4,s}^\top + \mathbb{1} \beta_{2,s}^\top \right), \\
 \rho_n^{(i)}(A, (X_s)_{s=0}^S) &= \left(A, \sigma \left(\sum_{s=0}^S n^{-s} A^s X_s \right) \right)
 \end{aligned} \tag{19}$$

1822 where α, θ, β are learnable parameters, and $\sigma: \mathbb{R} \rightarrow \mathbb{R}$ is an arbitrary L -Lipschitz entrywise nonlin-
 1823 earity.

1824 Consider the input and output spaces as (variants of) $\mathbb{V}_{\text{dup}}^G$, the duplication consistent sequences for
 1825 graph signals. The linear layer W_n in (19) parameterizes all linear S_n -equivariant maps between
 1826 these two spaces that are also compatible with the duplication embedding.

1827 The GGNN model is a modification of the 2-IGN (17) with the following key differences:

- 1828 • We treat the adjacency matrix and node features separately so that each layer has a graph and a
 1829 signal component. Moreover, we explicitly require the matrix component to be symmetric.
- 1830 • We impose the compatibility with respect to the duplication embedding on the linear layers.
 1831 This leads to both proper normalization of each basis function and a reduction in the total
 1832 number of basis functions. Particularly, all basis functions that output a diagonal matrix are
 1833 removed.
- 1834 • For the nonlinearity ρ_n , instead of the entrywise nonlinearity used in IGN, we adopt a message-
 1835 passing-like nonlinearity. This structure mirrors the GNN model studied in [56]. Therefore, our
 1836 model is at least as expressive as the GNN in [56].

1837 **Transferability analysis of GGNN.** Even though we only impose compatibility by design, we can
 1838 still prove that GGNN is Lipschitz transferable with respect to some norm, even though this norm is
 1839 arguably too weak.

1840 **Corollary G.4.** Consider one layer of GGNN, $(\text{GGNN}_n(A, X) = \rho_n^{(i)} \circ W_n^{(i)})$, as defined in (19),
 1841 where the entrywise nonlinearity σ is L_σ -Lipschitz. Endow the space of duplication-consistent
 1842 sequences with the ∞ -norm as defined in (10) (with respect to $\|\cdot\|_\infty$ on $\mathbb{R}^{d'_i}$). Then, the sequence of
 1843 maps (GGNN_n) is locally Lipschitz transferable.

1844 Therefore, (GGNN_n) extends to a function GGNN_∞ on the space of graphon signals, which is
 1845 $L(r)$ -Lipschitz on $B(0, r)$ with respect to the symmetrized metric defined in (11) with $p = \infty$.

1846 *Proof.* The sequence of linear maps $(W_n^{(i)})$ in (19) is Lipschitz transferable because

$$\begin{aligned}
 \|W_n^{(i)}\|_{\text{op}} &\leq \max \left\{ |\alpha_1| + |\alpha_2| + |\alpha_3| + 2|\alpha_4| + 2|\alpha_5| + \sum_{j=1}^{d'_i} (|\alpha_{6,j}| + |\alpha_{7,j}|), \right. \\
 &\quad \left. \|\Theta_{1,s}\|_{\text{op},1,1} + \|\Theta_{2,s}\|_{\text{op},1,1} + \|\theta_{1,s}\|_\infty + \|\theta_{2,s}\|_\infty + \|\theta_{3,s}\|_\infty + \|\theta_{4,s}\|_\infty \right\},
 \end{aligned}$$

1847 where $\|\Theta\|_{\text{op},1,1} = \max_j \sum_k |\theta_{k,j}|$ is the max ℓ_1 norm of a column.

1848 For the nonlinearity $(\rho_n^{(i)})$, we consider its Fréchet derivative (since all norms are equivalent in
1849 finite-dimensional vector spaces, this is independent of the norm chosen):

$$D\rho_n^{(i)}(A, X_0, \dots, X_S)[H, H_0, \dots, H_S] = \left(H, \sum_{s=0}^S n^{-s} \left(\sum_{k=0}^{s-1} A^k H A^{s-1-k} \cdot X_s + A^s H_s \right) \right).$$

1850 Hence,

$$\begin{aligned} & \|D\rho_n^{(i)}(A, X_0, \dots, X_S)[H, H_0, \dots, H_S]\|_\infty \\ & \leq \max \left(\|H\|_\infty, \sum_{s=0}^S \left(\sum_{k=0}^{s-1} \|A\|_\infty^{s-1} \|H\|_\infty \|X_s\|_\infty + \|A\|_\infty^s \|H_s\|_\infty \right) \right) \\ & \leq \max \left(\|H\|_\infty, \sum_{s=0}^S s r^s \|H\|_\infty + r^s \|H_s\|_\infty \right) \\ & \leq \underbrace{\left(1 \vee \sum_{s=0}^S (s r^s + r^s) \right)}_{=: L(r)} \cdot \|(H, H_0, \dots, H_S)\|_\infty. \end{aligned}$$

1851 Therefore, for all n , $\rho_n^{(i)}$ is $L_\sigma L(r)$ -Lipschitz on the set

$$B_n(0, r) = \{(A, X_0, \dots, X_S) : \|A, X_0, \dots, X_S\|_\infty < r\}.$$

1852 Applying Proposition 5.1, the sequence of maps (GGNN_n) is locally Lipschitz transferable, where
1853 the extension GGNN_∞ is $(L_\sigma L(\|W_n\|_{\text{op}} r) \|W_n\|_{\text{op}})$ -Lipschitz on the set

$$B(0, r) = \{(W, f) : \|(W, f)\|_\infty < r\}.$$

1854

□

1855 **Continuous GGNN architecture.** We aim to further restrict GGNN to construct a model that is
1856 transferable with respect to the cut norm. By Proposition G.1, we consider endowing the consistent
1857 sequence with the operator 2-norm, which is easier to analyze. The *Continuous GGNN* is a variant
1858 of GGNN with an additional constraint on the linear layers $W_n^{(i)}$, requiring them to have bounded
1859 operator norm: $\|W_n\|_{\text{op}} < \infty$ (with respect to the operator 2-norm). This constraint effectively leads
1860 to a further reduction in the set of basis functions:

$$\begin{aligned} W_n^{(i)}(A, X) = (A', (X'_s)_{s=0}^S) = & \left(\alpha_1 A + \alpha_2 \frac{\mathbb{1}^\top A \mathbb{1}}{n^2} \mathbb{1} \mathbb{1}^\top + \alpha_4 \frac{1}{n} (A \mathbb{1} \mathbb{1}^\top + \mathbb{1} \mathbb{1}^\top A) \right. \\ & + \sum_{j=1}^k \left[\alpha_{6,j} (X_{:,j} \mathbb{1}^\top + \mathbb{1} X_{:,j}^\top) + \alpha_{7,j} \frac{1}{n} (\mathbb{1}^\top X_{:,j}) \mathbb{1} \mathbb{1}^\top \right], \quad (20) \\ & \left. X \Theta_{1,s} + \frac{1}{n} \mathbb{1} \mathbb{1}^\top X \Theta_{2,s} + \frac{1}{n} A \mathbb{1} \theta_{1,s}^\top + \frac{\mathbb{1}^\top A \mathbb{1}}{n^2} \mathbb{1} \theta_{4,s}^\top \right), \end{aligned}$$

1861 Therefore, the hypothesis class of continuous GGNN forms a strict subset of that of GGNN, with the
1862 additional constraint enabling improved transferability. We use (cGGNN_n) to denote the sequence
1863 of functions of continuous GGNN.

1864 **Transferability analysis of Continuous GGNN.**

1865 **Corollary G.5.** Consider one layer of the continuous GGNN, $(\text{cGGNN}_n(A, X) = \rho_n^{(i)} \circ W_n^{(i)})$, as
1866 defined in (20), where the entrywise nonlinearity σ is L_σ -Lipschitz. Endow the space of duplication-
1867 consistent sequences with the operator 2-norm as defined in (12) (with respect to $\|\cdot\|_\infty$ on \mathbb{R}^{d_i}).
1868 Then, the sequence of maps (cGGNN_n) is locally Lipschitz transferable.

1869 Therefore, (cGGNN_n) extends to a function cGGNN_∞ on the space of graphon signals, which is
1870 $L(r)$ -Lipschitz on $B(0, r)$ with respect to the symmetrized operator 2-metric defined in (13) with
1871 $p = 2$.

1872 **Remark G.6.** By Proposition G.1, the sequence of maps (cGGNN_n) is continuously transferable
 1873 with respect to the cut distance (15) on the space

$$\{(W, f) : \|(W, f)\|_{\text{op},2} < r, \|W\|_\infty, \|f\|_\infty < r\}.$$

1874 Moreover, for the convergence and transferability results stated in Proposition D.4, D.6, one can
 1875 additionally obtain quantitative rates of convergence with respect to the cut distance.

1876 *Proof.* First, by construction, the sequence of maps $(W_n^{(i)})$ is Lipschitz transferable because

$$\|W_n^{(i)}\|_{\text{op}} \leq \max \left\{ |\alpha_1| + |\alpha_2| + 2|\alpha_4| + \sum_{j=1}^{d'_i} (|\alpha_{6,j}| + |\alpha_{7,j}|), \right. \\ \left. \|\Theta_{1,s}\|_{\text{op},1} + \|\Theta_{2,s}\|_{\text{op},1} + \|\theta_{1,s}\|_\infty + \|\theta_{4,s}\|_\infty \right\} < \infty.$$

1877 For the nonlinearity $(\rho_n^{(i)})$, we again consider its Fréchet derivative:

$$D\rho_n^{(i)}(A, X_0, \dots, X_S)[H, H_0, \dots, H_S] = \left(H, \sum_{s=0}^S n^{-s} \left(\sum_{k=0}^{s-1} A^k H A^{s-1-k} \cdot X_s + A^s H_s \right) \right).$$

1878 Hence,

$$\|D\rho_n^{(i)}(A, X_0, \dots, X_S)[H, H_0, \dots, H_S]\|_{\text{op},2} \\ \leq \max \left(n^{-1} \|H\|_{\text{op},2}, \sum_{s=0}^S \left(\sum_{k=0}^{s-1} \frac{\|A\|_{\text{op},2}^{s-1}}{n^{s-1}} \cdot \frac{\|H\|_{\text{op},2}}{n} \cdot \|X_s\|_2 + \frac{\|A\|_{\text{op},2}^s}{n^s} \cdot \|H_s\|_2 \right) \right) \\ \leq \max \left(n^{-1} \|H\|_{\text{op},2}, \sum_{s=0}^S sr^s \cdot \frac{\|H\|_{\text{op},2}}{n} + r^s \|H_s\|_2 \right) \\ \leq \underbrace{\left(1 \vee \sum_{s=0}^S (sr^s + r^s) \right)}_{=: L(r)} \cdot \|(H, H_0, \dots, H_S)\|_{\text{op},2}.$$

1879 Therefore, for all n , $\rho_n^{(i)}$ is $L_\sigma L(r)$ -Lipschitz on the set

$$\{(A, X_0, \dots, X_S) : \|A, X_0, \dots, X_S\|_{\text{op},2} < r\}.$$

1880 Applying Proposition 5.1, the sequence of maps (f_n) is locally Lipschitz transferable, where the
 1881 extension cGGNN_∞ is $(L_\sigma L(\|W_n\|_{\text{op}} r) \|W_n\|_{\text{op}})$ -Lipschitz on the set

$$B(0, r) = \{(W, f) : \|(W, f)\|_{\text{op},2} < r\}.$$

1882 □

1883 **Related Work.** We discuss two closely related works, [7] and [45], that address the transferability
 1884 of IGNs. Interpreting their results within our theoretical framework offers a better understanding of
 1885 IGN transferability.

1886 As shown in our work, the normalized 2-IGN is not compatible with the duplication-consistent
 1887 subspace $\mathbb{V}_{\text{dup}}^G$, and thus fails to satisfy the convergence and transferability in Proposition 3.2. At
 1888 first glance, this observation may seem to contradict [7, Theorem 2]. However, this is not the case.
 1889 While [7] introduces cIGN, a “graphon analogue of IGN,” and proves its continuity in the graphon
 1890 space, it is crucial to note that the discrete IGN does not extend to cIGN in general:

$$\text{IGN}_n(A_n, X_n) \neq \text{cIGN}([A_n], [X_n]).$$

1891 Therefore, the convergence of cIGN established in Theorem 2 of [7] does not imply the convergence
 1892 or transferability of the discrete IGN model.

Moreover, [7, Definition 6] introduces a constraint that resembles our compatibility condition, formulated through a restricted variant termed “IGN-small.” Our definition of compatibility clarifies this notion and enables explicit constructions and practical implementations of compatible, transferable versions of IGNs.

In a more recent work, [27] adopts an approach similar to ours by placing additional constraints on the linear layers of IGN, specifically requiring them to have bounded operator norm. This leads to the Invariant Graphon Network (IWN) model. However, unlike our construction, IWN retains the standard point-wise nonlinearity. As shown in [27, Proposition 5.5], this point-wise nonlinearity results in discontinuity with respect to the cut norm. Our continuous CCGN model resolves this issue by replacing the point-wise nonlinearity with a message-passing-like operator, yielding continuity with respect to the cut norm.

Finally, while [27, Theorem 5.6] argues that IWN is still transferable due to its approximability by continuous functions, we emphasize that the form of “transferability” in [27] does not guarantee a vanishing error. In particular, we suspect that in [27, Theorem 5.6], as the approximation precision ϵ decreases, $C_{\epsilon, \mathcal{N}}$ diverges, implying that the approximation error does not vanish in the asymptotic limit. This distinction underscores the necessity of continuity for true transferability. As illustrated in Figure 2(c), discontinuous models exhibit an irreducible asymptotic error, in contrast to the continuous and transferable models depicted in Figures 2(a) and (d).

H Example 3 on point clouds: details and missing proofs from Section 5.3

H.1 Duplication consistent sequence for point clouds

The duplication consistent sequence for point clouds $\mathbb{V}_{\text{dup}}^P = \{(V_n), (\varphi_{N,n}), (G_n)\}$ is defined as follows. The index set is again $\mathbb{N} = (\mathbb{N}, \cdot \mid \cdot)$. For each n , the vector spaces are $V_n = \mathbb{R}^{n \times k}$, with the group $G_n = S_n \times O(k)$ acting on V_n by

$$(g, h) \cdot X = gXh^\top.$$

For any $n \mid N$, the embedding is given by,

$$\begin{aligned} \varphi_{N,n}: \mathbb{R}^{n \times k} &\hookrightarrow \mathbb{R}^{N \times k} \\ X &\mapsto X \otimes \mathbb{1}_{N/n}, \end{aligned}$$

and the group embedding is

$$\begin{aligned} \theta_{N,n}: S_n \times O(k) &\hookrightarrow S_N \times O(k) \\ (g, h) &\mapsto (g \otimes I_{N/n}, h). \end{aligned}$$

Analogous to the case of sets, we can identify each matrix $X \in \mathbb{R}^{n \times k}$ with a step function $f_X: [0, 1] \rightarrow \mathbb{R}^k$, thereby interpreting V_∞ as the space of step functions with discontinuities at rational points \mathbb{Q} . We also view the orbit of X as an empirical probability measure $\frac{1}{n} \sum_{i=1}^n \delta_{X_{i:}}$. Equivalently, this identifies the orbit of a step function $f \in V_\infty$ with $\mu_f = \text{Law}(f(T))$ where $T \sim \text{Unif}[0, 1]$.

The orthogonal group $O(k)$ acts on probability measures via push-forward: for $g \in O(k)$ and a measure μ , the action is given by $g \cdot \mu = g^\# \mu$, where $g^\# \mu(B) = \mu(g^{-1}(B))$ for all measurable sets $B \subseteq \mathbb{R}^k$. The orbit space of V_∞ can be identified with the orbit space of empirical probability measures on \mathbb{R}^k under the action of $O(k)$.

Norm on $\mathbb{V}_{\text{dup}}^P$. Consider Euclidean norm $\|\cdot\|_2$ on \mathbb{R}^k which corresponds to the inner product preserved by elements of $O(k)$. We equip each V_n with the normalized ℓ_p norm:

$$\|X\|_{\bar{p}} = \begin{cases} \left(\frac{1}{n} \sum_{i=1}^n \|X_{i:}\|_2^p \right)^{1/p} & p \in [1, \infty) \\ \max_{i=1}^n \|X_{i:}\|_2 & p = \infty \end{cases}$$

By Proposition C.8, it is straightforward to verify that this norm extends naturally to V_∞ , and that the limit space in this case can be identified with $\overline{V_\infty} = L^p([0, 1]; \mathbb{R}^k)$ of functions $f: [0, 1] \rightarrow \mathbb{R}^k$ with norm $\|f\| = \left(\int_0^1 \|f(t)\|_2^p dt \right)^{1/p} < \infty$.

1931 Analogous to the case of sets, the corresponding space of orbit closures can be identified with
 1932 the space of orbit closures of probability measures on \mathbb{R}^k (with finite p -th moments) under the
 1933 $O(k)$ -actions. The symmetrized metric is given by:

$$\bar{d}_p(f, g) = \inf_{g \in O(k)} W_p(g \cdot \mu_f, \mu_g) \quad \text{for } f, g \in \overline{V_\infty}, \quad (21)$$

1934 where W_p is the Wasserstein p -distance with respect to the ℓ_2 -norm on \mathbb{R}^k .

1935 H.2 Invariant networks on point clouds

1936 H.2.1 DeepSet for Conjugation Invariance (DS-CI)

1937 The DS-CI model proposed in [2] is given by

$$\begin{aligned} \text{DS-CI}_n : \mathbb{R}^{n \times k} &\rightarrow \mathbb{R} \\ V &\mapsto \text{MLP}_c \left(\text{DeepSet}_{(1)} \left(\{f_j^d(VV^\top)\}_{j=1, \dots, n} \right), \right. \\ &\quad \text{DeepSet}_{(2)} \left(\{f_\ell^o(VV^\top)\}_{\ell=1, \dots, n(n-1)/2} \right), \\ &\quad \left. \text{MLP}_{(3)} \left(f^*(VV^\top) \right) \right), \end{aligned}$$

1938 where for symmetric matrix $X \in \mathbb{R}_{\text{sym}}^{n \times n}$, the invariant features are given by $f_j^d(X)$ = the j -th largest
 1939 of the numbers X_{11}, \dots, X_{nn} , by $f_\ell^o(X)$ = the ℓ -th largest of the numbers X_{ij} , $1 \leq i < j \leq n$, and
 1940 by $f^*(X) = \sum_{i \neq j} X_{ii} X_{ij}$.

1941 We define *normalized DS-CI* with appropriate normalization: replacing $\text{DeepSet}_{(1)}$, $\text{DeepSet}_{(2)}$
 1942 with their normalized version (i.e. replacing the sum aggregation with the mean aggregation), and
 1943 replacing $f^*(X)$ with $\bar{f}^*(X) = \frac{1}{n(n-1)} \sum_{i \neq j} X_{ii} X_{ij}$. We denote the sequence of functions of
 1944 normalized DS-CI by $(\overline{\text{DS-CI}}_n)$.

1945 **Transferability analysis of normalized DS-CI.** Normalized DS-CI is not compatible with respect
 1946 to $\mathbb{V}_{\text{dup}}^P$. To see this, observe that under duplication, we have

$$(V \otimes \mathbb{1}_{N/n})(V \otimes \mathbb{1}_{N/n})^\top = (VV^\top) \otimes (\mathbb{1}_{N/n} \mathbb{1}_{N/n}^\top),$$

1947 Therefore, diagonal elements of VV^\top become off-diagonal elements in $(VV^\top) \otimes (\mathbb{1}_{N/n} \mathbb{1}_{N/n}^\top)$, so

$$\overline{\text{DeepSet}}_{(2)} \left(\{f_\ell^o((V \otimes \mathbb{1}_{N/n})(V \otimes \mathbb{1}_{N/n})^\top)\}_{\ell=1}^{N(N-1)/2} \right) \neq \overline{\text{DeepSet}}_{(2)} \left(\{f_\ell^o(VV^\top)\}_{\ell=1}^{n(n-1)/2} \right),$$

1948

$$\bar{f}^*((V \otimes \mathbb{1}_{N/n})(V \otimes \mathbb{1}_{N/n})^\top) \neq \bar{f}^*(VV^\top).$$

1949 However, we can make some additional adjustments to ensure compatibility: we define the *compatible*
 1950 *DS-CI* by modifying the inputs of $\overline{\text{DeepSet}}_{(2)}$ to be $\{f_l^a(VV^\top)\}_{l=1, \dots, n^2}$, where $f_l^a(X)$ denotes the
 1951 l -th largest value among the entries X_{ij} for $1 \leq i, j \leq n$. Additionally, we replace $\bar{f}^*(X)$ with

$$\tilde{f}^*(X) := \frac{1}{n^2} \sum_{i,j} X_{ii} X_{ij}.$$

1952 We denote the sequence of functions of compatible DS-CI by (C-DS-CI_n) . We prove that this model
 1953 is locally Lipschitz transferable.

1954 **Corollary H.1.** Endow $\mathbb{V}_{\text{dup}}^P$ with the normalized ℓ_p norm with $p \in [1, \infty]$. Assume that all activation
 1955 functions in the MLPs used for DS-CI are Lipschitz. Then the sequence of maps (C-DS-CI_n) is $L(r)$ -
 1956 locally Lipschitz transferable with respect to $\|\cdot\|_{\bar{p}}$. In particular, it is locally Lipschitz transferable
 1957 with respect to $\|\cdot\|_\infty$.

1958 Therefore, (C-DS-CI_n) defines a function C-DS-CI_∞ on $\mathcal{P}_p(\mathbb{R}^k)$ which is $L(r)$ -Lipschitz on $B(0, r)$
 1959 with respect to the symmetrized Wasserstein metric defined in (21).

1960 *Proof.* By Proposition 5.1, it is sufficient to verify the compatibility and Lipschitz continuity of each
 1961 individual layer.

1962 • The sequence of maps

$$(\mathbb{R}^{n \times k}, \|\cdot\|_{\bar{p}}) \rightarrow (\mathbb{R}^n, \|\cdot\|_{\bar{p}}), \quad V \mapsto \text{diag}(VV^\top)$$

1963 is $(2r)$ -locally Lipschitz transferable. Indeed, it is S_n -equivariant, $O(k)$ -invariant, and

$$\begin{aligned} \text{diag}((V \otimes \mathbb{1}_m)(V \otimes \mathbb{1}_m)^\top) &= \text{diag}((VV^\top) \otimes (\mathbb{1}_m \mathbb{1}_m^\top)) \\ &= \text{diag}(VV^\top) \otimes \mathbb{1}_m, \end{aligned}$$

$$\begin{aligned} \|\text{diag}(VV^\top) - \text{diag}(WW^\top)\|_{\bar{p}} &= \left(\frac{1}{n} \sum_{i=1}^n \|\|V_{i:}\|_2^2 - \|W_{i:}\|_2^2\|^p \right)^{1/p} \\ &= \left(\frac{1}{n} \sum_{i=1}^n |\langle V_{i:} - W_{i:}, V_{i:} + W_{i:} \rangle|^p \right)^{1/p} \\ &\leq \left(\frac{1}{n} \sum_{i=1}^n \|V_{i:} - W_{i:}\|_2^p \cdot (2r)^p \right)^{1/p} \\ &\quad \text{(by Cauchy-Schwarz)} \\ &= 2r \|V - W\|_{\bar{p}}. \end{aligned}$$

1964 • The sequence of maps

$$(\mathbb{R}^{n \times k}, \|\cdot\|_{\bar{p}}) \rightarrow (\mathbb{R}^{n^2}, \|\cdot\|_{\bar{p}}), \quad V \mapsto \text{vec}(VV^\top)$$

1965 is $(2r)$ -locally Lipschitz transferable, where the codomain is equipped with a consistent se-
 1966 quence structure as follows: for $g \in S_n$, define $\pi(g) = g^\top \otimes g \in S_{n^2}$, and let g act on \mathbb{R}^{n^2}
 1967 by $g \cdot x = \pi(g)x$. The symmetric groups (S_n) are embedded into each other as usual, and the
 1968 vector spaces are embedded by $\varphi_{nm,n}: \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{(nm)^2}$ where

$$\varphi_{nm,n}(x) = \text{vec}(\text{reshape}_n(x) \otimes \mathbb{1}_m \mathbb{1}_m^\top),$$

1969 and $\text{reshape}_n: \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n \times n}$ is the inverse of vec on $n \times n$ matrices. Since these are all linear
 1970 maps, so is $\varphi_{nm,n}$. We then have for all $V \in \mathbb{R}^{n \times k}, g \in S_n, h \in O(k)$ that

$$\begin{aligned} \text{vec}((V \otimes \mathbb{1}_m)(V \otimes \mathbb{1}_m)^\top) &= \text{vec}((VV^\top) \otimes (\mathbb{1}_m \mathbb{1}_m^\top)) = \varphi_{nm,n}(\text{vec}(VV^\top)), \\ \text{vec}((gVh^\top)(gVh^\top)^\top) &= \text{vec}(gVV^\top g^\top) = \pi(g)\text{vec}(VV^\top). \end{aligned}$$

1971 Furthermore,

$$\begin{aligned} \|\text{vec}(VV^\top) - \text{vec}(WW^\top)\|_{\bar{p}} &= \left(\frac{1}{n^2} \sum_{i,j} |\langle V_{i:}, V_{j:} \rangle - \langle W_{i:}, W_{j:} \rangle|^p \right)^{1/p} \\ &\leq \left(\frac{1}{n^2} \sum_{i,j} (|\langle V_{i:}, V_{j:} - W_{j:} \rangle| + |\langle V_{i:} - W_{i:}, W_{j:} \rangle|)^p \right)^{1/p} \\ &\leq \left(\frac{1}{n^2} \sum_{i,j} 2^{p-1} r^p (\|V_{j:} - W_{j:}\|_2^p + \|V_{i:} - W_{i:}\|_2^p) \right)^{1/p} \\ &\quad \text{(Using } (a+b)^p \leq 2^{p-1}(a^p + b^p) \text{ and Cauchy-Schwarz)} \\ &= 2r \|V - W\|_{\bar{p}}. \end{aligned}$$

1972 • The scalar maps

$$(\mathbb{R}^{n \times k}, \|\cdot\|_{\bar{p}}) \rightarrow (\mathbb{R}, |\cdot|), \quad V \mapsto \tilde{f}^*(VV^\top)$$

1973 is $(4r^3)$ -locally Lipschitz transferable. Indeed, it is $S_n \times O(k)$ invariant and

$$\begin{aligned} \tilde{f}^*((V \otimes \mathbb{I}_m)(V \otimes \mathbb{I}_m)^\top) &= \tilde{f}^*(VV^\top), \\ \left| \tilde{f}^*(VV^\top) - \tilde{f}^*(WW^\top) \right| & \\ &\leq \frac{1}{n^2} \sum_{i,j} |\langle V_{i:}, V_{j:} \rangle \|V_{i:}\|_{\mathbb{R}^k}^2 - \langle W_{i:}, W_{j:} \rangle \|W_{i:}\|_{\mathbb{R}^k}^2| \\ &\leq \frac{1}{n^2} \sum_{i,j} |\langle V_{i:}, V_{j:} \rangle - \langle W_{i:}, W_{j:} \rangle| \|V_{i:}\|_{\mathbb{R}^k}^2 + |\langle W_{i:}, W_{j:} \rangle| |\|V_{i:}\|_{\mathbb{R}^k}^2 - \|W_{i:}\|_{\mathbb{R}^k}^2| \\ &\leq 4r^3 \|V - W\|_{\bar{1}} \quad (\text{by the result of the previous two computations}) \\ &\leq 4r^3 \|V - W\|_{\bar{p}}. \end{aligned}$$

1974 • If the activation functions used are Lipschitz, then MLPs are Lipschitz. By Corollary F.4, the
1975 normalized DeepSet is Lipschitz transferable, assuming the constituent MLPs are Lipschitz.

1976 Thus, our compatible DS-CI architecture is a composition of locally Lipschitz layers. \square

1977 Finally, we conclude that the normalized DS-CI is “approximately transferable” since it is asymptoti-
1978 cally equivalent to the compatible DS-CI up to an error of $O(n^{-1})$.

1979 **Lemma H.2.** *If the activations in all the MLPs used are Lipschitz, then for any sequence of inputs*
1980 $V^{(n)} \in \mathbb{R}^{n \times k}$, $|\text{C-DS-CI}_n(V^{(n)}) - \overline{\text{DS-CI}}_n(V^{(n)})| = O(n^{-1})$

1981 *Proof.* Assume for $x \in \mathbb{R}^n$, $\overline{\text{DeepSet}}_{(2)}(x) = \sigma(\frac{1}{n} \sum_i \rho(x_i))$, and σ, ρ are L_σ, L_ρ Lipschitz respec-
1982 tively. Then,

$$\begin{aligned} &\left| \overline{f^*}(VV^\top) - \tilde{f}^*(VV^\top) \right| \\ &\leq \left(\frac{1}{n(n-1)} - \frac{1}{n^2} \right) \sum_{i \neq j} |(VV^\top)_{ii}(VV^\top)_{ij}| + \frac{1}{n^2} \sum_i (VV^\top)_{ii}^2 = O(n^{-1}) \\ &\left| \overline{\text{DeepSet}}_{(2)}(\{f_\ell^o(VV^\top)\}_{\ell=1, \dots, n(n-1)/2}) - \overline{\text{DeepSet}}_{(2)}(\{f_\ell^a(VV^\top)\}_{\ell=1, \dots, n^2}) \right| \\ &\leq L_\sigma \left(\left(\frac{1}{n(n-1)} - \frac{1}{n^2} \right) \sum_{i \neq j} |\rho((VV^\top)_{ij})| + \frac{1}{n^2} \sum_i |\rho((VV^\top)_{ii})| \right) = O(n^{-1}). \end{aligned}$$

1983 Since every layer is Lipschitz, this leads to an overall error of $O(n^{-1})$. \square

1984 H.3 Constructing new transferable models: SVD-DS

1985 We propose the SVD-DS model defined as follows:

$$\overline{\text{SVD-DS}}_n : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}, \quad X \mapsto \overline{\text{DeepSet}}_n(XV),$$

1986 where $X = UDV^\top$ is the singular value decomposition (SVD) for X with ordered singular values.
1987 We proceed to show that it is locally transferable almost everywhere on its domain, and that its
1988 performance is competitive with DS-CI while being more computationally efficient.

1989 **Transferability analysis of SVD-DS.** We extend the SVD to elements in the limit space $\overline{V_\infty} =$
 1990 $L^2([0, 1], \mathbb{R}^k)$ and analyze its continuity, yielding the following transferability result.

1991 **Corollary H.3.** *Endow the duplication consistent sequences with the normalized ℓ_2 norm induced by*
 1992 *$\|\cdot\|_2$ on \mathbb{R}^k . Observe that*

$$\|X\|_2 := \left(\frac{1}{n} \sum_{i=1}^n \|X_{i:}\|^2 \right)^{1/2} = \frac{1}{\sqrt{n}} \|X\|_F,$$

1993 *where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Then, the sequence of maps $(\overline{\text{SVD-DS}}_n)$ is*
 1994 *compatible, and extends to a map $\overline{\text{SVD-DS}}_\infty$ defined on the limit space $\overline{V_\infty} = L^2([0, 1], \mathbb{R}^k)$, which*
 1995 *is Lipschitz at every point with distinct singular values.*

1996 **Remark H.4.** *This transferability result is weaker than the “ $L(r)$ -locally Lipschitz transferability”*
 1997 *defined in Definition 3.1, since our model may be discontinuous at points with non-distinct singular*
 1998 *values. Therefore, in this case our transferability results in Propositions D.4 and D.6 only apply to*
 1999 *sequences (x_n) converging to a limit $x \in \overline{V_\infty}$ with distinct singular values.*

2000 *Proof.* Decompose $\overline{\text{SVD-DS}}_n$ as the composition

$$\mathbb{V}_{\text{dup}}^P = \{\mathbb{R}^{n \times k}\} \xrightarrow{X \mapsto XV} \mathbb{U} = \{\mathbb{R}^{n \times k}\} \xrightarrow{\text{DeepSet}} \mathbb{V}_{\mathbb{R}},$$

2001 where $\mathbb{V}_{\mathbb{R}}$ is the trivial consistent sequence over \mathbb{R} , and the consistent sequence \mathbb{U} consists of vector
 2002 spaces $U_n = \mathbb{R}^{n \times k}$ under the duplication embedding $\otimes \mathbb{I}$. The group $S_n \times O(k)$ acts on U_n by
 2003 $(g, h) \cdot X = gX$, i.e., the action of $O(k)$ is trivial.

2004 By Corollary F.4, the normalized DeepSet map is Lipschitz transferable, assuming the constituent
 2005 MLPs are Lipschitz.

2006 It remains to show that the SVD-based map $X \mapsto XV$ extends to a function that is Lipschitz at every
 2007 point except on a set of measure zero. We show this in Proposition H.5 below after extending the
 2008 SVD to all of $\overline{V_\infty}$ and considering its ambiguities. \square

2009 **Functional SVD and its local Lipschitz continuity.** We can identify the space $\overline{V_\infty} = L^2([0, 1], \mathbb{R}^k)$
 2010 with $\mathcal{L}(L^2([0, 1]), \mathbb{R}^k)$, the space of bounded linear maps $L^2([0, 1]) \rightarrow \mathbb{R}^k$ endowed with the Hilbert-
 2011 Schmidt norm $\|\cdot\|_{HS}$. In more detail, each $X \in L^2([0, 1], \mathbb{R}^k)$ can be written as a sequence
 2012 of rows $X = (f_1, \dots, f_k)^\top$ where $f_i \in L^2([0, 1])$, and such X defines the bounded linear map
 2013 $Xf = (\langle f_1, f \rangle, \dots, \langle f_k, f \rangle)^\top$. Conversely, any bounded linear map $X: L^2([0, 1]) \rightarrow \mathbb{R}^k$ is of
 2014 the form $Xf = (\langle f_1, f \rangle, \dots, \langle f_k, f \rangle)^\top$ for some $f_1, \dots, f_k \in L^2([0, 1])$ which we view as the
 2015 columns of X , and $\|X\|_{HS}^2 = \sum_{i=1}^k \|f_i\|_2^2$. Here $V_n = \mathbb{R}^{n \times k}$ is viewed as the subspace of $\overline{V_\infty}$ with
 2016 piecewise-constant columns f_i on consecutive intervals of length $1/n$.

2017 Note that X vanishes identically on $\mathcal{V}_k = \text{span}\{f_1, \dots, f_k\}^\perp$, while $X: \mathcal{V}_k \rightarrow \mathbb{R}^k$ is a linear map
 2018 between finite-dimensional vector spaces and therefore admits a singular value decomposition. Thus,
 2019 there exists positive numbers $\sigma \in \mathbb{R}_{\geq 0}^k$, orthonormal $v_1, \dots, v_k \in \mathbb{R}^k$, and orthonormal functions
 2020 $u_1, \dots, u_k \in L^2([0, 1])$ satisfying

$$X = \sum_{i=1}^k \sigma_i \langle u_i, \cdot \rangle v_i. \quad (22)$$

2021 If $X \in \mathbb{R}^{n \times k}$ and $X = \sum_{i=1}^k \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^\top$, is the usual SVD of X , then $\sigma_i = \tilde{\sigma}_i / \sqrt{n}$, $v_i = \tilde{v}_i$, and
 2022 $u_i(t) = \sqrt{n} [\tilde{u}_i]_{\lceil nt \rceil}$ is the functional SVD of X as in (22). Conversely, if (22) is the functional SVD
 2023 of such an X then $X = \sum_{i=1}^k (\sigma_i \sqrt{n}) ([u_i(j/n)]_{j=1}^k / \sqrt{n}) v_i^\top$ is the usual SVD of X . Note that the
 2024 right singular vectors V are the same in both SVDs.

2025 If for any $X \in V_n$ we let $\sigma(X)$ be its (functional) singular values from (22) and $\tilde{\sigma}(X)$ be its usual
 2026 singular values, then

$$\|X\|_2^2 = \frac{1}{n} \|X\|_F^2 = \frac{1}{n} \sum_{i=1}^k \tilde{\sigma}_i(X)^2 = \sum_{i=1}^k \sigma_i(X)^2, \quad (23)$$

2027 and by Mirsky's inequality [50],

$$\|\sigma(X) - \sigma(Y)\|_2 = \frac{1}{\sqrt{n}} \|\tilde{\sigma}(X) - \tilde{\sigma}(Y)\|_2 \leq \frac{1}{\sqrt{n}} \|X - Y\|_F = \|X - Y\|_{\bar{2}}. \quad (24)$$

2028 Furthermore, whenever $V \in \mathbb{R}^{k \times k}$ and $X \in V_n$ we have

$$\|XV\|_{\bar{2}} = \frac{1}{\sqrt{n}} \|XV\|_F \leq \frac{1}{\sqrt{n}} \|V\|_F \tilde{\sigma}_1(X) = \|V\|_F \sigma_1(X). \quad (25)$$

2029 Note that the final bounds in all of the above inequalities are independent of n , continuous in $\|\cdot\|$,
 2030 and hold for all $X \in V_\infty$. We therefore conclude that they also hold for all $X \in \overline{V_\infty}$, with $\|\cdot\|_{\bar{2}}$
 2031 replaced with $\|\cdot\|_{HS}$.

2032 There is an ambiguity in the above decomposition, since if $(u_i), (v_i)$ satisfy (22) then so do
 2033 $(s_i u_i), (s_i v_i)$ for any choice of signs $s_i \in \{\pm 1\}$. Furthermore, if the singular values (σ_i) are
 2034 distinct then this is the only ambiguity in (22), see [9]. To disambiguate the SVD, we therefore
 2035 choose signs so that $v_i > -v_i$ in lexicographic order. When the entries of the v_i are all nonzero, this
 2036 amounts to requiring the first row of $V = [v_1, \dots, v_k]$ to be positive. We proceed to prove that the
 2037 map $X \mapsto V(X) = [v_1, \dots, v_k]$ with this choice of signs is locally Lipschitz continuous on a dense
 2038 subset of $\overline{V_\infty}$.

2039 **Proposition H.5.** Fix $X_0 \in \overline{V_\infty}$ with distinct singular values and all-nonzero entries in its right
 2040 singular vectors (v_i) . Let $\text{gap}_p(X_0) = \min_{2 \leq i \leq k} \{\sigma_{i-1}(X_0)^p - \sigma_i(X_0)^p\}$ be the minimum gap
 2041 between p -th powers of (functional) singular values of X_0 , and set

$$B(X_0) = \frac{\sqrt{8}(2\sigma_1(X_0) + 1)}{\text{gap}_2(X_0)}. \quad (26)$$

2042 For any $\hat{X} \in \overline{V_\infty}$ satisfying

$$\|X_0 - \hat{X}\|_{HS} \leq R(X_0) := 1 \wedge \frac{\text{gap}_1(X_0)}{2\sqrt{k}} \wedge \frac{1}{2B(X_0)} \min_{1 \leq i, j \leq k} |[v_i(X_0)]_j|,$$

2043 we have:

2044 (1) \hat{X} has distinct singular values, and all nonzero entries of $v_i(\hat{X})$ have the same sign as those of
 2045 $v_i(X_0)$.

2046 (2) We have

$$\|V(X_0) - V(\hat{X})\|_F \leq kB(X_0)\|X_0 - \hat{X}\|_{HS}, \quad (27)$$

2047 and

$$\|X_0 V(X_0) - \hat{X} V(\hat{X})\|_{HS} \leq (k\sigma_1(X_0)B(X_0) + 1)\|X_0 - \hat{X}\|_{HS}. \quad (28)$$

2048 *Proof.* (1) If $\|X_0 - \hat{X}\|_{HS} \leq \frac{\text{gap}_1(X_0)}{2\sqrt{k}}$, then for any $2 \leq i \leq k$ we have by

$$\begin{aligned} \sigma_{i-1}(\hat{X}) - \sigma_i(\hat{X}) &= \sigma_{i-1}(\hat{X}) - \sigma_{i-1}(X_0) + \sigma_{i-1}(X_0) - \sigma_i(X_0) + \sigma_i(X_0) - \sigma_i(\hat{X}) \\ &\geq (\sigma_{i-1}(X_0) - \sigma_i(X_0)) - \sum_{i=1}^k |\sigma_i(\hat{X}) - \sigma_i(X_0)| \\ &\geq \text{gap}_1(X_0) - \sqrt{k} \|\sigma(X_0) - \sigma(\hat{X})\|_2 && \text{(Cauchy-Schwarz)} \\ &\geq \frac{\text{gap}_1(X_0)}{2} > 0. && \text{(Mirsky's inequality (24))} \end{aligned}$$

2049 Thus, \hat{X} has distinct singular values.

2050 Next, for $X_0, \hat{X} \in V_n$, the result [72, Thm. 4] shows that for each $i \in [k]$

$$\begin{aligned}
& \min_{s \in \{\pm 1\}} \|v_i(X_0) - s \cdot v_i(\hat{X})\|_2 \\
& \leq \frac{\sqrt{8}(2\tilde{\sigma}_1(X_0) + \tilde{\sigma}_1(X_0 - \hat{X}))\|X_0 - \hat{X}\|_F}{\widetilde{\text{gap}}_2(X_0)} \\
& = \frac{\sqrt{8n}(2\sigma_1(X_0) + \sigma_1(X_0 - \hat{X}))\|X_0 - \hat{X}\|_F}{n\text{gap}_2(X_0)} & (\text{since } \tilde{\sigma}_i(X) = \sigma_i(X)\sqrt{n}) \\
& = \frac{\sqrt{8}(2\sigma_1(X_0) + \sigma_1(X_0 - \hat{X}))\|X_0 - \hat{X}\|_{\bar{2}}}{\text{gap}_2(X_0)} & (\text{since } \|X\|_{\bar{2}} = \|X\|_F/\sqrt{n}) \\
& \leq \frac{\sqrt{8}(2\sigma_1(X_0) + \|X_0 - \hat{X}\|_{\bar{2}})\|X_0 - \hat{X}\|_{\bar{2}}}{\text{gap}_2(X_0)} & (\text{by (23), } \sigma_1(X)^2 \leq \|X\|_{\bar{2}}^2 = \sum_{i=1}^k \sigma_i(X)^2) \\
& \leq \frac{\sqrt{8}(2\sigma_1(X_0) + 1)\|X_0 - \hat{X}\|_{\bar{2}}}{\text{gap}_2(X_0)} & (\text{since } \|\hat{X} - X_0\|_{\bar{2}} \leq R(X_0) \leq 1) \\
& = B(X_0)\|X_0 - \hat{X}\|_{\bar{2}}.
\end{aligned}$$

2051 The final bound is independent of n and hence applies on all of V_∞ . It is also continuous in $\|\cdot\|_{\bar{2}}$ on
2052 the dense subset of V_∞ consisting of operators with distinct singular values, so taking closures we
2053 conclude that this bound applies for any $X_0, \hat{X} \in \overline{V_\infty}$ such that $\text{gap}_2(X_0) > 0$.

2054 Finally, combining the last line above with our bound on $R(X_0)$, we get

$$\min_{s \in \{\pm 1\}} \max_{1 \leq j \leq k} |[v_i(X_0)]_j - s \cdot [v_i(\hat{X})]_j| \leq \min_{s \in \{\pm 1\}} \|v_i(X_0) - s \cdot v_i(\hat{X})\|_2 \leq \frac{1}{2} \min_{1 \leq i, j \leq k} |[v_i(X_0)]_j|. \quad (29)$$

2055 Thus, the sign s achieving the above minimum is the one making *all* entries of $v_i(\hat{X})$ have the same
2056 sign as those of $v_i(X_0)$. Since the first entries of $v_i(X_0)$ and of $v_i(\hat{X})$ are positive, we conclude that
2057 $s = 1$ achieves the above minimum.

2058 (2) Combining the above results,

$$\begin{aligned}
\|V(X_0) - V(\hat{X})\|_F & \leq \sum_{i=1}^k \|v_i(X_0) - v_i(\hat{X})\|_2 \\
& = \sum_{i=1}^k \min_{s \in \{\pm 1\}} \|v_i(X_0) - s \cdot v_i(\hat{X})\|_2 \\
& \leq kB(X_0)\|X_0 - \hat{X}\|_{HS},
\end{aligned}$$

2059 as claimed.

2060 Finally, we have

$$\begin{aligned}
\|X_0 V(X_0) - \hat{X} V(\hat{X})\|_{HS} & \leq \|X_0(V(X_0) - V(\hat{X}))^\top\|_{HS} + \|(X_0 - \hat{X})V(\hat{X})\|_{HS} \\
& \leq kB(X_0)\|X_0 - \hat{X}\|_{\bar{2}}\sigma_1(X_0) + \|X_0 - \hat{X}\|_{\bar{2}}, & (\text{by (25)})
\end{aligned}$$

2061 yielding the last claim. \square

2062 **SVD-DS is more computationally efficient than DS-CI** When $k \ll n$ (for example, for us $k = 2$
2063 or 3), to evaluate SVD-DS on a given input of size $n \times k$, we need to compute its SVD at a cost of
2064 $O(n)$ and then evaluate DeepSet on the output, which takes $O(n)$ again. Moreover, during training
2065 we can compute the SVD of the dataset once in advance. In contrast, for DS-CI we need to form
2066 VV^\top at a cost of $O(n^2)$, and this needs to be differentiated-through during training. Thus, SVD-DS
2067 is much faster to train and to deploy compared to DS-CI.

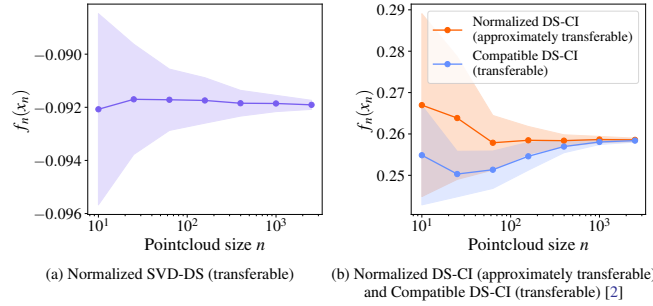


Figure 4: Transferability of invariant models on point clouds with respect to $(\mathbb{V}_{\text{dup}}^P, \|\cdot\|_{\overline{P}})$. The plot shows outputs of untrained, randomly initialized models for a sequence of point clouds $X_n \in \mathbb{R}^{n \times k}$, where each point is sampled i.i.d. from $\mathcal{N}(0, I_k)$. The error bars extend from the mean to \pm one standard deviation over 100 random samples. (a)(b): The transferable models generate convergent outputs. (b): Showing the asymptotic equivalence between the normalized DS-CI and compatible DS-CI as proved in Lemma H.2

2068 H.4 Transferability plots

2069 The numerical experiments illustrating the transferability of SVD-DS and normalized and compatible
 2070 DS-CI is shown in Figure 4. There we also illustrate the asymptotic equivalence of the latter two
 2071 models.

2072 I Size generalization experiments: details from Section 6

2073 In this section, we provide details of our size generalization experiments. In all cases, the training
 2074 dataset consists of inputs with a fixed, small dimension n_{train} . For evaluation, we use a series of test
 2075 datasets where the input dimension n_{test} is progressively larger than n_{train} .

2076 All experiments are implemented using the PyTorch framework and are trained on a single NVIDIA
 2077 A5000 GPU. Specific training and model configurations are provided in the descriptions of the
 2078 individual experiments.

2079 I.1 Size generalization on sets

2080 We consider two any-dimensional learning tasks on sets, where the target functions have different
 2081 properties, so that different models are expected to perform better.

2082 **Model and training details.** In both experiments, we compare the size generalization of three
 2083 models: DeepSet, normalized DeepSet, and PointNet as analyzed in Appendix F.2. Both σ and ρ
 2084 parts of the model are parametrized by three fully connected layers with hidden dimension 50 and
 2085 ReLU activation. Training was performed by minimizing the MSE loss using the AdamW optimizer
 2086 with an initial learning rate of 0.001 and a weight decay of 0.1. The learning rate was automatically
 2087 halved if the validation loss did not improve for 50 consecutive epochs. Each model was trained for
 2088 1000 epochs, with each run taking less than 3 minutes to complete.

2089 I.1.1 Experiment 1: Population statistics

2090 We adopt the experimental setup from [73, Section 4.1.1], which comprises four distinct tasks on
 2091 population statistics. In all four tasks, the datasets consist of sets where each set contains i.i.d.
 2092 samples from a distribution μ , where μ itself is sampled from a parameterized distribution family.
 2093 The objective is to learn either the entropy or the mutual information of the distribution μ .

2094 While the original experiment in [73] focused on training and testing with set sizes $n_{\text{train}} = n_{\text{test}} =$
 2095 $[300, 500]$, we instead evaluate size generalization. During the training stage, the dataset consists of
 2096 $N = 2048$ sets, each of size $n_{\text{train}} = 500$. This dataset is randomly split into training, validation,
 2097 and test data in the ratio 50%, 25%, 25%. During the evaluation stage, the trained model is tested

on a sequence of datasets with set sizes $n_{\text{test}} \in \{500, 1000, 1500, \dots, 4500\}$, each consisting of $N = 512$ sets.

The descriptions of the four tasks, as originally presented in [73], are provided below:

- **Rotation:** Generate N datasets of size M from $\mathcal{N}(0, R(\alpha)\Sigma R(\alpha)^T)$ for random Σ and $\alpha \in [0, \pi]$. Learn marginal entropy of the first dimension.
- **Correlation:** For $d = 16$, generate sets from $\mathcal{N}(0, [\Sigma, \alpha\Sigma; \alpha\Sigma, \Sigma])$ for random Σ and $\alpha \in (-1, 1)$. Learn mutual information between first d and last d dimensions.
- **Rank 1:** Generate sets from $\mathcal{N}(0, I + \lambda vv^T)$ for random $v \in \mathbb{R}^{32}$ and $\lambda \in (0, 1)$. Learn mutual information.
- **Random:** Generate sets from $\mathcal{N}(0, \Sigma)$ for random $d \times d$ covariance matrices Σ , $d = 32$. Learn mutual information.

In all tasks, the target functions are scalar functions on the underlying probability measure μ , and are continuous with respect to the Wasserstein p -distance. Based on Appendix F.2.1, normalized DeepSet is well-aligned with the task at the level of continuity. PointNet aligns at the compatibility level, and DeepSet lacks alignment. The results are summarized in Figure 5, showing that stronger task-model alignment improves both in-distribution and size-generalization performance.

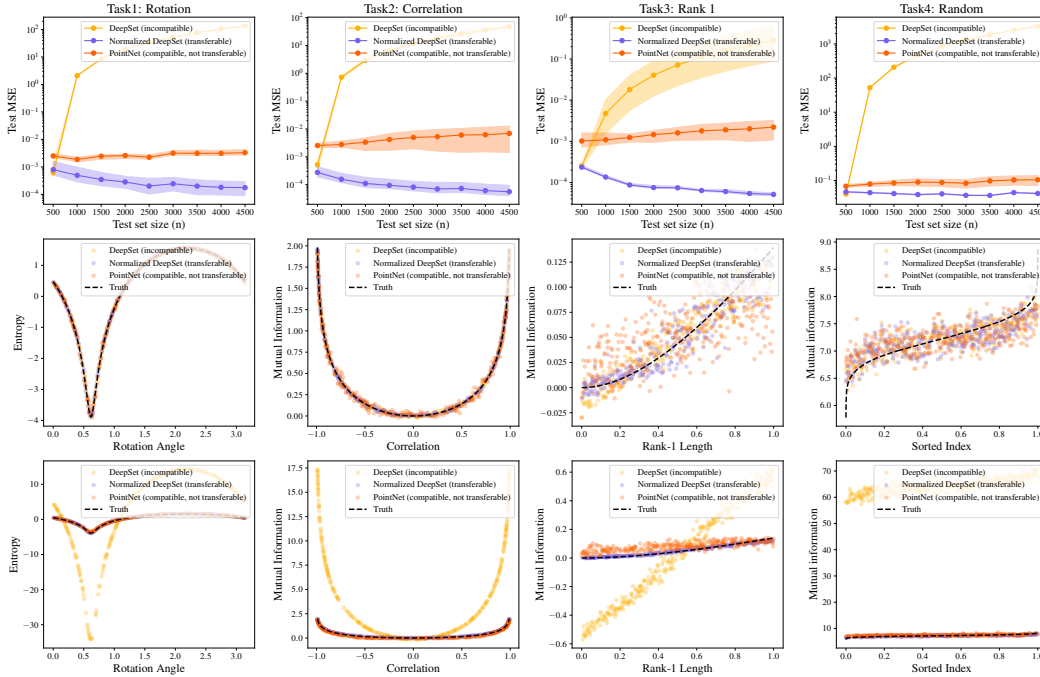


Figure 5: Size generalization on population statistic experiment. All models are trained on set size $n_{\text{train}} = 500$ and tested on set sizes $n_{\text{test}} \in \{500, 1000, \dots, 4500\}$. **Top:** MSE (log scale) vs. test set size. The solid line denotes the mean, and the error bars extend from the minimum to the maximum test MSE over 10 randomly initialized trainings. In terms of size generalization performance: Normalized DeepSet performs better than PointNet, which in turn outperforms DeepSet. **Middle:** Test-set predictions of the three models vs. ground truth for $n_{\text{test}} = n_{\text{train}} = 500$. All models have similar performances. **Bottom:** Test-set predictions vs. ground truth for $n_{\text{test}} = 4500 \gg n_{\text{train}} = 500$. DeepSet has blown-up outputs due to scaling; In Task 3, Normalized DeepSet clearly outperforms PointNet.

I.1.2 Experiment 2: Maximal distance from the origin

Recall that we consider the following data and task: each dataset consists of sets where each set contains n two-dimensional points sampled as follows. First, a center is sampled from $\mathcal{N}(0, I_2)$ and a radius is sampled from $\text{Unif}([0, 1])$, which together define a circle. The set then consists of n points sampled uniformly along the circumference. The goal is to learn the maximum Euclidean

2119 norm among the points in each set. Equivalently, this is the Hausdorff distance $d_H(\{0\}, X)$. Hence
 2120 the target function depends only on the support of the point cloud and is continuous with respect to
 2121 the Hausdorff distance.

2122 We again evaluate size generalization. During training stage, the dataset consists of $N = 5000$ sets,
 2123 each of size $n_{\text{train}} = 20$. This dataset is randomly split into training, validation and test data in
 2124 the ratio 80%, 10%, 10%. During the evaluation stage, the trained model is tested on a sequence of
 2125 datasets with set sizes $n_{\text{test}} \in \{20, 40, \dots, 200\}$, each consisting of $N = 1000$ sets.

2126 For this task, PointNet aligns at the level of continuity, normalized DeepSet at the level of compatibil-
 2127 ity, and DeepSet is not aligned. The result is summarized in Figure 6 showing model performance
 2128 again improves with better task-model alignment.

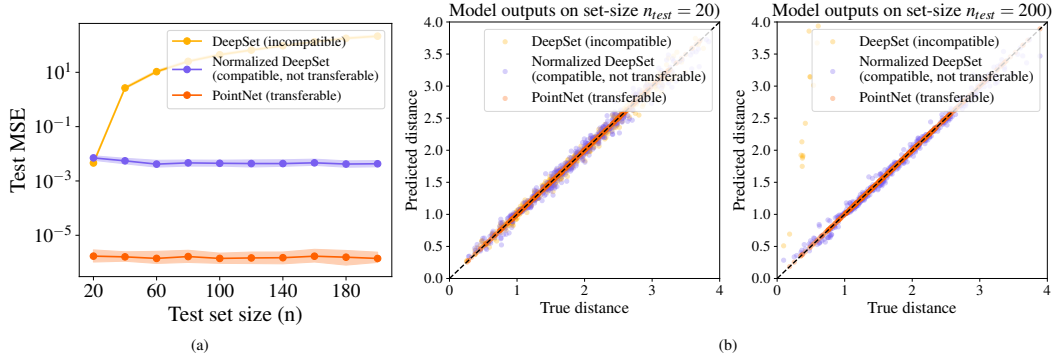


Figure 6: Size generalization on max-distance-from-origin task. All models are trained on set size $n_{\text{train}} = 20$ and tested on set sizes $n_{\text{test}} \in \{20, 40, \dots, 200\}$. **(a):** Test MSE (log scale) vs. test set size. The solid line denotes the mean, and the error bars extend from the minimum to the maximum test MSE over 10 randomly initialized trainings. In terms of size generalization performance: PointNet performs better than normalized DeepSet, which in turn outperforms DeepSet. **(b):** Test-set predictions vs. ground truth for set size $n_{\text{test}} = n_{\text{train}} = 20$, and $n_{\text{test}} = 200 \gg n_{\text{train}} = 20$. PointNet is very accurate in both cases. DeepSet has blown-up outputs due to scaling in the latter case.

2129 I.2 Size generalization on graphs

2130 The dataset consists of N attributed graphs (A, x) generated according to the following two proce-
 2131 dures (the first is described and reported in the main paper):

- 2132 1. Each graph is a fully connected weighted graph whose adjacency matrix has entries $A_{ij} =$
 2133 $A_{ji} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([0, 1])$ for $i \leq j$, and node features $x_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([0, 1])$.
- 2134 2. First, sample the number of clusters K uniformly from $\{10, \dots, 20\}$, and construct a $K \times K$
 2135 symmetric probability matrix P with entries sampled uniformly from $[0, 1]$. The resulting
 2136 stochastic block model (SBM) is used to generate an undirected, simple graph with edges
 2137 sampled as $A_{ij} = A_{ji} \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(P_{z_i, z_j})$ for $i \leq j$, where $z_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \dots, K\})$ are cluster
 2138 assignments. Node features are given by $x_i = \gamma_{z_i}$, where $\gamma \in \mathbb{R}^K$ has entries sampled
 2139 uniformly from $[0, 1]$.

2140 The task is to learn the rooted, signal-weighted homomorphism density of degree three, defined as

$$\mathbb{R}_{\text{sym}}^{n \times n} \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad (A, x) \mapsto y,$$

2141 where

$$y_i = \frac{1}{n^2} \sum_{j, k \in [n]} A_{ij} A_{jk} A_{ki} x_i x_j x_k.$$

2142 Note that when $x = \mathbb{1}$, y_i corresponds to a normalized count of triangles centered at node i . Thus,
 2143 this can be interpreted as a signal-weighted triangle density. This formulation is related to the signal-
 2144 weighted homomorphism density studied in [27], which generalizes the notion of homomorphism
 2145 density extensively studied in graphon theory [44].

We conduct experiments to evaluate the size generalization performance of the following models: the GNN from [56], the normalized 2-IGN [45], and our proposed GGNN and continuous GGNN. ReLU is used as the entry-wise activation function in all models. We choose the number of layers and hidden dimensions such that each model has approximately 2k parameters to ensure a fair comparison.

During training, we use a dataset of $N = 5000$ graphs, each with $n_{\text{train}} = 50$ nodes. This dataset is randomly split into training, validation, and test sets in a 60%, 20%, 20% ratio. For evaluation, we test the trained models on datasets of graph sizes $n_{\text{test}} \in \{50, 200, 500, 1000, 2000\}$, each containing $N = 1000$ graphs.

Training is performed by minimizing the MSE loss using the AdamW optimizer with an initial learning rate of 0.001 and weight decay of 0.1. Each model is trained for 500 epochs. Training a single run takes less than 3 minutes for the GNN model, and approximately 6–9 minutes for the IGN, GGNN, and continuous GGNN models, which are more computationally intensive. We note that evaluation on large graphs is particularly time- and memory-intensive for these models, taking up to several hours. Memory limitations restrict the maximum graph size we can evaluate to $n = 2000$.

The results of the size generalization experiments are summarized in Figure 7. Since the target function naturally extends to the graphon-level signal-weighted triangle density $(W, f) \mapsto g$, given by

$$g(x) = \int_{[0,1]^2} W(x, y)W(y, z)W(z, x)f(x)f(y)f(z) dy dz,$$

which is continuous with respect to the cut norm, models that are continuous under this topology—such as the GNN and continuous GGNN—are aligned with the task at the level of continuity. Our proposed continuous GGNN, which is provably transferable and likely more expressive than the GNN, achieves the best performance. Although GGNN is not transferable under the cut norm, it is transferable under a weaker topology (see Appendix G.3), enabling it to perform reasonably well. In contrast, the 2-IGN model, even after proper normalization, exhibits divergent outputs for larger graph sizes, indicating a lack of compatibility with the task.

Finally, we remark that the expressive power of various GNN architectures with respect to homomorphism densities has been extensively studied. Prior work has shown that common GNNs—including those considered in this study—are generally unable to express homomorphism densities of degree ≥ 3 [10, 27]. However, our results demonstrate that GNNs can still achieve strong performance on this task when evaluated over certain large parametric families of random graph models. This does not contradict prior theoretical findings, as our results pertain to an *average-case* evaluation, while the negative results in the literature are established in the *worst-case* setting.

I.3 Size generalization on 3D point clouds

We follow the setup of Section 7.2 in [2]. From ModelNet10, we select 80 point clouds from class 2 (chair) and 80 from class 7 (sofa), split into 40 training and 40 testing samples per class. Each dataset has 40×40 cross-class pairs. The objective is to learn the third lower bound of the Gromov-Wasserstein distance [49]. We prove in Appendix I.4 that it is continuous with respect to the Wasserstein p -distance. Unlike [2], which downsampled all point clouds to 100 points, we focus on size generalization: training is done on $n_{\text{train}} = 20$, and testing is done on $n_{\text{test}} \in \{20, 100, 200, 300, 500\}$.

We compare the size generalization of 3 models: unnormalized SVD-DS, (normalized) SVD-DS and normalized DS-CI. For each pair of inputs V, V' , we predict the GW lower bound via:

$$\widehat{\text{GW}}(V, V') = a\|W(f(V) - f(V'))\|^2 + b,$$

where $f : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}^t$ is the $S_n, O(k)$ -invariant model, $t = 10$, and $W \in \mathbb{R}^{t \times t}$, $a, b \in \mathbb{R}$ are learnable. All DeepSet components (σ, ϕ) are fully connected ReLU networks. We choose the number of layers and hidden dimensions such that each model has approximately 2k parameters to ensure a fair comparison.

Training is performed by minimizing the MSE loss using the AdamW optimizer with an initial learning rate of 0.01 and weight decay of 0.1. Each model is trained for 3000 epochs. Training a single run takes less than 3 minutes.

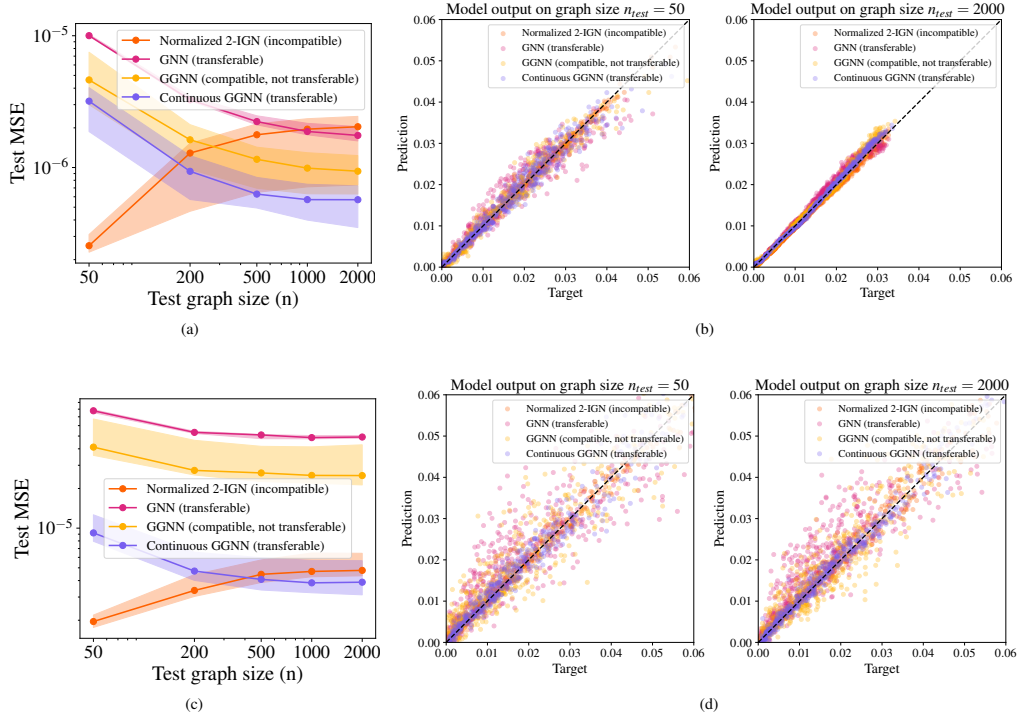


Figure 7: GNN size generalization results on weighted triangle density. (a)(b) shows the results on fully connected weighted graphs (the first data generation procedure), and (c)(d) shows the results on simple graphs sampled from SBM (the second data generation procedure). **(a)(c)**: Test MSE vs. test graph size. The solid line denotes the mean, and the error bar extend from the 20% to 80% percentile test MSE over 10 randomly initialized trainings. Continuous GGNN performs the best for both random graph models. **(b)(d)**: Test-set predictions v.s. ground truth for graph size $n_{test} = n_{train} = 50$, and $n_{test} = 2000 \gg n_{train} = 50$.

2194 The experiment results are summarized in Figure 8. Normalized DS-CI and normalized SVD-DS, both
 2195 aligned with the target at the continuity level, achieve good performance. While normalized SVD-DS
 2196 underperforms compared to normalized DS-CI, it offers superior time and memory efficiency.

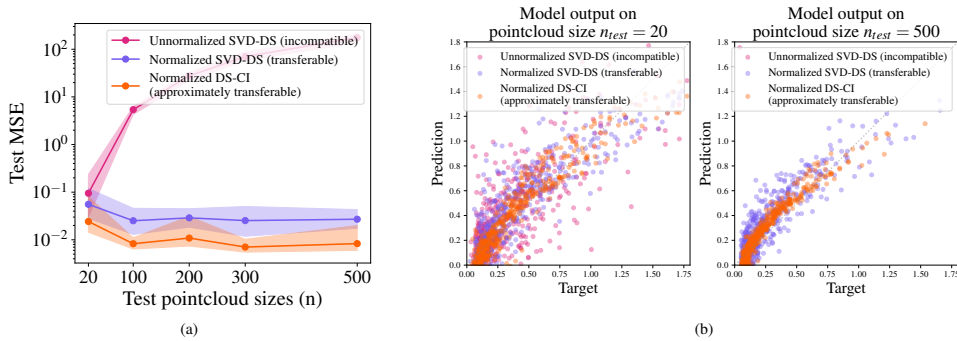


Figure 8: Size generalization results for point clouds. **(a)**: Test-set MSE (log scale) vs. point cloud size n . The solid line denotes the mean and the error bars extend from the min to max test MSE over 10 trials, each taking the best of 5 random initializations. **(b)**: Test-set predictions vs. ground truth for graph size $n_{test} = n_{train} = 20$, and $n_{test} = 500 \gg n_{train} = 20$.

2197 I.4 Continuity of Gromov-Wasserstein distance and its third lower bound

2198 The following is based on [49], though these continuity results are not stated there. Let $\mu \in$
 2199 $\mathcal{P}_p(\mathbb{R}^k)$ be a probability measure, and associate with it the “metric measure space” $(\mathcal{X}, d_{\mathcal{X}}, \mu)$
 2200 where $\mathcal{X} = \text{supp}(\mu)$ and $d_{\mathcal{X}}(x, y) = \|x - y\|_p$. Given two such measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^k)$, the
 2201 Gromov-Wasserstein distance between their associated metric spaces is defined by

$$\mathfrak{D}_p(\mu, \nu) = \inf_{\pi \in \mathcal{M}(\mu, \nu)} \|\Gamma_{X,Y}\|_{L^p(\pi \otimes \pi)} \quad (30)$$

$$= \inf_{\pi \in \mathcal{M}(\mu, \nu)} \left(\int \left| \|x - x'\|_p - \|y - y'\|_p \right|^p d\pi(x, y) d\pi(x', y') \right)^{1/p}, \quad (31)$$

2202 where $\Gamma_{X,Y}(x, y, x', y') = \left| \|x - x'\|_p - \|y - y'\|_p \right|$ and $\mathcal{M}(\mu, \nu)$ is the set of couplings between μ
 2203 and ν . The G-W distance admits the following lower bound [49, Def. 6.3].

$$\begin{aligned} \mathfrak{D}_p &\geq \text{TLB}_p(\mu, \nu) = \inf_{\pi \in \mathcal{M}(\mu, \nu)} \|\Omega_{\mu, \nu}\|_{L^p(\pi)}, \\ &\quad \text{where } \Omega_{\mu, \nu}(x, y) = \inf_{\pi' \in \mathcal{M}(\mu, \nu)} \|\Gamma(x, y, \cdot, \cdot)\|_{L^p(\pi')}. \end{aligned} \quad (32)$$

2204 We aim to show that both \mathfrak{D}_p and TLB_p are continuous with respect to the Wasserstein- p metric on
 2205 \mathcal{P}_p . More precisely, the following Lipschitz bounds hold.

2206 **Proposition I.1.** *Let $\mu, \nu, \mu', \nu' \in \mathcal{P}_p(\mathbb{R}^k)$. Then*

$$|\mathfrak{D}_p(\mu, \nu) - \mathfrak{D}_p(\mu', \nu')| \leq W_p(\mu, \mu') + W_p(\nu, \nu'). \quad (33)$$

2207 *Proof.* By the triangle inequality for \mathfrak{D}_p , which holds by [49, Thm. 5.1(a)], we have

$$\mathfrak{D}_p(\mu, \nu) \leq \mathfrak{D}_p(\mu, \mu') + \mathfrak{D}_p(\mu', \nu') + \mathfrak{D}_p(\nu, \nu').$$

2208 By [49, Thm. 5.1(d)], we further have $\mathfrak{D}_p(\mu, \mu') \leq W_p(\mu, \mu')$ and similarly for $\mathfrak{D}_p(\nu, \nu')$. Combin-
 2209 ing these inequalities and interchanging the roles of (μ, ν) and (μ', ν') , we get the claim. \square

2210 The above proof only used the triangle inequality and the bound $\mathfrak{D}_p \leq W_p$ for the G-W distance.
 2211 Since $\text{TLB}_p \leq \mathfrak{D}_p \leq W_p$, the latter property also holds for TLB. Thus, it suffices to prove TLB_p
 2212 satisfies the triangle inequality, hence is similarly Lipschitz in W_p .

2213 **Lemma I.2** (Triangle inequality for $\Omega_{\mu, \nu}$). *Let $\mu, \nu, \xi \in \mathcal{P}_p(\mathbb{R}^k)$. We have $\Omega_{\mu, \nu}(x, y) \leq \Omega_{\mu, \xi}(x, z) +$
 2214 $\Omega_{\xi, \nu}(z, y)$ for all x, y, z in the relevant supports. Furthermore, we have $\text{TLB}_p(\mu, \nu) \leq \text{TLB}_p(\mu, \xi) +$
 2215 $\text{TLB}_p(\xi, \nu)$.*

2216 *Proof.* Note that the usual triangle inequality for $\|\cdot\|_p$ gives $\Gamma(x, y, x', y') \leq \Gamma(x, z, x', z') +$
 2217 $\Gamma(z, y, z', y')$ for any $x, y, z, x', y', z' \in \mathbb{R}^k$. For any couplings $\pi_1 \in \mathcal{M}(\mu, \xi)$ and $\pi_2 \in \mathcal{M}(\xi, \nu)$,
 2218 the Gluing Lemma [65, Lemma 7.6] guarantees the existence of a coupling $\pi \in \mathcal{M}(\mu, \xi, \nu)$ whose
 2219 corresponding marginals are π_1, π_2 . Let $\pi_3 \in \mathcal{M}(\mu, \nu)$ be the marginal of π on its first and third
 2220 coordinates. Then

$$\begin{aligned} \Omega_{\mu, \nu}(x, y)^p &\leq \|\Gamma(x, y, \cdot, \cdot)\|_{L^p(\pi_3)}^p = \|\Gamma(x, y, \cdot, \cdot)\|_{L^p(\pi)}^p \\ &\leq \|\Gamma(x, z, \cdot, \cdot) + \Gamma(z, y, \cdot, \cdot)\|_{L^p(\pi)}^p \leq \|\Gamma(x, z, \cdot, \cdot)\|_{L^p(\pi_1)}^p + \|\Gamma(z, y, \cdot, \cdot)\|_{L^p(\pi_2)}^p. \end{aligned}$$

2221 Since this holds for all couplings π_1, π_2 as above, we obtain the first claim.

2222 The second claim is proved analogously. For couplings π_1, π_2, π_3, π as above, we have

$$\text{TLB}_p(\mu, \nu) \leq \|\Omega_{\mu, \nu}\|_{L^p(\pi_3)} = \|\Omega_{\mu, \nu}\|_{L^p(\pi)} \leq \|\Omega_{\mu, \xi} + \Omega_{\xi, \nu}\|_{L^p(\pi)} \leq \|\Omega_{\mu, \xi}\|_{L^p(\pi_1)} + \|\Omega_{\xi, \nu}\|_{L^p(\pi_2)},$$

2223 and taking infs over π_1, π_2 completes the proof. \square

2224 **Proposition I.3.** *Let $\mu, \nu, \mu', \nu' \in \mathcal{P}_p(\mathbb{R}^k)$. Then*

$$|\text{TLB}_p(\mu, \nu) - \text{TLB}_p(\mu', \nu')| \leq W_p(\mu, \mu') + W_p(\nu, \nu'). \quad (34)$$

2225 *Proof.* The proof is now identical to that of Proposition I.1. \square

2226 The above argument generalizes to measures on different abstract metric spaces, we did not use the
 2227 fact that all measures involved were over \mathbb{R}^k .

2228 **J Limitations of this work**

2229 This work provides a theoretical framework for transferability based on consistent sequences. It
2230 applies to several machine learning models that use a fixed number of parameters to define functions
2231 on any-dimensional inputs. However, this theory does not capture all possible ways for inputs to grow
2232 in dimension. In particular, it does not capture settings where there is no limiting space containing
2233 all finite-sized inputs, and where such inputs can be compared. For example, how do we compare
2234 natural language inputs of different lengths to each other?

2235 While we believe the general framework may extend to settings such as images (with varying
2236 resolutions and sizes), partial differential equations (across different scales), and sequences (of
2237 varying lengths), we did not explore these directions. We leave these investigations for future work.

2238 Finally, as discussed briefly in the related work, we do not consider the expressive power of neural
2239 networks on the limiting space. If the target function is not expressible, mere alignment in terms of
2240 compatibility and continuity—as discussed in Section 4—is insufficient to ensure good performance.
2241 Studying universal approximation theory on the limiting space is an important and promising direction
2242 for future research.

2243