Monotone and Separable Set Functions: Characterizations and Neural Models

Soutrik Sarangi* IIT Bombay
Yonatan Sverdlov* Technion
Nadav Dym Technion
Abir De IIT Bombay

Abstract

Motivated by applications for set containment problems, we consider the following fundamental problem: can we design set-to-vector functions so that the natural partial order on sets is preserved, namely $S \subseteq T$ if and only if $F(S) \leq F(T)$. We call functions satisfying this property *Monotone and Separating (MAS)* set functions. We establish lower and upper bounds for the vector dimension necessary to obtain MAS functions, as a function of the cardinality of the multisets and the underlying ground set. In the important case of an infinite ground set, we show that MAS functions do not exist, but provide a model called MASNET which provably enjoys a relaxed MAS property we name "weakly MAS" and is stable in the sense of Holder continuity. We also show that MAS functions can be used to construct universal models that are monotone by construction and can approximate all monotone set functions. Experimentally, we consider a variety of set containment tasks. The experiments show the benefit of using our MASNET model, in comparison with standard set models which do not incorporate set containment as an inductive bias. Our code is available in GitHub.

1 Introduction

A multiset $\{\!\{x_1, \ldots, x_n\}\!\}$ is an unordered collection of vectors, where order does not matter (like sets) and repetitions are allowed (unlike sets). In recent years, there has been increased interest in neural networks that can map multisets to vectors, with applications for physical simulations [1], processing point clouds [2], and graph neural networks [3]. Another important application of multiset-to-vector maps is set-containment search [4, 5, 6]. Here the goal is to check whether a given multiset S is (approximately) a subset of T, and this is often carried out by learning a multiset-to-vector mapping F, and then checking whether $F(S) \leq F(T)$ (element-wise), in which case, one deduces that S is a subset of T. In this paper, we look into this problem from a theoretical perspective.

Monotone and Separable functions: We begin our analysis with some definitions: we say that a function F mapping sets to vectors is *monotone* if $S \subseteq T \subseteq V$ implies that $F(S) \leq F(T)$, and we will say that F is *separable* if $F(S) \leq F(T)$ implies that $S \subseteq T$. When F is simultaneously *monotone and separable*, we call it a MAS function. When F is a MAS function, we can safely test whether $F(S) \leq F(T)$, and this will be fully equivalent to checking whether $S \subseteq T$.

Motivation for MAS functions: It is natural to ask why we're interested in functions that are MAS instead of just being monotone or just separable. For this, we begin with a couple of real-life applications based on the set-containment task: (I) In recommendation system design, one problem is to recommend item having a particular set of features, S. Here, we represent each item as a set of corresponding features T, and the problem is to find all sets T such that approximately, $S \subseteq T$. (II) Text entailment where we are given a small query sentence q nd the goal is to find the set of corpus items c from a large corpus C, where $q \implies c$ or c entails q. Here, q and c are typically represented as sets of contextual embeddings (S for q and T for c) and the entailment problem can be cast as the problem of checking if $S \subset T$.

*Soutrik and Yonatan contributed equally. Contact emails of the authors: soutriksarangi14@gmail.com, yonatans@campus.technion.ac.il, nadavdym@technion.ac.il, abir@cse.iitb.ac.in

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

In a neural network setting, such problems require us to design a function F such that the "order" between $F(S)$ and $F(T)$ can serve as a boolean test for whether S is a subset of T. For this, F needs to satisfy two conditions: (A) $F(S) \leq F(T)$ implies $S \subseteq T$ (B) $F(S) \not\leq F(T)$ implies $S \not\subseteq T$. Here, we like to emphasize that, for high accuracy, F needs to satisfy both conditions A and B as above. Monotone functions satisfy condition (B) without necessarily satisfying condition (A). As a result, if we predict $S \subseteq T$ based on $F(S) \leq F(T)$ when F is monotone but not separable, it will give large number of false positives. Similarly, if F is separable but not monotone, then F satisfies condition (A) without satisfying condition (B). Thus, using such for checking $S \subseteq T$ results in a large number of false negatives. Hence, both monotonicity and separability are necessary to develop an accurate test for set-containment

Related works: Monotone set functions have been extensively studied from theoretical perspective, *e.g.*, in the context of theory of capacities [7], game theory [8], fuzzy systems [9], combinatorial auctions [10, 11, 12] and learning theory [13, 14, 15, 16, 17, 18]. However, to the best of the knowledge, the notion of functions which are jointly monotone and separable multiset (MAS), is entirely new, albeit some related works on partial order of sets can be traced in mathematics literature, such as [19]. Our goal in this paper is to understand whether it is possible to construct MAS multiset functions, and if so, develop differentiable models that are MAS by construction and are appropriate for multiset containment tasks.

Our inquiry is also motivated by the study of multiset injective functions. This has been extensively studied in recent works [20, 3], including characterization of the vector dimension needed to enable injectivity [21, 22, 23] and the development of differentiable injective multiset models [24, 25, 26]. In this work, we aim to attain similar results for the more challenging problem of MAS multiset functions (more details in Subsection 2.1).

An additional goal of this work is stability: in most learning scenarios, we are looking for S, which is only approximately a subset of T. Accordingly, we would like to design functions F that are not only MAS but also stable, in the sense that when S is approximately a subset of T, then $F(S)$ is approximately dominated by $F(T)$.

Summary of our goal In summary, in this work, we aim to characterize (MAS) set functions that output finite-dimensional set representations and subsequently design neural networks for such functions. At a high level, we seek to address the following questions: **(1)** What are the conditions for existence of a finite dimensional MAS functions? **(2)** What are the permissible relaxations to monotonicity and separability if the existential conditions identified in (1) are not satisfied? **(3)** What are the possible neural architectures for these functions? **(4)** Are these functions stable? We expect that the design of trainable models with built-in guarantees of monotonicity, separability, and stability will significantly enhance the inductive bias for set containment applications.

1.1 Main Results

We address the goal specified in the previous subsection by providing a detailed characterization and neural architecture for monotone and separable set functions. Our main results are as follows:

Existential characterization of MAS functions Our first objective is to determine whether it is possible to obtain a MAS set function with finite output dimension m. We begin our analysis in the case where set elements are assumed to be taken from a finite ground set V, and show that in this case a MAS function exists if, and only if, $m \geq |V|$. Next, we add the assumption that the cardinality of the input multisets is bounded by k, and provide lower and upper bounds for m in this case. When the ground set is infinite, we show that MAS set functions do not exist.

Weakly MAS functions In most applications, the ground set is $V = \mathbb{R}^D$, and our analysis shows that in this case, MAS functions do not exist in general. To address this, we introduce the notion of weakly MAS functions by extending the set functions to parametric set functions $F(S, w)$, where w is a parameter. Weakly MAS functions requires that (i) $F(:, w)$ is monotone, for all w, and (ii) for any $S \not\subset T$, there exists at least one $w \in \mathcal{W}$ such that $F(:, w)$ separates S and T. We explain how, by choosing suitable activations, it is possible to construct simple deep sets [20] models which are weakly MAS, and name the resulting model MASNET.

Stability We define the notion of stability as discussed above, separability notions are Boolean and fail to capture graded distance. We would like to guarantee that $F(T)$ is 'almost' larger than $F(S)$ when S is 'almost' a subset of T. To address this, we propose a novel asymmetric set distance

and present a Holder separability condition [27] which ensures stability in terms of this asymmetric distance.

Monotone Functions Finally, we show that MAS functions can be used to provide universal models for computing set-to-vector functions. Thus, MAS functions may be a useful concept also outside of the set containment setting.

Experiments We provide a number of experiments showing that for set containment problems, using our weakly MAS model leads to stronger results in comparison with standard multiset models like DeepSets and SetTransformer, which do not incorporate MAS considerations as an inductive bias.

Summary of contributions Our main contributions in this paper are: **(1)** We introduce the novel notion of *monotone and separating* (MAS) multiset functions. **(2)** We discuss lower and upper bounds for the embedding dimension for MAS functions, and suggest a model MASNET, which is weakly MAS. **(3)** We show the stability of MASNET. **(4)** Experimentally, we prove the effectiveness of MASNET for set containment tasks.

2 Existential characterization of MAS functions

We begin by stating the notation we will use for our paper.

Notation We denote by V to be the ground set and $\mathcal{P}_{<\infty}(V)$ to be the collections of all multisets from V with finite cardinality $S = \{\!\{x_1, \cdots, x_s\}\!\}$ where $x_i \in V$. We denote the space of all multisets with at most k elements by $\mathcal{P}_{\leq k}(V)$. We use $[d]$ to denote the set $\{1, \cdots, d\}$. Given two vectors $v, u \in \mathbb{R}^d$, we say $v \leq u$ if $v[i] \leq u[i]$ for all $i \in [d]$. Also, for an element $v \in V$ and a multiset $S \in \mathcal{P}_{<\infty}(V)$, we use $c_S(v)$ to denote the number of times v occurs in S. We say $S \subseteq T$ if $c_S(v) \leq c_T(v)$ for all $v \in V$.

Proofs All proofs for the results stated in the paper are provided in Appendix 8.

In this section, our goal is to understand when an MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ or $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^m$ exists, and if it does, what is the smallest m for which such a mapping exists. We denote this minimal dimension as $m^\star(V, k)$, and $m^\star(V, \infty)$, respectively.

2.1 Prologue: Relation to injectivity

Our inquiry is related to the notion of injective multiset functions, as every MAS function F is, in particular, injective. Indeed, suppose for two sets S, T we have $F(S) = F(T)$. Then, we have both $F(S) \leq F(T)$ and $F(T) \leq F(S)$. Hence, we have both $S \subseteq T$ and $T \subseteq S$ by separability, which implies $S = T$. Thus, injectivity follows from separability.

Injective multiset functions $f : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ are attainable even for $m = 1$, providing that V is finite or even infinite [20, 3, 21, 24]. As we will see, attaining the strong condition of MAS multiset functions requires high dimension, and in some cases (infinite V), it does not even exist. This can be seen even in the following very simple example:

Example 1. *Assume the ground set consists of only two elements $V = \{0, 1\}$, and we only look for multisets of cardinality $\leq k = 1$. In this case, there are only three multisets in the space $\mathcal{P}_{\leq 1}(V)$: the empty set, $S = \{0\}$, and $T = \{1\}$. As discussed previously, there exists a multiset function $F : \mathcal{P}_{\leq 1}(V) \to \mathbb{R}$ which is injective. However, no such function can be separated. This is because neither S nor T is a subset of the other. However, because real numbers are totally ordered, we have either $F(S) \leq F(T)$ or $F(T) \leq F(S)$, violating the separability condition.*

We note that the monotonicity can be satisfied bu using functions of the form $F(S) = \sum_{x \in S} f(x)$ for non-negative f. Thus, achieving separation is harder than achieving monotonicity.

The key reason for the non-existence of a scalar MAS set function is that multisets (or sets) are not totally ordered, but scalars are. However, this issue does not arise for multidimensional set functions, where vectors, similar to sets, are partially ordered. So, when mapping to vectors, when and how can MAS functions be constructed? We will now discuss this.

2.2 Existence of MAS functions: finite ground set, unbounded cardinality

We study MAS functions $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^m$, where input multisets may have an arbitrarily large finite cardinality, and the ground set is finite $|V| = n$. In this case, we can show that the smallest possible dimension $m^\star(V, k)$ of a MAS function is exactly n:

Theorem 1 (Dimension of MAS Function)**.** *For a finite ground set V of size n, there exists a MAS function $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^n$. In addition, any MAS function must have a dimension of at least n. In other words, $m^\star(V, \infty) = n$.*

Proof Sketch The construction of a monotone embedding with $n = |V|$ simply uses one-hot encoding. Namely, we identify V with $[n]$. For every $S \in \mathcal{P}_{<\infty}(V)$, we define

$$F(S) = \sum_{s \in S} e_s \in \mathbb{R}^n \tag{1}$$

where $e_s \in \mathbb{R}^n$ is the vector with $e_s[s] = 1$ and $e_s[j] = 0$ for all $j \neq s$. It's clear F satisfies all conditions. For the second part, assume we have an embedding of dimension m, for every output dimension $i \in [m]$, there is a "maximal singleton element" $v_i^* \in V$ such that $F(\{v_i^*\})[i] \geq F(\{v\})[i], \forall v \in V$. The value of the set function F applied to $T = \{v_1^*, \ldots, v_m^*\}$ will dominate any singleton, due to the monotonicity of F. However, if $m < |V|$, then we can select a $u \in V \setminus T$ and then $F(\{u\}) \leq F(T)$, which contradicts separability.

2.3 Existence of MAS functions: finite ground set, finite cardinality

We now consider the case where the ground set V is finite, as before, but now the multisets have bounded cardinality $k < |V|$. Indeed, in many practical applications, input multiset cardinality is much smaller than $|V|$, even when $|V|$ is large. In this setting, we show that we can get a lower embedding dimension m than in Theorem 1. In fact, we show the minimal output dimension m of a MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ can scale logarithmically with the size of the ground set $|V|$. However, this comes at the price of an exponential dependence on k:

Theorem 2 (Upper Bound on $m^\star(V, k)$)**.** *Let $k < n$ be natural numbers, and let V be a ground set with $|V| = n$. Then there exists a MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ with embedding dimension $m = (k + 2)^{k+2} \log(n)$.*

Proof sketch We construct the MAS function by taking the one-hot embedding $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^n$ from (1), and then applying m random projections defined by vectors in $\mathbb{R}^n$ with non-negative entries. The non-negativity ensures monotonicity, and we prove that for the stated value of m, the probability of achieving a MAS function using this procedure is strictly positive.

The theorem shows that $m^\star(V, k) \leq (k + 2)^{k+2} \log(n)$. We now give two lower bounds on the embedding dimension: we show that m must depend at least linearly on k, and at least double logarithmically on n:

Theorem 3 (Lower bounds on $m^\star(V, k)$)**.** *Let $k \geq 2$ and n be natural numbers, and let V be a ground set with $|V| = n$. Then the smallest possible dimension $m^\star(V, k)$ of a MAS function satisfied $m^\star(V, k) \geq \log_2(\log_3 n)$. Moreover, if $k \leq \frac{n-1}{2}$ then $m^\star(V, k) \geq 2k$*

Proof Sketch: To obtain the first lower bound, we consider the sequence of singleton-set embeddings $\mathcal{M} := (F(\{1\}), \ldots, F(\{n\})$. By the Erdös-Szekers theorem [28], every scalar sequence with n elements has a monotone subsequence of length $\sim n^{\frac{1}{2}}$. Applying this recursively to vectors of length m gets us a monotone subsequence in all m coordinates of the vector, of length $\sim n^{\frac{1}{2^m}}$. If this subsequence is of length ≥ 3 then we will have three distinct elements $v_1, v_2, v_3 \in V$ such that $F(\{v_1\})[i] \leq F(\{v_2\})[i] \leq F(\{v_3\})[i]$ or $F(\{v_3\})[i] \leq F(\{v_2\})[i] \leq F(\{v_1\})[i], \forall i \in [m]$ Monotonicity implies that $F(\{v_2\})[i]$ is dominated by $F(\{v_1, v_3\})[i]$, for all $i \in [m]$, thus contradicting separability. It follows that MAS existence can only happen when $n^{\frac{1}{2^m}} \leq 2$, which leads to the double logarithmic lower bound.

For the lower bound in terms of k, the argument is similar to the proof of Theorem 1, where we select a "maximal singleton element" v_i^* for every dimension $i \in [m]$. Since $m < |V| - 1$, we can find elements $\{u_1, u_2\}$ disjoint from the collection $\bigcup_{i=1}^m \{v_i^*\}$, and WLOG $F(\{u_1\})[i] \geq F(\{u_2\})[i]$ for at least half of the indices in $[m]$. We then construct $S = \{u_2\}$ and T to be union of u_1 and all the v_i^* for all dimensions i, where $F(\{u_1\})[i] \leq F(\{u_2\})[i]$. Then monotonicity implies $F(S) \leq F(T)$, but $S \not\subseteq T$. This implies $k \leq |T| - 1 \leq m/2$, as otherwise separability of F in $\mathcal{P}_{\leq k}(V)$ is violated.

2.4 Non-existence of MAS functions: for infinite ground set

For most practical applications, we deal with infinite (often uncountable) ground sets, for example $V = \mathbb{R}^d$. Thus, we would like to analyze whether MAS functions exists when $|V| = \infty$. The answer to this question is negative:

Corollary 4. *Given a ground state V with $|V| = \infty$, and $k \geq 2$, there does not exist a MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ for any finite $m \in \mathbb{N}$.*

This corollary is a simple consequence of the result in Theorem 3 that the embedding dimension cannot be larger than $\log \log |V|$ when V is finite. We note that this result makes the minimal assumption $k \geq 2$. In Appendix 8 we show this assumption is necessary, and that in the degenerate case $k = 1$ there is a MAS function $F : \mathcal{P}_{\leq 1}(V) \to \mathbb{R}^2$ for $V = [-1, 1]$.

Summary In this section, we provided lower and upper bounds for the smallest value $m^\star(V, k)$ for which an MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ exists. Our results are summarized in Table 1. We note that some of these results require weak assumptions, which are stated in the theorems but not in the table.

Ground set size :	$\lvert V\rvert = n < \infty$		$\lvert V\rvert = n = \infty$	
Input multiset size : $k < \infty$		$k = \infty$	$k = 1$	$k \geq 2$
$m^\star(V,k) \leq \min\{n, (k+2)^{k+2}\log(n)\}$, $m^\star(V,k) \geq \max\{2k, \log_2 \log_3 n\}$		$m^\star(V,k) = n$	$m^\star(V,k) = 2$	Not possible

Table 1: *This table summarizes the lower and upper bounds for the smallest value $m^\star(V, k)$ for which an MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ exists, where k is maximal set cardinality and n denotes the cardinality of the ground set V.*

3 Relaxations of MAS functions

Our definitions so far did not address differential parameterized set functions $F(S, w)$, which are the natural object of interest for set learning applications. One natural way to address this could be to require that there exists w such that $F(\bullet, w)$ is a MAS function; however, since our theoretical results show that MAS functions do not exist when $|V| = \infty$, this requirement is too restrictive. Instead, we retain monotonicity as a hard constraint for every parameter $w \in \mathcal{W}$, while treating separability as an existential condition over the parameter space.

3.1 The notion of weakly MAS functions

Formally, given a ground set V, and a probability space $(\mathcal{W}, \mathcal{B}, \mu)$, we consider parametric set functions $F : \mathcal{P}_{<\infty}(V) \times \mathcal{W} \to \mathbb{R}^m$, where $F(S, \bullet)$ is measurable for all S. We then define weakly MAS functions as:

Definition 5 (weakly MAS function)**.** *The set function $F : \mathcal{P}_{<\infty}(V) \times \mathcal{W} \to \mathbb{R}$ is a weakly MAS function if the following two conditions are satisfied:*
(1) Pointwise monotonicity: *For any $w \in W$ and $S \subseteq T$, we have that $F(S, w) \leq F(T, w)$.*
(2) Weak separability: *If $S \not\subseteq T$, then there exists $w \in \mathcal{W}$ such that $F(S; w) > F(T; w)$.*

Which condition, monotonicity or separability, should be relaxed? In our definition, we relax the separability requirement so that different multiset pairs can be separated by different parameters, while we maintain strict requirements of pointwise monotonicity. This is because, as discussed in Example 1, constructing monotone set functions is straightforward, even in one dimension. In contrast, separability is a much stronger and more difficult condition to satisfy. This is a key reason why MAS functions do not exist when $|V| = \infty$. Indeed, as shown in Appendix 8, there can be no continuous separable set function even without the monotonicity assumption when the ground set is $V = \mathbb{R}^D$. Therefore, separability is the natural condition for relaxation. Moreover, monotonicity remains a useful and often necessary constraint in many applications.

Are scalar deep sets weakly MAS? Equipped with the new notion of weakly MAS functions, we ask the question of how to construct such a class of functions. As a first step to achieve this goal, we look into a simple instance of DeepSets [20]. DeepSets defines a set function by applying an 'inner' MLP M_1 to each set element, summing over the result, and then applying an 'outer' MLP M_2. We consider the case where M_1 is a shallow MLP $\mathrm{M}_1(x) = \sigma(Ax + b)$ which gives us models of the form

$$F(S; (A, b)) = \mathrm{M}_2\big(\textstyle\sum_{x \in S} \sigma(Ax + b)\big) \tag{2}$$

To ensure monotonicity for every parameter choice A, b, we will require that M_2 is a monotone vector-to-vector function, and the activation σ is non-negative. While these choices automatically ensure monotonicity, they may not lead to weak MAS functions as for many activations there will be no separation:

Proposition 6. *If the deep set model in* (2) *is implemented with a vector-to-vector monotonously increasing function M_2 and a non-negative activation σ, then $F(\bullet; (A, b))$ is monotone for every*

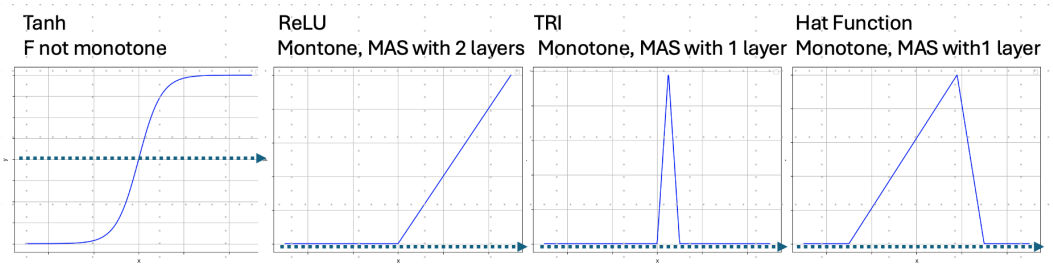

Figure 2: *Using a multiset model as in* (2) *with activations which are not always non-negative, like* $\sigma = \mathrm{Tanh}$ *will not be monotone. ReLU will be monotone, but to be weakly MAS, two layers are required. TRI and more general hat functions are weakly MAS even with a single layer.*

(A, b). *If in addition* σ *is monotone (increasing or decreasing), then there exists* $S \not\subseteq T$ *with* $F(S; (A, b)) \leq F(T; (A, b))$ *for all* A, b.

Proof idea The monotonicity is rather straightforward. To prove lack of separation when σ is monotone, choose $x \neq y$, $z = \frac{1}{2}(x + y)$, $S = \{z\}$, $T = \{x, y\}$, and then $Az + b$ is the average of $Ax + b$ and $Ay + b$, which can be used to show that $F(T; (A, b)) \geq F(S; (A, b))$ for all A, b.

Are Set Transformers weakly MAS? Now we ask the quedtion if Set Transformers [29] are weakly MAS. We notice in the following that Set Transformers are not even monotone in the following result. For this, we consider Set Transformer with sum-pooling, namely $F(S; W_Q, W_K, W_V) = \mathrm{SumPool}(\mathrm{Attn}(S))$. Explicity, this is defined for a give multiset $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ by first defining query, key, values per-point: $\mathbf{q}_i = W_Q\mathbf{x}_i, \mathbf{k}_i = W_K\mathbf{x}_i, \mathbf{v}_i = W_V\mathbf{x}_i$, and then taking a weighted average, defined by weights: $\alpha_{i,j} = \frac{e^{\mathbf{q}_i \cdot \mathbf{k}_j}}{\sum_s e^{\mathbf{q}_i \cdot \mathbf{k}_s}}$ to obtain a permutation-invariant function $F(S; W_Q, W_K, W_V) := \sum_{i,j} \alpha_{i,j}\mathbf{v}_i$

Proposition 7. *Let* $V = \mathbb{R}^d$ *and let* $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^D$ *be a Set Transformer defined by* $F(S; W_Q, W_K, W_V) :=$ *SumPool*$($*Attn*$(S))$. *Then* F *is not a point-wise monotone function.*

Proof Idea: We consider full rank matrices $W_Q = W_K$ and any non-zero W_V. We choose $\mathbf{x}_1 \in \mathbb{R}^d$ such that $\mathbf{v}_1 := W_V\mathbf{x}_1 \neq 0$, and consider the sets $S = \{\mathbf{x}_1\}, T = \{\mathbf{x}_1, \mathbf{0}_d\}$. Then, $F(S; W_Q, W_K, W_V) = \mathbf{v}_1$ and $F(T; W_Q, W_K, W_V) = (\alpha_{1,1} + \alpha_{1,2})\mathbf{v}_1$. Note that, $\alpha_{1,1} + \alpha_{1,2} > 1$ and thus, for any negative $\mathbf{v}_1[j]$, we have $F(T; W_Q, W_K, W_V)[j] < F(S; W_Q, W_K, W_V)[j]$ and thus, monotonicity is violated.

Thus, unlike Deep Sets, it's not obvious how to make Set Transformers monotone, let alone weakly separable. Thus, in the following sections, we shall focus on obtaining weakly-MAS functions from DeepSet-like models only.

3.2 The Hat activation class

Note that most commonly used non-negative activation functions, such as ReLU or sigmoid, are monotonically non-decreasing and therefore, from Proposition 6, using them as σ above will fail weak separability. Here, we consider a novel class of activation functions we call 'Hat Activations', which will all make F in Eq. (2) weakly MAS.

Definition 8 (The Hat activation function)**.** *We call a function* $\sigma : \mathbb{R} \to \mathbb{R}$ *a hat activation if it is (a) non-negative (b) compactly supported (c) not identically zero and (d) continuous.*

Examples of hat functions are the third and fourth functions in Figure 2. When σ is a hat function, Equation (2) is weakly MAS, even when the output of F is taken to be scalar. For this result, we will need to require that M_2 is strictly monotone. This can be handled by simply setting M_2 to be the identity.

Proposition 9. *If* M_2 *is strictly monotone increasing, and* σ *is a hat activation function, then* $\mathrm{M}_2(\sum_{x \in S} \sigma(a^\top x + b))$ *is weakly MAS function.*

Proof idea If S is not a subset of T, there is an element s whose multiplicity in S is larger than its multiplicity in T. We can then choose a, b so that s is in the support of $\sigma(ax + b)$ but the elements of T are not.

3.3 Slightly larger ReLU networks

So far, we have considered simple models as in (2) which apply a single linear layer before applying activation and summation. In this case, we saw that ReLU and other activations do not create weakly

MAS functions, and defined the notion of hat activations, which are weakly MAS functions. But what happens if we use ReLU activations but allow deeper networks?

We can use our previous analysis to show that very small ReLU networks with two layers can already be MAS functions. To see this, we consider a candidate from the hat function class, namely the TRI function from Figure 2, and observe that: $\mathrm{TRI}(x) = \mathrm{ReLU}\left(\mathrm{ReLU}(2x) + \mathrm{ReLU}(2x-2) - \mathrm{ReLU}(4x-2)\right)$ We use this to deduce the following result:

Proposition 10. *Given $V \subseteq \mathbb{R}^d$, we consider affine transformations $\mathcal{A}_2 : \mathbb{R}^d \to \mathbb{R}^3, \mathcal{A}_1 : \mathbb{R}^3 \to \mathbb{R}$. Thus, here $\mathcal{A}_2(t) := A_2 t + b_2$ and $\mathcal{A}_1(z) = a_1^\top z + b_1$, where $A_2 \in \mathbb{R}^{3\times d}, b_2 \in \mathbb{R}^3, a_1 \in \mathbb{R}^3, b_1 \in \mathbb{R}$ are the respective parameters. Then, the set functions of the form*

$$F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}, F(S; \mathcal{A}_1, \mathcal{A}_2) = \mathrm{M}_2\left(\textstyle\sum_{x\in S} \mathrm{ReLU} \circ \mathcal{A}_1 \circ \mathrm{ReLU} \circ \mathcal{A}_2(x)\right) \tag{3}$$

are weakly MAS functions.

Our results so far are summarized in Figure 2. If σ is not a non-negative function, then F will not be monotone. We can attain weakly MAS functions with a two-layer ReLU network, or with a one-layer network with hat activations.

3.4 Asymmetric distance induced Holder separability

Why weak separability is not sufficient While weak separability is a key relaxation for weakly MAS functions (Definition 5), it has two key limitations: **(1)** Separation should ideally hold over a non-negligible subset of $\mathcal{W}$, not just a single point w; **(2)** it treats separability as a boolean condition, ignoring approximate containment. We would like to guarantee that if S is almost a subset of T, then $F(S, w)$ will be close to dominated by $F(T, w)$. This motivates us to create an *asymmetric (pseudo) metric* to quantify the size of set difference and create stronger notions of separability based on that.

Asymmetric set distance For a ground set $V \subseteq \mathbb{R}^d$ and two sets $S, T \in \mathcal{P}_{<\infty}(V)$ such that $|S| \leq |T|$. We define the asymmetric distance from S to T as the Earth Mover Distance (EMD) between S and the subset of T that is *closest* to S with the same cardinality as S.

$$\Delta(S, T) = \min_{T' \subseteq T : |S| = |T'|} \mathrm{EMD}(S, T') \tag{4}$$

$\Delta(\bullet, \bullet)$ captures graded set containment: if $S \not\subseteq T$, but is "very close" to a subset of T in EMD metric, then $\Delta(S, T)$ is small. Moreover, $\Delta(S, T) = 0 \iff S \subseteq T$. Based on Δ, we propose a newer separability notion for parametric functions.

Lower Hölder separability of set functions Davidson and Dym [27] introduced Hölder continuity through expectation over parameters. Unlike their symmetric EMD, we define Hölder separability using the asymmetric distance Δ(4).

Definition 11 (Lower Hölder separability)**.** *Let $V \subseteq \mathbb{R}^d$ be the ground set and $(\mathcal{W}, \mathcal{B}, \mu)$ be a probability space, and a constant $\lambda > 0$. A parametric set function $F(\bullet, w) : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ with $w \sim \mu(\cdot)$ is λ lower Hölder separable if there exists $c > 0$ such that:*

$$\mathbb{E}_{w\sim\mu(\cdot)} \|[F(S; w) - F(T; w)]_+\|_1 \geq c \cdot \Delta(S, T)^\lambda, \text{ for all } S, T \in \mathcal{P}_{\leq k}(V) \tag{5}$$

3.5 Hölder separable Set functions

Hölder separability using Hat activation Proposition 9 shows how to construct weakly MAS functions using Hat activations. We now show that these constructions are also lower Hölder, under additional weak assumptions on the hat functions and M_2:

Theorem 12 (F is lower Hölder)**.** *Let $V \subset \mathbb{R}^d$ be a compact set, and σ a Hat activation function which is piecewise continuously differentiable supported in some interval $[\gamma_1, \gamma_2]$, and satisfying the condition: $\lim_{t\to\gamma_1^+} \frac{\mathrm{d}\sigma}{\mathrm{d}t} > 0$. Let $\mathrm{M}_2 : \mathbb{R} \to \mathbb{R}$ be a lower Lipschitz function. Consider the function $F(S; (a, b, c)) = \mathrm{M}_2\left(\sum_{x\in S} \sigma\left(\frac{a^\top x + b}{c}\right)\right)$, where the multisets S come from $\mathcal{P}_{\leq k}(V)$, and $a \sim \mathrm{Unif}(\mathcal{S}^{d-1}), b \sim \mathrm{Unif}([-1, 1]), c \sim \mathrm{Unif}((0, 2])$. Then, $F(\bullet, (a, b, c))$ is Monotone Hölder separable with exponent $\lambda = 2$.*

Probability of successful separation Given $S, T \subseteq V$, we can consider $[F(S; (a, b, c)) - F(T; (a, b, c))]_+$ to be a real-valued non-negative random variable, where the randomness is over the parameter space $(a, b, c) \in \mathcal{W} = \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ equipped with a probability measure. The above result on lower Hölder stability gives us a lower bound on the probability of a parameter tuple (a, b, c) separating two non-subsets S, T proportional to the set distance $\Delta(S, T)$. Moreover, the

functions we have considered so far are scalar-valued set functions. By considering independent copies of the parameters across multiple dimensions, we can increase the probability of separation, since $F(S), F(T)$ are not separated iff $F(S)[i] \leq F(T)[i]$ across each embedding dimension i. We formalize the above observations as follows:

Theorem 13 (Probability bounds on separation)**.** *Let $V \subseteq \mathbb{R}^d$ and σ be as in Theorem 12, and let $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m, c \in \mathbb{R}^m$ whose m columns (respectively entries) are drawn independently from the distribution on a_j, b_j, c_j described in Theorem 12, and consider the function*

$$F(S; A, b, c) = \sum_{x \in S} \sigma\left(c^{-1} \odot (Ax + b)\right). \tag{6}$$

Then there exists $C > 0$, so that for all $S \not\subseteq T$, $\mathbb{P}\left(F(S) \leq F(T)\right) \leq \left(1 - C\Delta^2(S,T)\right)^m$.

In Equation (6) c^{-1} stands for the elementwise inverse and $\odot$ stands for elementwise multiplication.

Note that the probability of failed separation goes to zero exponentially as m increases, and also becomes smaller as $\Delta(S,T)$ increases. This supports the idea that a larger measure of parameters separates sets with larger asymmetric distance and that separability becomes easier with larger embedding dimensions (which we saw previously in the existential results)

Hölder separable set functions using ReLU As seen in Proposition 10, one can construct weakly MAS functions with a two layer neural network with ReLU activation. Our proof used that two layer ReLU networks can basically "simulate" a Hat function. Under the light of the above results on Hölder separability, one can argue that two-layer ReLU networks would also have such guarantees. Also, we show in Appendix 8 that under certain structure on the ground set V (for example, V being a hypersphere), even with *one layer* ReLU networks, functions of the form in Equation (2) will also have Lower Hölder separability guarantees.

Upper Lipschitz bounds In Appendix. 8 we complement our results on *lower* Holder stability by showing that the construction in Theorem 13 is also *upper* Lipschitz in an appropriate sense.

3.6 Monotone Universality and the role of M_2

Besides set containment-based applications, there are other interesting scenarios where one would like to construct monotone multiset-to-vector functions. The following theorem shows that on finite ground sets, the combination of a MAS function and a universal monotone vector-to-vector (such as [30]) can approximate *all* monotone multiset-to-vector functions.

Theorem 14 (Universality)**.** *Let V be a finite ground state, and let $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ be a MAS function. Then for every multiset-to-vector monotone function $f : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^s$, there exists a vector-to-vector monotone function $M : \mathbb{R}^m \to \mathbb{R}^s$ such that $F(S) = M \circ f(S)$.*

4 MASNET: Neural Modeling of MAS functions

Now, our goal is to leverage our theoretical analysis to design neural network-based multiset-to-vector models that preserve monotonicity and weak separability. Motivated by the formulation in DeepSets and our analysis so far, we wish to design neural set functions of the form:

$$\mathrm{MASNET}(S) = M_{\theta_2}\left(\sum_{x \in S} \sigma\left(M_{\theta_1}(x)\right)\right) \tag{7}$$

In our analysis, we required M_{θ_2} to be a monotone vector-to-vector function. In most of our experiments, we enforce this simply by choosing $M_2(x) = x$, but this can also be enforced by using monotone activations and non-negative parameters.

MASNET-Hat To enforce weakly MAS in Equation (7),our results from Prop. 9 and Theorem 12 suggests using a Hat activation as σ. In our experiment, we do it by not choosing σ to be a specific hat activation, but rather the parametric form:

$$\sigma_{\alpha,\beta,\gamma}(x) = \mathrm{ReLU}\left(\tfrac{x-\alpha}{\gamma \cdot \beta}\right) + \mathrm{ReLU}\left(\tfrac{x-(\alpha+\beta)}{(1-\gamma)\cdot\beta}\right) - \mathrm{ReLU}\left(\tfrac{x-(\alpha+\gamma\cdot\beta)}{\gamma\cdot(1-\gamma)\cdot\beta}\right) \tag{8}$$

For all $\alpha, \beta > 0$, and $\gamma \in (0,1)$, $\sigma_{\alpha,\beta,\gamma}$ is a hat function (Definition 8) with support $[\alpha, \alpha+\beta]$ and peak at $\alpha + m\beta$. Examples include the third and fourth functions in Figure 2. Applying this to an m-dimensional output with independent α, β, γ per dimension yields $3m$ parameters total.

MASNET with ReLU Motivated by our theoretical results from Theorem 10 of 2-Layer ReLU networks being weakly MAS and from our discussion of ReLU networks being Hölder separable (under appropriate assumptions) in Section 3.5, we propose use ReLU in MASNET, which case M_{θ_1} will have to be a network with at least 2 layers to ensure weakly MAS as in Theorem 10. This gives one more way to design MASNET and we refer to this option as MASNET-ReLU.

Other variants of MASNET Definition 8 establishes a general hat function class extending beyond the piecewise linear functions as in (8). We present MASNET-INT in Appendix 9, which achieves universal approximation of the Hat class by modeling its derivative via neural networks and approximating the integral. Additionally, using the specific hat activation TRI (Figure 2) as σ yields an alternative formulation called MASNET-TRI.

Which MASNET to choose and when? We have given multiple recipes of MASNET in the above discussion, and the question arises: which one to choose? We show via our experiments: For parameter-constrained scenarios requiring single-layer M_{θ_1} networks, MASNET-Hat provides optimal separability (both theoretically and empirically). For deeper M_{θ_1} (≥ 2 layers), both MASNET-ReLU and Hat-based variants perform similarly, with MASNET-ReLU often providing more stable training and better performance overall.

5 Experiments

We evaluate the MASNET variants (Section 4) on synthetic, text and point-cloud datasets to characterize monotonicity and separability. Specifically, we focus on (exact and approximate) set containment. Given sets S and T, with binary label $y(S,T) \in \{0,1\}$ indicating (exact or approximate) set containment, we evaluate how accurately MASNET predicts $y(S,T)$.

5.1 Set Containment

We perform experiments on a synthetically generated dataset, four text datasets, and one image dataset. In each case, we split the dataset into 5:2:2 train, test, and dev folds. We minimize the following fixed-margin hinge loss that enforces vector dominance, to train the parameters of MASNET.

$$\sum_{S,T}(1-y(S,T)) \min_{i\in[m]} [F(S)[i]-F(T)[i]+\delta]_+ + y(S,T) \max_{i\in[m]} [F(S)[i]-F(T)[i]+\delta]_+ \quad (9)$$

Analysis on synthetic datasets To generate our synthetic dataset, we first sample a target set $T \subset \mathbb{R}^d$, where each element $x \in T$ is drawn from $\mathcal{N}(0, \boldsymbol{I})$, i.i.d. Given T, we compute S with a fixed size $|S| = s$ as follows. We first obtain S with positive labels $y(S,T)=1$ by drawing a subset from T, uniformly at random, without replacement. To generate S with $y(S,T)=0$, we sample s from $\mathcal{N}(0, \boldsymbol{I})$ independently. We chose $d=4$ and the set embeddings have $m=256$ dimensions, the class-ratio of subsets to non-subsets was taken to be 1:1.

$\lvert S\rvert$	$\lvert T\rvert$	DS	ST	M-ReLU	M-Hat
1	2	0.98	0.98	**0.99**	**0.99**
1	10	0.59	0.59	**0.99**	**0.99**
10	30	0.89	0.80	**1.00**	**1.00**
10	100	0.54	0.57	0.98	**0.99**

Table 3: **Accuracy on synthetic data**

We compare two variants of MASNET— MASNET-ReLU (M-ReLU) and MASNET-Hat (M-Hat)— against Deep Sets (DS) [20] and Set Transformer (ST) [29]. Table 3 summarizes the results for varying values of $|S|$ and $|T|$. We observe that: (1) Both variants of our method outperform the baselines; (2) the baselines degrade significantly as $|T|$ increases (with fixed $|S|$), as predicting separability becomes harder as the gap between $|T|$ and $|S|$ increases; (3) The performance of M-ReLU and M-Hat is comparable. We now compare 1-layer pointwise models ending in ReLU and Hat activation respectively, followed by aggregation: we call them 1layer-ReLU and 1layerM-Hat. We see in Fig. 4 that: (1) 1layerM-Hat performs significantly than 1layer-ReLU, validating the results from Propn. 6. (2) acc. declines for both as $|T|$ increases, consistent with the intuition that larger $|T \setminus S|$ gaps make the task harder.

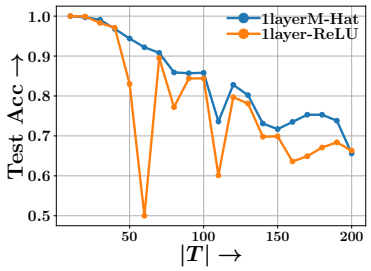

Figure 4: Acc vs $|T|$ for 1-layer MLP

Set containment on Text datasets We evaluate on MSWEB, MSNBC, and Amazon Registry datasets, which exhibit natural set containment from user behavior. Each datapoint is a wordbag; we compute BERT embeddings for all unique words to form the ground set. (S,T) pairs are sampled from these (Appendix 10), and labeled via set containment. For inexact containment, small Gaussian noise is added to S without changing $y(S,T)$. We now compare the variants of MASNET: MASNET-ReLU and MASNET-Hat against several baselines, DeepSets [20], Set Transformers [29], FlexSubNet [31], Neural SFE [32], for inexact set containment on the datasets: MSWEB, MSNBC, Amazon-Feeding, Amazon-Bedding. Gaussian noise with 0.01 std was used to generate noisy pairs. The class ratio of subsets to non-subsets was taken to be 1:1. Here,

Model	**Bedding**	**Feeding**	**MSWEB**	**MSNBC**
DeepSets	0.51	0.50	0.88	0.67
SetTransformer	0.77	0.79	0.90	0.94
FlexSubNet	0.88	0.85	0.92	0.91
Neural SFE	0.52	0.53	0.88	0.66
MASNET-ReLU	**0.98**	**0.98**	**0.99**	**0.97**
MASNET-Hat	0.94	0.93	0.97	0.95

Table 5: Acc in text datasets with 1:1 ratio

set embeddings have dim $m = 50$, and set elements have dim $d = 768$. The results in table 5 show: (1) MASNET-ReLU and MASNET-Hat have similar acc., with MASNET-ReLU being marginally better (2) Both are generally better than baselines.

Set containment on point clouds: We use ModelNet40 [33], a dataset of 3D CAD models across 40 categories. For each object in category C_1, we sample 1024 points to form T. For true subsets, we sub-sample $S \subseteq T$ and for non-subsets, S is sampled from a different category C_2. Small white noise(std=0.005) is added to S for the task of inexact containment. We use a pointcloud encoder(such as pointnet) followed by a set-to-vector embedding model to produce $m = 50$ dimensional embeddings, and each point in the dataset is a 3D coordinate with $d = 3$ dimensions. The class ratio of subsets to non-subsets was taken to be 1:1. Results are in table 6. We observe that: (1) MASNET-ReLU and MASNET-Hat are generally better than baselines, with MASNET-ReLU being slightly better. (2) The accuracy of MASNET decreases as $|S|$ decreases, with the separability becoming hard for monotone models.

$\|S\| \rightarrow$	128	256	512
DeepSets	0.52	0.51	0.50
SetTransformer	0.62	0.51	0.50
FlexSubNet	0.84	0.75	0.60
Neural SFE	0.52	0.51	0.50
MASNET-ReLU	**0.98**	**0.94**	**0.72**
MASNET-Hat	0.87	0.81	0.65

Table 6: Performance on Point cloud for different values of $|S|$ with 1:1 class ratio.

5.2 Monotone Set function approximation

In this experiment, we examine if a scalar variant of MASNET can approximate a monotone function F^*, by getting trained from set, value pairs in the form of $\{(S, F^*(S))\}$. Once trained, we measure $|MAS(S) - F^*(S)|$ for each S and average over them to compute MAE. We compare MASNET against baselines: DeepSets, SetTransformer. We trained using MSE loss, and report test MAE results in Table 7 where we see: (1) MASNET-Hat is best performing, followed by MASNET-ReLU (2) Both are better than baselines.

Model	**MAE**
DeepSets	0.01024 ± 0.00030
SetTransformer	0.01026 ± 0.00033
MASNET-ReLU	0.01023 ± 0.00107
MASNET-Hat	$\mathbf{0.00959 \pm 0.00016}$

Table 7: Monotone function approximation

6 Conclusion and Future Work

In this work, we study the design of set functions that are useful in set containment through Monotone and Separating (MAS) properties. We derive bounds on the embedding dimensions required for MAS functions and show their nonexistence in infinite domains. To address this, we introduce a relaxed model, MASNET, which satisfies a weak MAS property and is provably stable. Experiments demonstrate that MASNET outperforms standard models on set containment tasks by leveraging monotonicity as an inductive bias. It would be interesting to check if our method can be used for applications in graphs, *e.g.*, subgraph matching. Another potential direction is to consider other possible relaxations of separability.

7 Limitations

Our work provides a theoretical analysis of monotonicity and separability for set functions in the context of set containment, and introduces neural models that outperform standard set-based architectures such as DeepSets and Set Transformers on set containment tasks. Our work has the following limitations, addressing which are interesting directions for future work:

- After proving the impossibility of constructing MAS functions over infinite ground sets, we proposed parametric set functions defined with respect to a probability measure over the parameter space. We provided probabilistic guarantees: for non-subsets S, T, if their asymmetric distance $\Delta(S, T)$ is large, then a randomly sampled function from the parameter space will separate them with high probability. However, these guarantees hold only in the randomized setting—analogous to guarantees for neural networks at random initialization—and do not extend to models after training via gradient-based optimization.
- Our universality result, which shows that MAS functions composed with monotone vector-to-vector mappings can approximate all monotone set-to-vector functions, was only established for finite ground sets. Extending this result to infinite ground sets remains an open problem.
- We currently do not generalize our analysis to the subgraph isomorphism problem, which is a natural and strictly harder extension of the set-containment task.

Acknowledgment

Nadav and Yonatan were supported by ISF grant 272/23, Soutrik and Abir were supported by Amazon and Google Research Grant; as well as Bhide Family Chair Endowment Fund.

References

[1] Hang Huang, J. M. Landsberg, and Jianfeng Lu. Geometry of backflow transformation ansatz for quantum many-body fermionic wavefunctions, 2021. URL https://arxiv.org/abs/2111.10314.

[2] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[3] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

[4] Indradyumna Roy, Rishi Agarwal, Soumen Chakrabarti, Anirban Dasgupta, and Abir De. Locality sensitive hashing in fourier frequency domain for soft set containment search. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rUf0GV5CuU.

[5] Aman Singh, Naman Agarwal, Prateek Jain, and Manik Varma. Explaining monotonic ranking functions. In *Proceedings of the VLDB Endowment*, volume 13, pages 2921–2934. VLDB Endowment, 2020.

[6] Joshua Engels, Benjamin Coleman, Vihan Lakshman, and Anshumali Shrivastava. Dessert: An efficient algorithm for vector set search with vector set queries. *Advances in Neural Information Processing Systems*, 36:67972–67992, 2023.

[7] Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953. URL http://www.numdam.org/item/AIF_1953__5__131_0/. Foundational work introducing Choquet capacities and the Choquet integral.

[8] Michel Grabisch. *Set Functions, Games and Capacities in Decision Making*. Number 978-3-319-30690-2 in Theory and Decision Library C. Springer, March 2016. ISBN ARRAY(0x694541d0). doi: 10.1007/978-3-319-30690-2. URL https://ideas.repec.org/b/spr/thdlic/978-3-319-30690-2.html.

[9] Hamzeh Agahi and Yong Ouyang. Monotone set functions and their generalizations. *Fuzzy Sets and Systems*, 393:1–19, 2020. doi: 10.1016/j.fss.2019.07.009. Generalizes monotonicity to fuzzy sets and hybrid integrals.

[10] Daniel Lehmann, Liadan Ita Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, pages 18–28. ACM, 2001.

[11] Shahar Dobzinski, Noam Nisan, and Michael Schapira. Approximation algorithms for combinatorial auctions with complement-free bidders. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 610–618. ACM, 2005.

[12] Uriel Feige. Maximizing welfare when utility functions are subadditive. *SIAM Journal on Computing*, 39(1):122–142, 2009.

[13] Maria-Florina Balcan, Florin Constantin, Satoru Iwata, and Lei Wang. Learning valuation functions. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 4.1–4.24, 2012.

[14] Haifeng You, Kevin Canini, Boyan Wang, Zhe Zhang, and Maya Gupta. Deep lattice networks and partial monotonic functions. In *Advances in Neural Information Processing Systems*, pages 2981–2989, 2017.

[15] Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 9–18, 1988.

[16] Ryan O'Donnell and Rocco A. Servedio. Learning monotone functions from random examples in polynomial time. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 448–456, 2003.

[17] Nader H. Bshouty. Exact learning boolean functions via the monotone theory. *Information and Computation*, 123(1):146–153, 1995.

[18] Ming Li, Chenyi Zhang, and Qin Li. Monotonic learning in the pac framework: A new perspective. *arXiv preprint arXiv:2501.05493*, 2025.

[19] Ben Dushnik and Edwin W Miller. Partially ordered sets. *American journal of mathematics*, 63 (3):600–610, 1941.

[20] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2018. URL https://arxiv.org/abs/1703.06114.

[21] Tal Amir, Steven Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42516–42551. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/84b686f7cc7b7751e9aaac0da74f755a-Paper-Conference.pdf.

[22] Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR, 2019.

[23] Peihao Wang, Shenghao Yang, Shu Li, Zhangyang Wang, and Pan Li. Polynomial width is sufficient for set representation with high-dimensional features. *arXiv preprint arXiv:2307.04001*, 2023.

[24] Tal Amir and Nadav Dym. Fourier sliced-wasserstein embedding for multisets and measures, 2024. URL https://arxiv.org/abs/2405.16519.

[25] Radu Balan, Naveed Haghani, and Maneesh Singh. Permutation invariant representations with applications to graph deep learning. *arXiv preprint arXiv:2203.07546*, 2022.

[26] Yonatan Sverdlov, Yair Davidson, Nadav Dym, and Tal Amir. Fsw-gnn: A bi-lipschitz wl-equivalent graph neural network. *arXiv preprint arXiv:2410.09118*, 2024.

[27] Yair Davidson and Nadav Dym. On the hölder stability of multiset and graph neural networks, 2025. URL https://arxiv.org/abs/2406.06984.

[28] Paul Erdös and George Szekeres. A combinatorial problem in geometry. *Compositio mathematica*, 2:463–470, 1935.

[29] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks, 2019. URL https://arxiv.org/abs/1810.00825.

[30] Joseph Sill. Monotonic networks. *Advances in neural information processing systems*, 10, 1997.

[31] Abir De and Soumen Chakrabarti. Neural estimation of submodular functions with applications to differentiable subset selection, 2022. URL https://arxiv.org/abs/2210.11033.

[32] Nikolaos Karalias, Joshua David Robinson, Andreas Loukas, and Stefanie Jegelka. Neural set function extensions: Learning with discrete functions in high dimensions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=39XK7VJ0sKG.

[33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015. URL https://arxiv.org/abs/1406.5670.

[34] James R. Munkres. *Elementary differential topology*, volume No. 54 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, revised edition, 1966. Lectures given at Massachusetts Institute of Technology, Fall, 1961.

[35] Antoine Wehenkel and Gilles Louppe. Unconstrained monotonic neural networks, 2021. URL https://arxiv.org/abs/1908.05164.

[36] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width, 2017. URL https://arxiv.org/abs/1709.02540.

[37] Bruno Andreis, Jeffrey Willette, Juho Lee, and Sung Ju Hwang. Mini-batch consistent slot set encoder for scalable set encoding, 2021. URL https://arxiv.org/abs/2103.01615.

[38] Jeffrey Willette, Seanie Lee, Bruno Andreis, Kenji Kawaguchi, Juho Lee, and Sung Ju Hwang. Scalable set encoding with universal mini-batch consistency and unbiased full set gradient approximation, 2023. URL https://arxiv.org/abs/2208.12401.

[39] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks, 2019. URL https://arxiv.org/abs/1808.04730.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: All claims were introduced were formally restated in the main text and proved formally in the appendix.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: In the appendix, we have a section named limitations.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings,

model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In the appendix all claims and theorems were formally proved.

   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the link of the Github repository containing the code in the paper. Detailed instructions on how to run the code to reproduce the results are provided in the README file of the repository.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case

of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We provide the link of the Github repository containing the code in the paper.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: In the appendix and in the README of our code repository, we added full implementation details and configurations.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
Answer: [Yes]
Justification: We added a full statistical explanation in the appendix.
Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: In the appendix, we provide our machines, including CPUs, GPUs.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: We did everything as in the rules.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [NA]
    Justification: We don't have any social impact.
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No danger in our code.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: All we used is open source.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We have no assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We used only for minor editing such as fixing typos, grammatical errors etc.

    Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

APPENDIX

8 Proofs of the technical results

8.1 Proofs of existential results on MAS functions

Theorem 1 (Dimension of MAS Function)**.** *For a finite ground set V of size n, there exists a MAS function $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^n$. In addition, any MAS function must have a dimension of at least n. In other words, $m^\star(V, \infty) = n$.*

Proof. The construction of a monotone embedding with $n = |V|$ simply uses one-hot encoding. Namely, we identify V with $[n]$. For every $S \in \mathcal{P}_{<\infty}(V)$, we define

$$F(S) = \sum_{s \in S} e_s \in \mathbb{R}^n \tag{10}$$

where $e_s \in \mathbb{R}^n$ is the vector with $e_s[s] = 1$ and $e_s[j] = 0$ for all $j \neq s$. We now show it satisfies all conditions: Be $S \subseteq T$, then by definition, $\forall v \in V, c_S(v) \leq c_T(v)$, then $F(S) = \sum_{s \in S} e_s \leq \sum_{t \in T} e_t = F(T)$ thus $F(S) \leq F(T)$. On the other hand, if $F(S) \leq F(T)$ then $\forall v \in V, c_S(v) \leq c_T(v)$, and thus $S \subseteq T$.

For the second direction, assume there exists MAS $F : V \to \mathbb{R}^m$ for $m \leq n - 1$. For every output dimension $i \in [m]$, there is a "maximal singleton element" $v_i^* \in V$ such that $F(\{v_i^*\})[i] \geq F(\{v\})[i], \forall v \in V$. The value of the set function over the union of such v_i^*-s across $i \in [m]$ will dominate any singleton that is disjoint from this collection, thanks to the monotonicity of F. As $m < |V|$, then $T := \bigcup_{i=1}^m \{v_i^*\}$ does not cover V. Thus, we can select an element disjoint from this collection named $s \in V$, which gives disjoint sets $S = \{s\}, T$ with $F(S) \leq F(T)$, which contradicts separability. □

Theorem 2 (Upper Bound on $m^\star(V, k)$)**.** *Let $k < n$ be natural numbers, and let V be a ground set with $|V| = n$. Then there exists a MAS function $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ with embedding dimension $m = (k + 2)^{k+2} \log(n)$.*

Proof. For convenience of notation, in this proof we will assume without loss of generality that $V = [n]$. As discussed in the main text, the function $F_{a_1,\ldots,a_m}$ will be weakly monotone for any choice of non-negative vectors. Our goal is to show that there exists a set of parameters $a_1, \ldots, a_m \in [0, 1]^n$, such that the obtained $F_{a_1,\ldots,a_m}$ is a monotone embedding, which means that, if $S \not\subseteq T$, then there exists some j such that $h_j(S) > h_j(T)$, or in other words $\langle f(S), a_j \rangle > \langle f(T), a_j \rangle$. We note that to prove this, it is sufficient to consider pairs (S, T) which we will call *extreme pairs*. This means that S consists of a single element $x \in V$, and T consists of k elements in V but does not contain x (but repetitions are allowed). Indeed, if the claim is true for such pairs, and $S, T \in \mathcal{P}_{\leq k}(V)$ such that $S \not\subset T$, then there exists an element x which is in S more times than it's in T. First, remove all copies of x from T, and we know that at least one copy of x will be in S, note that the difference between the function values remains the same. Then, we will 'replace' S with a smaller set $\tilde{S} = \{x\}$, and $\tilde{T}$ from adding more elements to T from $V \setminus \{x\}$ to obtain a new set $\tilde{T}$ of maximal cardinality k. We then will have that $(\tilde{S}, \tilde{T})$ is an extreme pair, and so for some j we will have that $h_j(\tilde{S}) > h_j(\tilde{T})$. Therefore, using the weak monotonicity of the function h_j, we will have

$$h_j(S) \geq h_j(\tilde{S}) > h_j(\tilde{T}) \geq h_j(T).$$

Given an extreme pair (S, T) we consider the set of all 'bad' a, namely

$$B_{S,T} := \{a \in [0, 1]^n, \langle f(S), a \rangle > \langle f(T), a \rangle\}$$

Our first goal will be to bound the Lebesgue measure of this set. Recall that S is a singleton $S = \{s\}$, and T does not contain this singleton. Accordingly,

$$B_{S,T} = \{a \in [0, 1]^n | \quad a_s \leq \sum_{t \in T} a_t\}$$

We will compute the probability of the complement of this set, namely vectors a such that $a_s > \sum_{t \in T} a_t$. First denote by

$$E_{S,T} = \{a \in [0, 1]^n | \forall t \in T, a_t < \frac{a_s}{k}\}$$

Note, that

$$E_{S,T} \subseteq \{a \in [0,1]^n | \quad a_s > \sum_{t \in T} a_t\}$$

So let's compute the measure of $E_{S,T}$:

$$\text{Leb}(E_{S,T}) = \text{Leb}(\{a \in [0,1]^n | \forall t \in T, a_t \leq \frac{a_s}{k}\}) = \int_0^1 (\frac{a_s}{k})^k da_s = \frac{1}{k^k \cdot (k+1)} \geq \frac{1}{(k+1)^{k+1}}$$

So,

$$\text{Leb}(B_{S,T}) \leq 1 - \text{Leb}(E_{S,T}) \leq 1 - \frac{1}{(k+1)^{k+1}}$$

Now, consider the set

$$B^m_{S,T} = \{a_1, \ldots, a_m \in B_{S,T}\},$$

Then we have

$$\text{Leb}(B^m_{S,T}) \leq \left(1 - \frac{1}{(k+1)^{k+1}}\right)^m$$

The vectors $a_1, \ldots, a_m$ will not define a monotone embedding, if there exists an extreme pair S, T which is in $B^m_{S,T}$. The probability of this happening can be bounded by a union bound. There are $n \cdot \binom{n-1}{k}$ extreme pairs, and for simplicity we will replace this number with a larger but simpler expression $n^{k+2} > n \cdot \binom{n-1}{k}$. The union bound will then give us

$$\text{Leb}\{a_1, \ldots, a_m \text{ do not define a monotone embedding } \} \leq n^{k+2} \left(1 - \frac{1}{(k+1)^{k+1}}\right)^m$$

It is sufficient to show that this expression is smaller than 1, for our m, which, by taking a logarithm, is equivalent to requiring that

$$(k+2)\ln(n) + m \ln\left(1 - \frac{1}{(k+1)^{k+1}}\right) < 0$$

Using Taylor's expansion it can be shown that $\ln(1-x) < -x$ for $x \in (0,1)$, and as a result it is sufficient to choose m so that

$$(k+2)\ln(n) - \frac{m}{(k+1)^{k+1}} < 0$$

or equivalently

$$m > (k+1)^{k+1} \cdot (k+2) \cdot \ln(n).$$

For convenience we slightly enlarge this lower bound to obtain $m \geq (k+2)^{k+2} \ln(n)$ □

Theorem 3 (Lower bounds on $m^\star(V,k)$)**.** *Let $k \geq 2$ and n be natural numbers, and let V be a ground set with $|V| = n$. Then the smallest possible dimension $m^\star(V,k)$ of a MAS function satisfied $m^\star(V,k) \geq \log_2(\log_3 n)$. Moreover, if $k \leq \frac{n-1}{2}$ then $m^\star(V,k) \geq 2k$*

Proof. The proof is partially inspired by the proof in [19].

Lower bound on k: We want to show the lower bound of $m^\star(V,k) \geq \min(2k, n-2)$. As in the proof of Theorem 1, we select a "maximal singleton element" $v_i^* \in V$ for every dimension $i \in \{1, 2, \cdots, m^\star(V,k)\}$. Thus we must have:

$$\text{for each } i \in [m^\star(V,k)]: \; F(\{v_i^*\})[i] \geq F(\{v\})[i], \forall v \in V$$

Now, if $m^\star(V,k) \leq |V| - 2$, we can find two elements $\{u_1, u_2\}$ disjoint from the collection $\mathcal{V} := \bigcup_{i=1}^{m^\star(V,k)} \{v_i^*\}$. Without loss of generality, we can assume that $F(\{u_1\})[i] \geq F(\{u_2\})[i]$ for most indices in $[m^\star(V,k)]$. We can then construct a set T containing u_1 and all the v_i^* for all indices i for which $F(\{u_1\})[i] < F(\{u_2\})[i]$, thus we define:

$$T := \{u_1\} \cup \left\{v_j^* : F(\{u_1\})[j] < F(\{u_2\})[j]\right\}$$

We then have that:

$$F(\{u_2\})[i] \leq \max\{F(\{u_1\})[i], F(\{v_i^*\})[i]\} \leq F(T)[i], \forall i \in [m^\star(V,k)]$$

where the last inequality uses the monotonicity of F. This implies that $F(S) \leq F(T)$ However, S is not a subset of T. It follows that $k \leq |T| - 1 \leq m^\star(V,k)/2$, as otherwise separability of F in $\mathcal{P}_{\leq k}(V)$ is violated. Hence, we get that: $m^\star(V,k) \geq 2k$

Lower bound on n: For this proof, we will use the Erdös-Szekeres theorem, which in particular states, for a natural $N \geq 2$, that a sequence of real numbers with $\geq N^2$ elements has a subsequence of cardinality N which is monotone (either monotonously increasing or monotonously decreasing). Assume we have $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}^m$ a monotone embedding, where $V = [n]$. Assume by way of contradiction that:

$$\log_2(\log_3 n) > m \text{ or equivalently } n > 3^{2^m}$$

Consider the first coordinate and look at the sequence

$$F(\{1\})[1], \ldots, F(\{n\})[1]$$

This is a real-valued sequence, and so there is a monotone subsequence $a_1 < a_2 < \ldots, < a_\ell$ of the original sequence $1, \ldots, n$, with $\ell = \sqrt{3^{2^m}} = 3^{2^{m-1}}$, such that:

$$F(a_1)[1] \leq F(a_2)[1] \leq \ldots \leq F(a_\ell)[1] \text{ or } F(a_1)[1] \geq F(a_2)[1] \geq \ldots \geq F(a_\ell)[1]$$

Next, we consider the second coordinate of this subsequence, namely $F(a_1)[2], \ldots, F(a_\ell)[2]$ and obtain a new subsequence $\{a_j\}$ such that both the sequences: $\{F(a_j)[1]\}$ and $\{F(a_j)[2]\}$ are ordered monotonically, and the size of the subsequence is the square root of the previous one, namely $3^{2^{m-2}}$. After doing this m times, we have a subsequence of size 3. Namely, we have three distinct elements $u, v, w \in V$ such that for all $i = 1, \ldots, m$,

$$\text{either } F(\{u\})[i] \leq F(\{v\})[i] \leq F(\{w\})[i] \text{ or } F(\{u\})[i] \geq F(\{v\})[i] \geq F(\{w\})[i]$$

It follows that for all i, we have:

$$F(\{v\})[i] \leq \max\left(F(\{u\})[i], F(\{w\})[i]\right) \leq F(\{u, w\})[i]$$

Thus, $F(\{v\}) \leq F(\{u, w\})$, which is a contradiction since $\{v\}$ is not contained in $\{u, w\}$. □

A refined lower bound: Now, we may extend the above proof idea to generalize the lower bound on $m^\star(V, k)$, which is useful specially when $|V| >> k$. In such cases, the following gives a tighter lower bound on $m^\star(V, k)$ which is quadratic in k.

Lemma 15. *For all $\ell \in [k]$, we have $m^\star(V, k) \geq \min\left(|V| - \ell, \ell k + 2\ell - \ell^2\right)$. More specifically, if $\ell = \frac{k+2}{2}$, then $m^\star(V, k) \geq \min\left\{|V| - \frac{k+2}{2}, \left(\frac{k+2}{2}\right)^2\right\}$*

Proof. Consider any $\ell \in [k]$. Like in the proof just before, we consider a "maximal singleton element" $v_i^* \in V$ for every dimension $i \in \{1, 2, \cdots, m^\star(V, k)\}$ and obtain the collection $\mathcal{V} := \bigcup_{i=1}^{m^\star(V,k)} \{v_i^*\}$. Let us assume $m^\star(V, k) \leq |V| - \ell$. Thus, we can produce the collection $\mathcal{U} := \{u_1, u_2, \cdots, u_\ell\}$ disjoint from $\mathcal{V}$, i.e $\mathcal{V} \cap \mathcal{U} = \emptyset$. For each $u_q \in \mathcal{U}$, let m_q be the number of co-ordinates $j \in [m^\star(V, k)]$ such that $F(\{u_q\})[j] > F(\{u_r\})[j], \forall r \in [\ell] \setminus \{q\}$. Clearly, $\sum_{q \in [\ell]} m_q \leq m^\star(V, k)$, thus, $\exists \ell_0 \in [\ell] \setminus$ such that $m_{\ell_0} \leq \frac{m^\star(V,k)}{\ell}$. Like before, we consider the sets $S = \{u_{\ell_0}\}$. We construct T as follows: we select $\mathcal{U} \setminus \{u_{\ell_0}\}$ and take union with all $\left\{v_j^*\right\}$ for each dimension j in which $F(\{u_{\ell_0}\})[j] > F(\{u_r\}), \forall r \in [\ell] \setminus \{\ell_0\}$. Since $\left\{v_j^*\right\}$ is the maximal singleton for the dimension j, it follows that $F(\{u_{\ell_0}\})[j] \leq F(\left\{v_j^*\right\})[j]$. Hence, if we define:

$$T = (\mathcal{U} \setminus \{u_{\ell_0}\}) \bigcup \left\{v_j^* : F(\{u_{\ell_0}\})[j] > F(\{u_r\}), \forall r \in [\ell] \setminus \{\ell_0\}\right\}$$

By the choice of ℓ_0, we have that $\left|\left\{v_j^* : F(\{u_{\ell_0}\})[j] > F(\{u_r\}), \forall r \in [\ell] \setminus \{\ell_0\}\right\}\right| \leq \frac{m^\star(V,k)}{\ell}$ Now, we must have that, $\forall j \in [m^\star(V, k)], \exists t \in T$ such that $F(\{t\})[j] \geq F(\{u_{\ell_0}\})[j]$. By monotonicity of F, we thus have that $F(\{u_{\ell_0}\})[j] \leq F(T)[j], \forall j \in [m^\star(V, k)]$ which implies $F(S) \leq F(T)$. But by design, $S \cap T = \emptyset$, which contradicts separability of F. Thus, we must have that, $|T| > k$, i.e $|T| - 1 \geq k$. But as stated before, $|T| \leq \frac{m^\star(V,k)}{\ell} + \ell - 1$. Combining these two, we get that: $\frac{m^\star(V,k)}{\ell} \geq k + 2 - \ell$, which implies $m^\star(V, k) \geq k\ell + 2\ell - \ell^2$, thus proving our result.

Now, we already have a sufficient upper bound on $m^\star(V, k)$ from Theorem 2 that's $O\left(k^{k+2}\log(n)\right)$. Now, when $n >> k$ we have that $O\left(k^{k+2}\log(n)\right) < n - O(k)$, in which case we may apply $\ell = \frac{k+2}{2}$ in the above lemma. And since we already know that $O\left(k^{k+2}\log(n)\right) < n - O(k)$, we thus get that $m^\star(V, k) < n - O(k)$. Thus, in this case the above lemma gives us $m^\star(V, k) \geq \left(\frac{k+2}{2}\right)^2$, which is a tigher quadratic lower bound in k. □

When $k = 1$ In the degenerate case where only a single set element is allowed, $k = 1$, it is possible to construct a MAS function even for the uncountable ground set $V = [-1, 1]$ by defining $F : \mathcal{P}_{\leq 1}([-1, 1]) \to \mathbb{R}^2$ via

$$F(S) = (-1, -1) \text{ if } S = \emptyset \quad \text{and} \quad F(S) = (-x, x) \text{ if } S = \{x\}, \tag{11}$$

Separable embeddings (even non monotone) don't exist: Now, as discussed in the main text, we give a justification here on why we choose to relax separability and not monotonicity in the definition of wekaly MAS embeddings. Here, we show that, with the added assumptions of injectivity and continuity, even non-monotone set-t-vector embeddings do not exist in any dimension for uncountable ground sets.

Theorem 16. *There does not exist a continuous injective separable set function taking values in $\mathbb{R}^m$ for any m when the ground set V is an open subset of $\mathbb{R}^d$*

Proof. We restrict ourselves to $\mathcal{P}_{=n}(V)$, i.e, only subsets of size n. Since F is continuous and permutation-invariant, according to Theorem-7 of [20], $f(\{x_1, \cdots, x_n\}) = \rho(\sum_{i=1}^{n} \phi(x_i))$, where ρ, ϕ are continuous and bijective(given in condition). Now, consider the space $\mathcal{S} \subset [0, 1]^n, \mathcal{S} = \{(x_1, \cdots, x_n) : x_1 < x_2 < x_3 < \cdots < x_n; 0 < x_i < 1, \forall i \in [n]\}$. Note that $\mathcal{S}$ is open in $[0, 1]^n$ (to see this, let $\epsilon = \min_{i \in [n-1]}(x_{i+1} - x_i)$. The ball of radius $\frac{\epsilon}{2}$ is contained in the domain). Since $F : \mathcal{S} \to \mathbb{R}^n$ is an injective, continuous map. Thus, by Invariance of Domain Theorem([34]), $F(\mathcal{S})$ is open in $\mathbb{R}^n$ as well. Note that each element of S uniquely corresponds to an element of $\mathcal{P}_{=n}(V)$. Thus, $F(\mathcal{P}_{=n}(V))$ is open in $\mathbb{R}^n$. Now, consider any set $A = \{a_1, \cdots, a_n\}$. Since F is an open map, we have that $\exists \delta > 0$ such that $B(F(A), \delta) \subseteq F(\mathcal{P}_{=n}(V))$. Thus, $\exists B \in \mathcal{P}_{=n}(V)$ such that $F(B)_i = F(A)_i + \frac{\delta}{2n}, \forall i \in [n]$. But then, $F(B) > F(A)$ implies $A \subset B$ but $|A| = |B| = n$ by construction. This gives a contradiction. □

8.2 Proofs of results on on weakly MAS functions

Proposition 6. If the deep set model in (2) is implemented with a vector-to-vector monotonously increasing function M_2 and a non-negative activation σ, then $F(\bullet; (A, b))$ is monotone for every (A, b). If in addition σ is monotone (increasing or decreasing), then there exists $S \not\subseteq T$ with $F(S; (A, b)) \leq F(T; (A, b))$ for all A, b.

Proof. Let $x < y$ and consider the two sets:

$$S = \{\frac{x + y}{2}\}, T = \{x, y\}$$

And we claim that for all A, b, although $S \not\subseteq T$, $F(S, (A, b)) \leq F(T, (A, b))$. Note that as $x < \frac{x+y}{2} < y$, then it must have been that

$$Ax + b \leq A \cdot \frac{x + y}{2} + b \leq Ay + b$$

Or

$$Ay + b \leq A \cdot \frac{x + y}{2} + b \leq Ax + b$$

Then, by (weakly) monotonicity of the activation σ and of M_2, it must be that

$$\sigma(A \cdot \frac{x + y}{2} + b) \leq \sigma(A \cdot x + b)$$

Or that

$$\sigma(A \cdot \frac{x + y}{2} + b) \leq \sigma(A \cdot y + b)$$

In any case, it's true that

$$\sigma(A \cdot \frac{x + y}{2}) \leq \sigma(A \cdot x + b) + \sigma(A \cdot y + b)$$

Then, by monotonicity of M_2,

$$F(S, (A, b)) = \mathrm{M}_2 \cdot (\sigma(A \cdot \frac{x + y}{2} + b)) \leq \mathrm{M}_2 \cdot (\sigma(A \cdot x + b) + \sigma(A \cdot y + b)) = F(T, (A, b))$$

Thus,

$$F(S, (A, b)) \leq F(T, (A, b))$$

□

Proposition 17. *Let $V = \mathbb{R}^d$ and let $F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}^D$ be a Set Transformer defined by $F(S; W_Q, W_K, W_V) :=$ SumPool(Attn(S)). Then F is not a point-wise monotone function.*

Proof. We need to cone up with a class of tuple of parameters (W_Q, W_K, W_V) such that the function $S \mapsto F(S; W_Q, W_K, W_V)$ is not monotone. Choose $W_Q = W_K = W$ such that W is full-rank. Choose any non-zero W_V. Now, choose some vector $\mathbf{x}_1$ such that $W_V\mathbf{x}_1$ is non-zero. By swapping $\mathbf{x}_1$ with $-\mathbf{x}_1$ if necessary, we can choose some index j such that for $\mathbf{v}_1 = W_V\mathbf{x}_1$, the j-th index is negative, *i.e.*$\mathbf{v}_1[j] < 0$. By assumption we also have that $\mathbf{q}_1 = \mathbf{k}_1$ is a non-zero vector, and thus has a positive norm. Now let S be the singleton set $S = \{\mathbf{x}_1\}$, which is a subset of $T = \{\mathbf{x}_1, \mathbf{0}_d\}$ where $\mathbf{0}_d$ is the $d-$dimensional all-zero vector. Note that, $F(S; W_Q, W_K, W_V) = \mathbf{v}_1$. And since $\mathbf{q}_2, \mathbf{k}_2, \mathbf{v}_2$ are all zero, we have: $F(T; W_Q, W_K, W_V) = (\alpha_{1,1} + \alpha_{1,2})\mathbf{v}_1$. Note that, $\frac{\alpha_{1,1}}{\alpha_{1,1}+\alpha_{1,2}} = \frac{e^{\|\mathbf{q}_1\|^2}}{e^{\|\mathbf{q}_1\|^2}+1} + \frac{1}{2} > 1$. Hence, it follows that $F(T; W_Q, W_K, W_V)[j] < F(S; W_Q, W_K, W_V)[j]$ and thus, the non-monotonicity follows □

Proposition 8. If M_2 is strictly monotone increasing, and σ is a hat activation function, then $F(S, (a, b)) = \mathrm{M}_2(\sum_{x\in S} \sigma(a^\top x + b))$ is weakly MAS function.

Proof. It's clear the function is a set function. Be $S \subseteq T$, as σ is positive, then for any a, b,

$$\sum_{x\in S} \sigma(a^T x + b) \leq \sum_{x\in S} \sigma(a^T x + b) + \sum_{x\in T\setminus S} \sigma(a^T x + b) = \sum_{x\in T} \sigma(a^T x + b)$$

So

$$F(S, (a, b)) \leq F(T, (a, b))$$

Denote by $[s_1, s_2]$ the support of σ,and choose some point in it such that $s \in [s_1, s_2] : \sigma(s) > 0$. Now, be $S \not\subseteq T$, then there exists $z \in V : C_S(z) > C_T(z)$. Now, remove z from both sets $C_T(z)$ times. This subtracts from both function values the same amount. Now, we have that $z \in S, z \notin T$. Denote by $\epsilon = \frac{min_{t\in T}|z-t|>0}{2}$. Choose

$$a = \frac{s_2 - s_1}{\epsilon}, b = -(a \cdot z - s)$$

We claim that $\forall t \in T, \sigma(a \cdot t + b) = 0$. Assume it's not zero, thus $s_1 < a \cdot t + b < s_2$. Thus

$$s_1 < \frac{s_2 - s_1}{\epsilon} \cdot t - a \cdot z + s < s_2$$

Thus,

$$s_1 - s < \frac{s_2 - s_1}{\epsilon} \cdot t - \frac{s_2 - s_1}{\epsilon} \cdot z < s_2 - s$$

Thus,

$$\frac{s_1 - s}{s_2 - s_1} \cdot \epsilon < t - z < \epsilon \cdot \frac{s_2 - s}{s_2 - s_1}$$

But,

$$|\frac{s_1 - s}{s_2 - s_1}| \leq 1$$
$$|\frac{s_2 - s}{s_2 - s_1}| \leq 1$$

Thus,

$$-\epsilon \leq t - z \leq \epsilon$$

A contradiction to the definition of ϵ. Thus, $F(T, (a, b)) = \mathrm{M}_2(0)$. But

$$\sigma(a \cdot z + b) = \sigma(a \cdot z - a \cdot z + s_1) = \sigma(s_1) > 0$$

Thus

$$F(S, (a, b)) = \mathrm{M}_2(\sum_{x\in S} \sigma(a^\top x + b)) \geq \mathrm{M}_2(\sigma(s_1)) > \mathrm{M}_2(0) = F(T, (a, b))$$

And we are done. □

Proposition 18. *Given $V \subseteq \mathbb{R}^d$, we consider affine transformations $\mathcal{A}_2 : \mathbb{R}^d \to \mathbb{R}^3, \mathcal{A}_1 : \mathbb{R}^3 \to \mathbb{R}$. Thus, here $\mathcal{A}_2(t) := A_2 t + b_2$ and $\mathcal{A}_1(z) = a_1^\top z + b_1$, where $A_2 \in \mathbb{R}^{3\times d}, b_2 \in \mathbb{R}^3, a_1 \in \mathbb{R}^3, b_1 \in \mathbb{R}$*

are the respective parameters. Then, the set functions of the form

$$F : \mathcal{P}_{<\infty}(V) \to \mathbb{R}, F(S; \mathcal{A}_1, \mathcal{A}_2) = \mathrm{M}_2\left(\sum_{x\in S} \mathrm{ReLU} \circ \mathcal{A}_1 \circ \mathrm{ReLU} \circ \mathcal{A}_2(x)\right) \tag{3}$$

are weakly MAS functions.

Proof. Consider the function TRI as in the third plot of Fig. 2. We know that TRI belongs to the class of Hat functions. From 9 we know that, for any $S \not\subseteq T \in \mathcal{P}_{<\infty}(V)$, $\exists \boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $\sum_{x\in S} \mathrm{TRI}(\boldsymbol{a}^\top x + b) > \sum_{y\in T} \mathrm{TRI}(\boldsymbol{a}^\top y + b)$. Now, consider the following parameters:

$$\begin{aligned} &\boldsymbol{A} = [\boldsymbol{a}, \boldsymbol{a}, \boldsymbol{a}]^\top \in \mathbb{R}^{3\times d}, \boldsymbol{b} = [b+1, b-1, b]^\top \in \mathbb{R}^3 \\ &\boldsymbol{a}_1 = [1, 1, -2]^\top \in \mathbb{R}^3, \text{ and } b_1 = 0 \end{aligned} \tag{12}$$

Then, for the above values of parameters, we have:

$$\mathrm{ReLU} \circ \mathcal{A}_2(x) = \left[\mathrm{ReLU}\left((\boldsymbol{a}^\top x + b) + 1\right), \mathrm{ReLU}\left((\boldsymbol{a}^\top x + b) - 1\right), \mathrm{ReLU}\left(\boldsymbol{a}^\top x + b\right)\right]$$

Thus, we have:

$$\begin{aligned} &\boldsymbol{a}_1^\top \left(\mathrm{ReLU} \circ \mathcal{A}_2(x)\right) + b_1 \\ &= \mathrm{ReLU}\left((\boldsymbol{a}^\top x + b) + 1\right) + \mathrm{ReLU}\left((\boldsymbol{a}^\top x + b) - 1\right) - 2\mathrm{ReLU}\left(\boldsymbol{a}^\top x + b\right) \\ &= \mathrm{TRI}\left(\boldsymbol{a}^\top x + b\right) \end{aligned}$$

Thus we have, $\mathrm{ReLU} \circ \mathcal{A}_1 \circ \mathrm{ReLU} \circ \mathcal{A}_2(x) = \mathrm{ReLU} \circ \mathrm{TRI}\left(\boldsymbol{a}^\top x + b\right) = \mathrm{TRI}\left(\boldsymbol{a}^\top x + b\right)$. Hence, $F(S) = \sum_{x\in S} \mathrm{TRI}\left(\boldsymbol{a}^\top x + b\right) > \sum_{y\in S} \mathrm{TRI}\left(\boldsymbol{a}^\top y + b\right) = F(T)$ and weak separability is proved. Since M_2 is monotone increasing and ReLU is non-negative, it follows that F is monotone as well, hence weakly MAS. □

8.3 Hölder separability of MAS functions

Theorem 12 (F is lower Hölder)**.** *Let $V \subset \mathbb{R}^d$ be a compact set, and σ a Hat activation function which is piecewise continuously differentiable supported in some interval $[\gamma_1, \gamma_2]$, and satisfying the condition: $\lim_{t\to\gamma_1^+} \frac{\mathrm{d}\sigma}{\mathrm{d}t} > 0$. Let $\mathrm{M}_2 : \mathbb{R} \to \mathbb{R}$ be a lower Lipschitz function. Consider the function $F(S; (a, b, c)) = \mathrm{M}_2\left(\sum_{x\in S} \sigma\left(\frac{a^\top x + b}{c}\right)\right)$, where the multisets S come from $\mathcal{P}_{\leq k}(V)$, and $a \sim \mathrm{Unif}(\mathcal{S}^{d-1}), b \sim \mathrm{Unif}([-1, 1]), c \sim \mathrm{Unif}((0, 2])$. Then, $F(\bullet, (a, b, c))$ is Monotone Hölder separable with exponent $\lambda = 2$.*

Proof. We consider $V \subset \mathbb{R}^d$ to be a compact set with maximum norm of 1. Also, we note that, if $\sigma(x)$ is a Hat function with support in $[\gamma_1, \gamma_2]$, then $\sigma(\frac{x-\gamma_1}{\gamma_2-\gamma_1})$ is a Hat function with support in $(0, 1)$. For this proof, we thus consider σ to have support in $[0, 1]$. For support in general $[\gamma_1, \gamma_2]$ and a general norm bound of V, the distributions on b, c needs to be scaled and shifted to get same Lower Hölder results with the Hölder constants scaled appropriately. [*]. We recall the definition of $\Delta(S, T)$ from equation 4. Let $S = \{x_1, \cdots, x_M\}$ and let $T = \{y_1, \cdots, y_N\}$, where $S, T \subset V$. Clearly by the definition of $\Delta(S, T)$, it is sufficient to consider $|S| \leq |T|$, i.e $M \leq N \leq k$.

Using the definition of EMD gives us the following equivalent definition of $\Delta(S, T)$: If $\Omega_{M,N} := \{\tau : [M] \to [N], \tau \text{ is injective}\}$, then $\Delta(S, T) = \min_{\tau\in\Omega_{M,N}} \sum_{i=1}^M \|x_i - y_{\tau(i)}\|_2$. Let τ^* be the optimal coupling between $[M]$ and $[N]$, i.e $\tau^* = \arg\min_{\tau\in\Omega_{M,N}} \sum_{i=1}^M \|x_i - y_{\tau(i)}\|_2$, which means $\Delta(S, T) := \sum_{i=1}^M \|x_i - y_{\tau^*(i)}\|_2$. Now, $\forall i \in [M]$, we define $\Delta_i := \|x_i - y_{\tau^*(i)}\|_2$. For any $x \in \mathbb{R}^d, r > 0$, let $\mathcal{B}(x, r)$ denote the Euclidean ball centered at x with radius r. We define $\Omega_i := \left\{\ell \in [M] : y_{\tau^*(\ell)} \in \mathcal{B}(x_i, \Delta_i)\right\}, \forall i \in [M]$. With these notations, we make the following observation:

Claim-1: Consider any $r \in (0, k)$. Then, $\exists i_0 \in [M]$ such that $\sum_{\ell\in\Omega_{i_0}} \Delta_\ell \leq \Delta_{i_0} r$ and $\Delta_{i_0} \geq \frac{r^k}{k^{k+1}} \Delta(S, T)$

[*] If $\sup_{x\in V} \|x\| \leq B$, and $\mathrm{supp}(\sigma) \subseteq [\gamma_1, \gamma_2]$ then distributions of b, c should be shifted by γ_1 and linearly scaled by $\delta := \frac{B}{\gamma_2-\gamma_1}$ and to be: $b \sim \mathrm{Unif}(-\gamma_1 - \delta, -\gamma_1 + \delta)$ and $c \sim \mathrm{Unif}(0, 2\delta)$ respectively

Proof. For the sake of contradiction, we assume otherwise, i.e $\forall i \in [M]$, we have: either $\Delta_i \leq \frac{r^k}{k^{k+1}}\Delta(S,T)$ or $\sum_{\ell\in\Omega_i}\Delta_\ell \geq \Delta_i r$. WLOG, assume the following order on Δ_i's: $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_M$. We build the following directed graph $\mathcal{G}$ on $[M]$. Start with the node $v_0 = 1$ and add we add one node to $\mathcal{G}$ at a time. At step t, we select $v_t := \arg\max_{u\in\Omega_{v_{t-1}}}\Delta_u$ and add the edge (v_{t-1}, v_t) to $\mathcal{G}$ along with the new node v_t.

Note that, by definition of $\Delta(S,T)$ we have that: $\Delta(S,T) = \sum_{i=1}^M \Delta_i$. This, implies that: $\Delta_{v_0} \geq \frac{\Delta(S,T)}{M}$, as we take argmax over entire $[M]$ .Thus, $\Delta_1 \geq \Delta(S,T)/M \geq \Delta(S,T)/k$. Thus, by the contradictory assumption, we must have: $\sum_{u\in\Omega_{v_0}}\Delta_u \geq \Delta_1 r$. Since $v_1 = \arg\max_{u\in\Omega_{v_0}}\Delta_u$, it thus follows that: $\Delta_{v_1} \geq \frac{r}{k}\Delta_1 \geq \frac{r}{k^2}\Delta(S,T)$. Following the similar argument, we thus have, $\Delta_{v_2} \geq \frac{r}{k})\Delta_{v_1}$ and so on. Thus, $\forall t \in \mathbb{N}$ we must have: $\sum_{u\in\Omega_{v_t}}\Delta_u \geq \Delta_{v_t} r$. Thus, for v_{t+1}, defined by: $\arg\max_{u\in\Omega_{v_t}}\Delta_u$, we must have: $\Delta_{v_{t+1}} \geq \frac{r}{k}\Delta_{v_t}$. Since this holds for every $t \in \mathbb{N}$ and since $r > 0$ by hypothesis, we must have: $\Delta_{v_t} \geq \frac{r}{k}\Delta_{v_{t-1}} \geq \left(\frac{r}{k}\right)^2\Delta_{v_{t-1}} \geq \cdots \geq \left(\frac{r}{k}\right)^t\Delta_{v_0}$.

As we work with finite M, N either $\mathcal{G}$ is a DAG and this process terminates; or $\mathcal{G}$ has a cycle. In former case, at termination step $\boldsymbol{t_t}$ we cannot add any more nodes, thus $\Omega_{\boldsymbol{t_t}} = \emptyset$. Since all nodes are unique in this case, we must have $\boldsymbol{t_t} \leq k$. So, $\Delta_{\boldsymbol{t_t}} \geq \left(\frac{r}{k}\right)^{\boldsymbol{t_t}}\Delta_{v_0} \geq \frac{1}{k}\left(\frac{r}{k}\right)^{\boldsymbol{t_t}}\Delta(S,T)$. Since $\frac{r}{k} \in (0,1)$ by hypothesis and since $\boldsymbol{t_t} \leq k$, we thus have: $\Delta_{\boldsymbol{t_t}} \geq \frac{1}{k}\left(\frac{r}{k}\right)^k\Delta(S,T)$, and hence done with this case.

For the second case, consider the smallest cycle in $\mathcal{G}$. Let this cycle of length p be $\mathcal{C} = \{v_\beta \to v_{\beta+1} \to \cdots \to v_{\beta+p-1} \to v_\beta\}$ for some $\beta \geq 0$. Then, we have: $\beta + i + 1 \in \Omega_{\beta+i}, \forall i \in \{0, \cdots p-1\}$, where addition of indices is modulo p. Thus we have: $\|x_{v_\beta+i} - y_{\tau^*(v_\beta+i+1)}\|_2 < \|x_{v_\beta+i} - y_{\tau^*(v_\beta+i)}\|_2, \forall i \in \{0, \cdots, p-1\}$. Taking sum, we have the following: $\|x_{v_\beta} - y_{\tau^*(v_\beta+1)}\|_2 + \|x_{v_\beta+1} - y_{\tau^*(v_\beta+2)}\|_2 + \cdots + \|x_{v_\beta+p-1} - y_{\tau^*(v_\beta)}\|_2 < \sum_{j=0}^{p-1}\|x_{v_\beta+j} - y_{\tau^*(v_\beta+j)}\|_2$. Hence, switching from the coupling $i \mapsto \tau^*(i)$ to $i \mapsto \tau^*(i+1)$ for $i \in \{v_\beta, \cdots, v_{\beta+p-2}\}$ and from $v_\beta + p - 1 \mapsto \tau^*(v_\beta + p - 1)$ to $v_\beta + p - 1 \mapsto \tau^*(v_\beta)$ strictly reduces the value of the sum $\sum_{i=1}^M\|x_i - y_{\tau^*(i)}\|_2$, but we started with the optimal coupling τ^*. This gives the reqd. contradiction. □

We also make the following observation regarding this "optimal coupling" τ^*:

Observation-2: For any given $i \in [M]$, if $y_j \in T \cap \Omega_i$ for some $j \in [N]$, then $\exists \ell \in [M]$ such that $j = \tau^*(\ell)$

Proof. The proof follows from the following observation: if $y_j \in \Omega_i$ for any $i \in [M]$, then we must have: $\|x_i - y_j\|_2 \leq \Delta_i = \|x_i - y_{\tau(i)}\|_2$. Thus, if y_j was "free", i.e if $\nexists \ell \in [M] : j = \tau^*(\ell)$, then we can switch from $i \mapsto \tau^*(i)$ to $i \mapsto j$, and this would reduce the sum $\sum_{i=1}^M\|x_i - y_{\tau^*(i)}\|_2$, but τ^* was the optimal coupling. This gives a contradiction □

Now, as per the starting discussion, we consider supp$(\sigma) = [0,1]$, and the proof for general support of $[\gamma_1, \gamma_2]$ requires the scaling and shifting as mentioned before.

By assumption we have, $\lim_{t\to 0^+}\sigma'(t) > 0$. We define $\kappa_1 := \frac{\lim_{t\to 0^+}\sigma'(t)}{2}$. Since σ is piecewise continuously differentiable as per assumption, we must have $\sigma'(t) \geq \kappa_1, \forall t \in (0, \omega)$ for some $\omega \in (0,1)$ by the continuity of σ'. Note that, by Lagrange's Mean Value Theorem, we have that:

$$\begin{aligned}
&\forall x, y \in (0,\omega), \frac{|\sigma(x) - \sigma(y)|}{|x-y|} = \sigma'(z), \text{ where } z \in (x,y) \\
&\implies \forall x, y \in (0,\omega), |\sigma(x) - \sigma(y)| \geq \kappa_1 |x-y|, \text{ since } z \in (0,\omega) \\
&\implies \lim_{y\to 0^+} |\sigma(x) - \sigma(y)| \geq \kappa_1 |x|, \text{ as } \sigma \text{ is cts.} \\
&\implies \sigma(x) \geq \kappa_1 x, \forall x \in (0,\omega), \text{ as } \sigma \geq 0
\end{aligned}$$

Also, σ is upper Lipschitz with constant $\kappa_2 > 0$ per definition of Hat fn. Now, consider the i_0 coming from the Claim as proved above. We define $\omega' := \min(\frac{\omega}{1-\omega}, 1)$. Let $P \in \mathbb{N}$ be such that $\frac{\kappa_1}{P} < \frac{1}{2}$

and $\frac{\kappa_1}{\kappa_2 P} < k$ and $\omega' > \frac{2}{P}$. Note that, such a $P \in \mathbb{N}$ exists, since taking $P \to \infty$ satisfies all the 3 conditions.

Now, for a given $a \in \mathcal{S}^{d-1}$, consider the sets $S_a = \left\{a^\top x : x \in S\right\}, T_a = \left\{a^\top y : y \in T\right\}$. Note that, since we have $\|x_i\|, \|y_j\| \leq 1, \forall i \in [M], j \in [N]$ as per earlier assumption, we get $S_a, T_a \subseteq (-1, 1)$. For given $a \in \mathcal{S}^{d-1}$, we define the "optimal coupling" between the real-valued sets S_a, T_a as: $\tau_a^* := \arg\min_{\tau \in \Omega_{M,N}} \sum_{i=1}^M \left|a^\top x_i - a^\top y_{\tau(i)}\right|$. Based on this, we define $\Delta_i^a := \left|a^\top x_i - a^\top y_{\tau_a^*(i)}\right|$ and $\Omega_i^a = \left\{\ell \in [M] : y_{\tau_a^*(\ell)} \in \mathcal{B}(a^\top x_i, \Delta_i^a)\right\}$. We have: $\Delta(S_a, T_a) := \sum_{i=1}^M \Delta_i^a$. Let $i_0 \in [M]$ be the index that comes from applying the claim to the sets S_a, T_a. Now, we define the regions $\boldsymbol{B}_a := (a^\top x_{i_0} - \omega' \Delta_{i_0}^a, a^\top x_{i_0} - \frac{\Delta_{i_0}^a}{P})$ and for a given pair $(a \in \mathcal{S}^{d-1}, b \in \mathbb{R})$, we consider $\boldsymbol{C}_{a,b} := (\frac{a^\top x_{i_0} - b}{\omega}, a^\top x_{i_0} + \Delta_{i_0}^a - b)$. Thus, we have:

$$
\begin{aligned}
&\left(a^\top x_{i_0} + \Delta_{i_0}^a - b\right) - \left(\frac{a^\top x_{i_0} - b}{\omega}\right) \\
=&\frac{(1-\omega)(b - a^\top x_{i_0}) + \omega \Delta_{i_0}^a}{\omega} \\
=&\frac{1-\omega}{\omega}\left((b - a^\top x_{i_0}) + \frac{\omega}{1-\omega}\Delta_{i_0}^a\right) \\
\geq&\frac{1-\omega}{\omega}\left(b - (a^\top x_{i_0} - \omega' \Delta_{i_0}^a)\right) > 0, \forall b > a^\top x_{i_0} - \omega' \Delta_{i_0}^a
\end{aligned}
$$

The above analysis shows that $\boldsymbol{C}_{a,b} \neq \emptyset, \forall b \in \boldsymbol{B}_a$, so the intervals are well defined. Now, if $b \in \boldsymbol{B}_a$ and $y < a^\top x_{i_0} - \Delta_{i_0}^a$ then we have: $y - b < a^\top x_{i_0} - \Delta_{i_0}^a - a^\top x_{i_0} + \omega' \Delta_{i_0}^a = (\omega' - 1)\Delta_{i_0}^a \leq 0$ by definition of ω'. On the other hand, if $b \in \boldsymbol{B}_a$ and $c \in \boldsymbol{C}_{a,b}$, for given any $y > a^\top x_{i_0} + \Delta_{i_0}^a$ we have: $\frac{y-b}{c} \geq \frac{a^\top x_{i_0} - b + \Delta_{i_0}^a}{a^\top x_{i_0} - b + \Delta_{i_0}^a} = 1$. Combining these, we get that if $\left|y - a^\top x_{i_0}\right| \geq \Delta_{i_0}^a$, then $\frac{y-b}{c} \notin (0, 1), \forall b \in \boldsymbol{B}_a, c \in \boldsymbol{C}_{a,b}$. Thus, conditioning on a given $a \in \mathcal{S}^{d-1}$, we can write the conditional expectation taken over b, c as follows:

$$
\begin{aligned}
&\mathbb{E}_{b,c}\left[\sum_{x \in S} \sigma\left(\frac{a^\top x - b}{c}\right) - \sum_{y \in T} \sigma\left(\frac{a^\top y - b}{c}\right)\right]_+ \\
\geq&\mathbb{E}_{b,c}\left[\left(\sum_{x \in S} \sigma\left(\frac{a^\top x - b}{c}\right) - \sum_{y \in T} \sigma\left(\frac{a^\top y - b}{c}\right)\right)\mathbf{1}_{\{b \in \boldsymbol{B}_a, c \in \boldsymbol{C}_{a,b}\}}\right]_+, \text{using non-negative RV} \\
\geq&\frac{1}{4}\int_{b=a^\top x_{i_0} - \omega' \Delta_{i_0}^a}^{a^\top x_{i_0}} \int_{c=\frac{a^\top x_{i_0} - b}{\omega}}^{a^\top x_{i_0} + \Delta_{i_0}^a - b} \Bigg[\sigma\big(\frac{a^\top x_{i_0} - b}{c}\big) - \\
&\sum_{\ell \in \Omega_{i_0}^a} \left|\sigma\big(\frac{a^\top y_{\tau_a^*(\ell)} - b}{c}\big) - \sigma\big(\frac{a^\top x_\ell - b}{c}\big)\right|\Bigg]_+ \mathrm{d}c\,\mathrm{d}b \\
\geq&\frac{1}{4}\int_{b=a^\top x_{i_0} - \omega' \Delta_{i_0}^a}^{a^\top x_{i_0}} \int_{c=\frac{a^\top x_{i_0} - b}{\omega}}^{a^\top x_{i_0} + \Delta_{i_0}^a - b} \left[\kappa_1 \frac{a^\top x_{i_0} - b}{c} - \kappa_2 \frac{\kappa_1}{P \kappa_2} \frac{\Delta_{i_0}^a}{c}\right]_+ \mathrm{d}c\,\mathrm{d}b, \text{from Lipschitz cts. } \sigma \\
=&\frac{\kappa_1}{4}\int_{b=a^\top x_{i_0} - \omega' \Delta_{i_0}^a}^{a^\top x_{i_0}} \left[a^\top x_{i_0} - b - \frac{\Delta_{i_0}^a}{P}\right]_+ \int_{c=\frac{a^\top x_{i_0} - b}{\omega}}^{a^\top x_{i_0} + \Delta_{i_0}^a - b} \frac{\mathrm{d}c}{c}\,\mathrm{d}b, \text{as } \mathrm{ReLU}(rx) = r\mathrm{ReLU}(x) \\
=&\frac{\kappa_1}{4}\int_{b=a^\top x_{i_0} - \omega' \Delta_{i_0}^a}^{a^\top x_{i_0} - \frac{\Delta_{i_0}^a}{P}} \left(a^\top x_{i_0} - b - \frac{\Delta_{i_0}^a}{P}\right) \ln\left(\frac{\omega(a^\top x_{i_0} + \Delta_{i_0}^a - b)}{a^\top x_{i_0} - b}\right) \mathrm{d}b \\
=&\frac{\kappa_1}{4}\int_{z=\frac{\Delta_{i_0}^a}{P}}^{\omega' \Delta_{i_0}^a} \left(z - \frac{\Delta_{i_0}^a}{P}\right) \ln\left(\frac{\omega(z + \Delta_{i_0}^a)}{z}\right) \mathrm{d}z, \text{putting } z := a^\top x_{i_0} - b
\end{aligned}
$$

$$\geq \frac{\kappa_1}{4} \int_{z=\frac{\Delta^a_{i_0}}{P}}^{\frac{\omega' \Delta^a_{i_0}}{2}} \left(z - \frac{\Delta^a_{i_0}}{P}\right) \ln\left(\omega\left(1 + \frac{\Delta^a_{i_0}}{z}\right)\right) \mathrm{d}z \text{.using non-negativity of integrand}$$

$$\geq \frac{\kappa_1}{4} \ln\left(\omega\left(1 + \frac{2}{\omega'}\right)\right) \int_{z=\frac{\Delta^a_{i_0}}{P}}^{\frac{\omega' \Delta^a_{i_0}}{2}} \left(z - \frac{\Delta^a_{i_0}}{P}\right) \mathrm{d}z$$

$$\geq \frac{\kappa_1}{8} \ln(2-\omega) \left(\frac{\omega'}{2} - \frac{1}{P}\right)^2 (\Delta^a_{i_0})^2, \text{ using } \omega' \leq \frac{\omega}{1-\omega}$$

$$\geq C \cdot \frac{1}{k} \left(\frac{\kappa_1}{\kappa_2 P}\right)^k (\Delta^a(S,T))^2, \text{ where } C > 0 \text{ as } \omega \in (0,1) \text{ and } P \geq \frac{2}{\omega'} \text{ by earlier choice}$$

Thus, we have shown that: $\mathbb{E}_{b,c}\left[\left[F(S) - F(T)\right]_+ \middle| a\right] \geq C' \cdot (\Delta^a(S,T))^2$. By law of total expectation, we have:

$$\mathbb{E}_{a,b,c}\left[F(S) - F(T)\right]_+ = \mathbb{E}_a \mathbb{E}_{b,c}\left[\left[F(S) - F(T)\right]_+ \middle| a\right]$$

$$\geq C' \mathbb{E}_a \left(\Delta^a(S,T)\right)^2 \geq C' \left(\mathbb{E}_a \Delta^a(S,T)\right)^2, \text{by Jensen's ineq.}$$

Thus, to show Lower Hölder separability with exponent $\lambda = 2$, it remains to show that, $\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{S}^{d-1})} \Delta^a(S,T) \geq c_1 \cdot \Delta(S,T)$, where $c_1 > 0$ is a constant. Now, we make the following observation:

Observation-3: Suppose we're given given $\ell \in \mathbb{N}$ and ℓ many non-zero vectors $u_1, u_2, \cdots, u_\ell \in \mathbb{R}^d$. If $a \sim \mathcal{S}^{d-1}$, then $\exists \delta(\ell) > 0$ such that: $\mathbb{P}\left\{\left|a^\top u_i\right| \geq \delta(\ell)\|u_i\|, \forall i \in [\ell]\right\} \geq \frac{1}{2}$

Proof. Firstly, we note that, for a fixed $i \in [\ell]$, $\left|a^\top u_i\right|$ is a continuous non-negative real-valued random variable. Thus, $\exists t_i \in \mathbb{R}_+$ such that $\mathbb{P}\left\{\left|a^\top u_i\right| \geq t_i\right\} \geq \frac{1}{2}$. Now, Choosing $\delta_i := \frac{t_i}{\|u_i\|} > 0$ gives us: $\mathbb{P}\left\{\left|a^\top u_i\right| \geq \delta_i \|u_i\|\right\} \geq \frac{1}{2}, \forall i \in [\ell]$. Since $\delta_i > 0, \forall i \in [\ell]$ and ℓ is finite, we thus have: $\delta := \min_{i \in \ell} \delta_i > 0$. For this particular choice of δ, we have that $\mathbb{P}\left\{\left|a^\top u_i\right| \geq \delta \|u_i\|\right\} \geq \frac{1}{2}, \forall i \in [\ell]$, and the observation is proved. □

. Now, consider all $\boldsymbol{\ell_0} := \binom{N}{M} M!$ vectors $(x_i - y_{\tau(i)}) \in \mathbb{R}^d$, where τ runs over all injective functions from $[M] \to [N]$. For this particular choice of $\boldsymbol{\ell_0}$, we get from the last observation that, $\exists \delta(N,M) > 0$ such that $\mathbb{P}\left\{\left|a^\top \left(x_i - y_{\tau(i)}\right)\right| \geq \delta(N,M)\|x_i - y_{\tau(i)}\|\right\} \geq \frac{1}{2}$, for all possible injective $\tau : [M] \to [N]$, we call this set of $a \in \mathcal{S}^{d-1}$ to be $A \subseteq \mathcal{S}^{d-1}$. Now, given a fixed $a \in \mathcal{S}^{d-1}$, we have a particular $\tau_a : [M] \to [N]$ that gives $\Delta^a(S,T) = \sum_{i=2}^M \left|a^\top \left(x_i - y_{\tau_a(i)}\right)\right|$. Also, let $\tau^* : [M] \to [N]$ be the "optimal coupling" such that $\Delta(S,T) = \sum_{i=1}^M \|x_i - y_{\tau^*(i)}\|$. Now, we can write:

$$\begin{aligned}
&\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{S}^{d-1})} \Delta^a(S,T) \\
=&\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{S}^{d-1})} \left[\sum_{i=1}^M \left|a^\top (x_i - y_{\tau_a(i)})\right|\right] \\
\geq&\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{S}^{d-1})} \left[\sum_{i=1}^M \left|a^\top (x_i - y_{\tau_a(i)})\right| \middle| a \in A\right] \mathbb{P}_a\{a \in A\} \\
\geq&\frac{\delta(N,M)}{2} \mathbb{E}_a \left[\sum_{i=1}^M \|x_i - y_{\tau_a(i)}\|_2\right] \\
\geq&\frac{\delta(N,M)}{2} \mathbb{E}_a \left[\sum_{i=1}^M \|x_i - y_{\tau^*(i)}\|_2\right], \text{as } \tau^* \text{ is optimal coupling} \\
=&\frac{\delta(N,M)}{2} \left(\sum_{i=1}^M \|x_i - y_{\tau^*(i)}\|_2\right) = \frac{\delta(N,M)}{2} \Delta(S,T)
\end{aligned}$$

Hence, F is Hölder separable with constant $\frac{C'\delta(N,M)}{2}$ and exponent $\lambda = 2$. □

Theorem 13 (Probability bounds on separation)**.** *Let $V \subseteq \mathbb{R}^d$ and σ be as in Theorem 12, and let $A \in \mathbb{R}^{m\times d}, b \in \mathbb{R}^m, c \in \mathbb{R}^m$ whose m columns (respectively entries) are drawn independently from the distribution on a_j, b_j, c_j described in Theorem 12, and consider the function*

$$F(S; A, b, c) = \sum_{x\in S} \sigma\left(c^{-1} \odot (Ax + b)\right). \tag{6}$$

Then there exists $C > 0$, so that for all $S \not\subseteq T$, $\mathbb{P}\left(F(S) \leq F(T)\right) \leq \left(1 - C\Delta^2(S,T)\right)^m$.

Proof. Denote by f the scalar-valued set function for a co-ordinate of F; we proved it's a lower Holder function. Let $S \not\subseteq T$, then by lower continuity we know that,

$$\mathbb{E}_{w\sim\mu(\cdot)}\|[f(S;w) - f(T;w)]_+\|_1 \geq C \cdot \Delta(S,T)^2, \text{ for all } S, T \in \mathcal{P}_{\leq k}(V) \tag{13}$$

We know that

$$\begin{aligned} c \cdot \Delta(S,T)^2 \leq& \mathbb{E}_{w\sim\mu(\cdot)}\|[f(S;w) - f(T;w)]_+\|_1 \\ \leq& \mathbb{P}(f(S;w) \geq f(T;w)) \cdot \sup_{w\in\mathbb{W}} |f(S;w) \geq f(T;w)| \end{aligned}$$

Now, note that as f is a bounded function, denote by M the supremum of f over all sets S and weights $w \in \mathbb{W}$. Thus by the triangle equality,

$$c \cdot \Delta(S,T)^2 \leq 2M \cdot \mathbb{P}(f(S;w) \geq f(T;w))$$

Thus, we obtain that

$$\frac{c \cdot \Delta(S,T)^2}{2 \cdot M} \leq \mathbb{P}(f(S;w) \geq f(T;w))$$

Thus, the complement satisfies:

$$\mathbb{P}(f(S;w) < f(T;w)) \leq 1 - \frac{c \cdot \Delta(S,T)^2}{2 \cdot M}$$

Now, taking m independeing copies of f, namely F, we have that

$$\mathbb{P}(\forall i \in [m], F(S;w)_i < F(T;w)_i) \leq (1 - \frac{c \cdot \Delta(S,T)^2}{2 \cdot M})^m$$

Denoting by $C := \frac{c}{2M}$ yield the desired:

$$\mathbb{P}(F(S;w) < F(T;w)) \leq (1 - C \cdot \Delta(S,T)^2)^m$$

□

Lower Hölder separability of ReLU **networks** So far, we have only given results about Lower Hölder separability of MAS functions with Hat activations. However, as seen in Proposition 10, we have seen that a two layer linear network with ReLU activation is weakly MAS, and as observed from the proof technique, such a 2 layer ReLU net can "simulate" a hat activation. In our arguments, we have shown Lower Hölder separability of Hat Activation based parametric set functions by finding a separating parameter point $w \in \mathcal{W}$ for a given pair of sets (S, T), and constructing an open set of non-zero measure, $\mathcal{W}' \subseteq \mathcal{W}$ around w. We lower bounded the expectation of the non-negative Random variable in the definition of Lower Hölder separability(in Equation (5)) by restricting the expectation only to $\mathcal{W}'$ and showing that it is proportional to $\Delta(S,T)^\lambda$. One can extend this exact same idea to construct the open set of parameters for the two later ReLU networks from the separable parameter point that we get from proof of Proposition 10.

Also, as obtained from Proposition 6, we see that one layer ReLU networks are not even weakly MAS. But a key key step in that proof uses that one can find 3 collinear elements in the ground set V and construct sets with them to obtain the necessary counter-example. Thus, a natural question is to ask: whether shallow 1-layer networks with ReLU activation are also weakly MAS with some additional assumptions on the domain. We show in the following result that, with the additional assumption of V being a hypersphere, we may guarantee Lower Hölder separability for even shallow ReLU networks.

Theorem 19. *Given a ground set $V \subseteq \mathcal{S}^{d-1}$. We define $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}$ as $F(S) = \sum_{x\in S} \mathrm{ReLU}\left(a^\top x + b\right)$, where $a \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ and $b \sim \mathrm{Unif}(-1, 1)$. Then, for any $S, T \in \mathcal{P}_{<\infty}(V)$, we have: $\mathbb{E}_{a,b}\left[F(S) - F(T)\right]_+ \geq C(k)\Delta(S,T)^{(d+3)2^k}$ for some constant $C(k) > 0$ depending only on k.*

Proof. Let $S = \{x_1, \cdots, x_M\}$ and $T = \{y_1, \cdots, y_N\}$, where $M \leq N \leq k$. And according to the definition of $\Delta(S,T)$ from Equation (4) we can equivalently write $\Delta(S,T) = \min_{\pi \in \mathcal{S}_N} \sum_{i=1}^{M} \|x_i - y_{\pi(i)}\|_2$. Let π^* be the optimal solution of the above problem, then we can define the map $\tau : [M] \to [N]$ as follows: $\tau(i) = \pi^*(i), \forall i \in [M]$. Essentially, x_i and $y_{\tau(i)}$ are coupled in the optimal alignment. Now, for any given $i \in [M]$, let $\Delta_i := \|x_i - y_{\tau(i)}\|_2$. Then, consider the open Euclidean ball $\mathcal{B}(x_i, \Delta_i)$. Then if $j \in [N]$ such that $y_j \in \mathcal{B}(x_i, \Delta_i)$ then, $j = \tau(\ell)$ for some $\ell \in [M]$. Otherwise switching from $i \mapsto \tau(i)$ to $i \mapsto j$ reduces the cost, which violates optimal alignment. Let $\Omega_i := \left\{\ell \in [M] : y_{\tau(\ell)} \in \mathcal{B}(x_i, \Delta_i)\right\}$. Here, we make the following claim:

Claim 20. *$\exists \alpha \in [M]$ such that $\sum_{j \in \Omega_i} \|y_j - x_{\tau^{-1}(j)}\|_2 \leq \frac{\Delta_\alpha^2}{16}$ and $\Delta_\alpha \geq \frac{16\Delta^{2^k}(S,T)}{(16k)^{2^k}}$*

Proof. Suppose not. WLOG, assume the following order $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_M$. We build the following directed graph G on $[M]$. Start with the node $v_0 = 1$ and add one node at a time. At step t, we select $v_t := \arg\max_{v \in \Omega_{v_{t-1}}} \Delta_v$ and add the edge (v_{t-1}, v_t). By the assumption, $\frac{\Delta_{v_t}}{16k} \geq (\frac{\Delta_{v_{t-1}}}{16k})^2 \geq \cdots \geq (\frac{\Delta_{v_0}}{(16k)})^{2^k}$. Since we work with finite M, either G is a DAG and this process terminates or G has a cycle. In former case at termination step T we cannot add any more nodes, thus $\Omega_T = \emptyset$, and $T \leq k$. So, $\Delta_T \geq \frac{16k}{k} \frac{\Delta^{2^T}(S,T)}{(16k)^{2^T}} \geq \frac{16\Delta^{2^k}(S,T)}{(16k)^{2^k}}$ and we're done. Otherwise, consider the smallest cycle in G. Let this cycle of length p be $\mathcal{C} = \{v_\beta \to v_{\beta+1} \to \cdots \to v_{\beta+p-1} \to v_\beta\}$. Then, we have $\|x_{v_\beta+i} - y_{\tau(v_\beta+i+1)}\|_2 < \|x_{v_\beta+i} - y_{\tau(v_\beta+i)}\|_2, \forall i \in \{0, \cdots, p-1\}$ and addition of sub-indices is modulo p. Thus, we have the following: $\|x_{v_\beta} - y_{\tau(v_\beta+1)}\|_2 + \|x_{v_\beta+1} - y_{\tau(v_\beta+2)}\|_2 + \cdots + \|x_{v_\beta+p-1} - y_{\tau(v_\beta)}\|_2 < \sum_{j=0}^{p-1} \|x_{v_\beta+i} - y_{\tau(v_\beta+j)}\|_2$. Hence, switching from the coupling $i \mapsto \tau(i)$ to $i \mapsto \tau(i+1)$ for $i \in \{v_\beta, \cdots, v_{\beta+p-2}\}$ and to $v_\beta + p - 1 \mapsto v_\beta$ strictly reduces the cost, but we started with the optimal coupling. This gives a contradiction. □

Consider α as in the above claim. Then we define the region: $\mathcal{A} := \left\{a \in \mathcal{S}^{d-1} : \|a - x_\alpha\| < \frac{\Delta_\alpha}{4}\right\}$. Then, $\forall a \in \mathcal{A}, \|a - x_\alpha\|_2 < \|a - y_{\tau(\alpha)}\|_2$. Thus, $a^\top x_\alpha > a^\top y_{\tau(\alpha)}$. Also, consider the region $\mathcal{B}_a := \left\{b \in \mathbb{R} : -1 + \frac{\Delta_\alpha^2}{16} + \frac{\Delta_\alpha^2}{64} + \frac{\|a-x\|^2}{2} \leq b \leq -1 + \frac{\Delta_\alpha^2}{8}\right\}$. For any given $a \in \mathcal{A}$, we have: $\frac{\|a-x\|^2}{2} \leq \frac{\Delta_\alpha^2}{32}$, thus $-1 + \frac{\Delta_\alpha^2}{16} + \frac{\Delta_\alpha^2}{64} + \frac{\|a-x\|^2}{2} \leq -1 + \frac{7}{64}\Delta_\alpha^2 \leq -1 + \frac{\Delta_\alpha^2}{8}$, so $\mathcal{B}_a \neq \emptyset$. Also note that, $\forall a \in \mathcal{A}, b \in \mathcal{B}_a$, we have:

$$
\begin{aligned}
[F(S) - F(T)]_+ &\geq \left[\mathrm{ReLU}\left(a^\top x_\alpha + b\right) - \left(\sum_{\beta \in \Omega_\alpha} \mathrm{ReLU}\left(a^\top y_{\tau(\beta)}\right) - \mathrm{ReLU}\left(a^\top x_\beta\right)\right)\right]_+ \\
&\geq \left[\mathrm{ReLU}\left(a^\top x_\alpha + b\right) - \sum_{\beta \in \Omega_\alpha} \left|\mathrm{ReLU}\left(a^\top y_{\tau(\beta)}\right) - \mathrm{ReLU}\left(a^\top x_\beta\right)\right|\right]_+ \\
&\geq \left[\mathrm{ReLU}\left(a^\top x_\alpha + b\right) - \sum_{\beta \in \Omega_\alpha} \left|a^\top (x - y)\right|\right]_+ \\
&\geq \left[\mathrm{ReLU}\left(a^\top x_\alpha + b\right) - \sum_{\beta \in \Omega_\alpha} \|x_\beta - y_{\tau(\beta)}\|\right]_+ \\
&\geq \left[a^\top x_\alpha + b - \frac{\Delta_\alpha^2}{16}\right]_+ = \left[1 + b - \frac{\|a-x\|^2}{2} - \frac{\Delta_\alpha^2}{16}\right]_+ \geq \frac{\Delta_\alpha^2}{64}
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
\mathbb{E}_{a,b}\left[F(S) - F(T)\right]_+ &\geq \mathbb{E}_{a,b}\left[(F(S) - F(T))\mathbf{1}_{\{a \in \mathcal{A}, b \in \mathcal{B}_a\}}\right]_+ \\
&\geq \frac{\Delta_\alpha^2}{64} \mathbb{P}\left\{a \in \mathcal{A}, b \in \mathcal{B}_a\right\} \\
&= \frac{\Delta_\alpha^2}{64 S_{d-1}} \int_{a \in \mathcal{S}^{d-1} : a_1 \geq 1 - \frac{\Delta_\alpha^2}{32}} \frac{1}{2}\left(\frac{3}{64}\Delta_\alpha^2 - (a_1 - 1)\right) da
\end{aligned}
$$

$$\geq \frac{\Delta_\alpha^4}{2^{13} S_{d-1}} \int_{a \in \mathcal{S}^{d-1}: a_1 \geq 1 - \frac{\Delta_\alpha^2}{32}} da$$
$$= \frac{\Delta_\alpha^4}{2^{13}} \int_{1-\frac{\Delta_\alpha^2}{32}}^{1} (1-x^2)^{\frac{d-3}{2}} dx \geq \frac{\Delta_\alpha^6}{2^{18}} \left(\frac{\Delta_\alpha^2}{32}\right)^{\frac{d-3}{2}} \left(2 - \frac{\Delta_\alpha^2}{32}\right)^{\frac{d-3}{2}}$$
$$\geq \frac{\Delta_\alpha^{d+3}}{2^{23}}$$

Now, using the bounds from 20, we get that, $\Delta_\alpha \geq C(k)\Delta^{2^k}(S,T)$. Thus we have that, $\mathbb{E}\left[F(S) - F(T)\right]_+ \geq C_1(k)\Delta(S,T)^{(d+3)2^k}$, and done $\square$

Upper Lipschitz property of weakly MAS functions We now show that, the MAS functions obtained by applying one layer neural network with Hat activation for each element of the set, followed by sum aggregation are also Upper Lipschitz continuous in expectation with respect to the augmented Wasserstein distance, following the framework of Davidson and Dym [27]. This shows stability of such MAS embeddings. We say that a parametric set function $F : \mathcal{P}_{\leq k}(V) \times \mathcal{W} \to \mathbb{R}^m$ is Upper Lipschitz in expectation if $\exists C > 0$ such that:

$$\mathbb{E}_{w \in \mathcal{W}} \|F(S;w) - F(T;w)\|_1 \leq C \cdot \mathbb{W}^{(k)}(S,T),\ \forall S,T \in \mathcal{P}_{\leq k}(V) \tag{14}$$

Where $\mathbb{W}^{(k)}(.,*)$ is the augmented-Wasserstein metric for sets in $\mathcal{P}_{\leq k}(V)$, where a padding z is added for $k - |S|$ times to any multiset S of size less than k. We work with a compact ground set V, and we pick a padding z that has a positive distance from V. For the following proofs, we work with $V \subseteq \mathbb{R}^d$ which is norm bounded by 1, and we choose the padding element $z \in \mathbb{R}^d$ such that $\|z\| \geq 3$. For a general V whose norm is bounded by B, we can scale the padding z by B. With this, we show that, real valued functions $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}$ given by $F(S) = \sum_{x \in S} \sigma\left(\frac{a^\top x - b}{c}\right)$ are Upper Lipschitz w.r.t $\mathbb{W}_k$. Now, if we take independent parameteric copies across m output dimensions and if M_2 is a vector-to-vector Lipschitz function, then functions of the form $\mathrm{M}_2 \circ F$, where F has the form in Equation (6) are Upper Lipschitz as well. This we state in the following theorem:

Theorem 21 (F is upper Lipschitz)**.** *Let σ belongs to the class of Hat activations, and $V \subset \mathbb{R}^d$ be a compact ground set such that $\sup_{v \in V} \|v\| \leq 1$. We consider $F : \mathcal{P}_{\leq k}(V) \to \mathbb{R}, F(S) = \sum_{x \in S} \sigma\left(\frac{a^\top x - b}{c}\right)$, with the distributions $a \sim \mathrm{Unif}(\mathcal{S}^{d-1}), b \sim \mathrm{Unif}(-1,1), c \sim \mathrm{Unif}(0,2)$. Then $\exists$ a constant $C > 0$, such that for any $S,T \in \mathcal{P}_{\leq k}(V)$:*

$$\mathbb{E}_{a,b,c} |F(S) - F(T)| \leq C \cdot \mathbb{W}^{(k)}(S,T)$$

Proof. Consider $S = \{x_1, \cdots, x_M\}, T = \{y_1, \cdots, y_N\}$. Since $|F(S) - F(T)|$ is symmetric in S,T, we consider WLOG that $M \leq N$. As in previous sections, we use τ to denote an injective function from $[M] \to [N]$, and we use $\Omega_{N,M}$ to denote the set of all injective functions from $[M] \to [N]$. Thus, we can write the augmented Wasserstein distance on $\mathcal{P}_{\leq k}(V)$ with padding z as:

$$\mathbb{W}^{(k)}(S,T) = \arg \min_{\tau \in \Omega_{N,M}} \left\{ \sum_{i=1}^{M} \|x_i - y_{\tau(i)}\| + \sum_{j \neq \tau(i)} \|y_j - z\| \right\}$$

Let τ^* be the argmin of the above expression that gives the $\mathbb{W}_k(S,T)$ Now, we can write:

$$\mathbb{E}_{a,b,c} |F(S) - F(T)| = \mathbb{E}_{a,b,c} \left| \sum_{i=1}^{M} \sigma\big(\frac{a^\top x_i - b}{c}\big) - \sigma\big(\frac{a^\top y_{\tau^*(i)} - b}{c}\big) - \sum_{j \neq \tau^*(i)} \sigma\big(\frac{a^\top y_j - b}{c}\big) \right|$$
$$\leq \sum_{i=1}^{M} \mathbb{E}_{a,b,c} \left| \sigma\left(\frac{a^\top x_i - b}{c}\right) - \sigma\left(\frac{a^\top y_{\tau^*(i)} - b}{c}\right) \right| + \sum_{j \neq \tau^*(i)} \mathbb{E}_{a,b,c} \left| \sigma\left(\frac{a^\top y_j - b}{c}\right) \right|$$

We now separately analyze coupled and un-coupled terms: i.e terms of the form $\mathbb{E}\left|\sigma\left(\frac{a^\top y_j - b}{c}\right)\right|$ and $\mathbb{E}\left|\sigma\big(\frac{a^\top x_i - b}{c}\big) - \sigma\big(\frac{a^\top y_{\tau^*(i)} - b}{c}\big)\right|$ in the following analysis. For the proof that follows, we consider $\mathrm{supp}(\sigma) \subseteq [0,1]$. For support in $[\gamma_1, \gamma_2]$, the proof follows similarly by sifting the distributions of b, c by γ_1 and scaling by $\gamma_2 - \gamma_1$.

Computing $\mathbb{E}\left|\sigma\left(\frac{\mathbf{a}^\top \mathbf{y_j} - \mathbf{b}}{\mathbf{c}}\right)\right|$:

Note that, $\sigma(\frac{a^\top y_j - b}{c}) = 0, \forall b \geq a^\top y_j$. Also, for a given $b \in (-1, a^\top y_j)$ we have that:

$$\forall c \in (0, a^\top y_j - b), \frac{a^\top y_j - b}{c} \geq 1 \implies \sigma\left(\frac{a^\top y_j - b}{c}\right) = 0$$

Thus, conditioned on a given $a \in \mathcal{S}^{d-1}$, we can write the conditional expectation over b, c as follows:

$$\begin{aligned}
&\mathbb{E}_{b,c}\left\{\sigma\big(\frac{a^\top y_j - b}{c}\big)\middle| a\right\} \\
=&\frac{1}{4}\int_{b=-1}^{a^\top y_j}\int_{c=a^\top y_j - b}^{2}\left|\sigma\left(\frac{a^\top y_j - b}{c}\right) - \sigma(0)\right| \mathrm{d}c\,\mathrm{d}b, \text{ using } \sigma(0) = 0 \\
\leq&\frac{\kappa_2}{4}\int_{b=-1}^{a^\top y_j}\int_{c=a^\top y_j - b}^{2}\left(\frac{a^\top y_j - b}{c}\right)\mathrm{d}c\,\mathrm{d}b, \text{as } \sigma \text{ is a Hat activation, it's Lipschitz} \\
=&\frac{\kappa_2}{4}\int_{b=-1}^{a^\top y_j}(a^\top y_j - b)\ln\left(\frac{2}{a^\top y_j - b}\right)\mathrm{d}b = \frac{\kappa_2}{4}\int_{a^\top y_j + 1}^{0} z\ln\left(\frac{2}{z}\right)(-\mathrm{d}z), z := a^\top y_j - b \\
=&\frac{\kappa_2}{4}\int_{0}^{a^\top y_j + 1} z\ln\left(\frac{2}{z}\right)\mathrm{d}z = \frac{\kappa_2}{4}\cdot\frac{a^\top y_j + 1}{2}\left(\frac{a^\top y_j + 1}{2} + \big(a^\top y_j + 1\big)\ln\left(\frac{2}{a^\top y_j + 1}\right)\right) \\
\leq&\frac{\kappa_2}{4}\cdot(a^\top y_j + a^\top a) = \frac{\kappa_2}{4}\left|a^\top(y_j + a)\right| \leq \frac{\kappa_2}{4}\|y_j + a\| \leq \frac{\kappa_2}{4}\|y_j - z\|
\end{aligned}$$

Now, taking an expectation over $a \sim \mathrm{Unif}(\mathcal{S}^{d-1})$, the inequality is preserved and we get that, for "isolated " y_j's, $\mathbb{E}\sigma\left(\frac{a^\top - b}{c}\right) \leq \frac{\kappa_2}{4}\|y_j - z\|$

Thus for "isolated" elements from the larger set that get coupled with the padded element z while computing augmented-$\mathbb{W}^{(k)}$, we have shown that they are upper bounded by their diatcne from the padding z. Hence, it remains to show a similar Lipschitz upper bound for the elements $x_1, \cdots x_m \in S_1$ which have a corresponding $y_{i(1)}, \cdots, y_{i(m)} \in S_2$ in the optimal coupling

Computing $\mathbb{E}\left|\sigma(\frac{\mathbf{a}^\top \mathbf{x_i} - \mathbf{b}}{\mathbf{c}}) - \sigma(\frac{\mathbf{a}^\top \mathbf{y}_{\tau^*(\mathbf{i})} - \mathbf{b}}{\mathbf{c}})\right|$:

Now, we have, for any $i \in [M]$, and a given $a \in \mathcal{S}^{d-1}$ the conditional expectation $\mathbb{E}_{b,c}\left|\sigma(\frac{a^\top x_i - b}{c}) - \sigma(\frac{a^\top y_{\tau^*(i)} - b}{c})\right|$ is symmetric in x_i and $y_{\tau^*(i)}$. Thus, WLOG we can assume that, $a^\top x_i \geq a^\top y_{\tau^*(i)}$. Note that we then have:

$$\begin{aligned}
&\left|\sigma(\frac{a^\top x_i - b}{c}) - \sigma(\frac{a^\top y_{\tau^*(i)} - b}{c})\right| \\
=&\begin{cases}
\left|\sigma(\frac{a^\top x_i - b}{c})\right|, \forall b \in (a^\top y_{\tau^*(i)}, a^\top x_i), \forall c \in (a^\top x_i - b, 2) \\
\left|\sigma(\frac{a^\top x_i - b}{c}) - \sigma(\frac{a^\top y_{\tau^*(i)} - b}{c})\right|, \forall b \in (-1, a^\top y_{\tau^*(i)}), \forall c \in (a^\top y_{\tau^*(i)} - b, 2) \\
0, \text{ otherwise}
\end{cases}
\end{aligned}$$

Hence, conditioned on a specific $a \in \mathcal{S}^{d-1}$, we can write the conditional expectation over b, c as:

$$\begin{aligned}
&\mathbb{E}_{b,c}\left|\sigma(\frac{a^\top x_i - b}{c}) - \sigma(\frac{a^\top y_{\tau^*(i)} - b}{c})\right| = \int_{b=a^\top y_{\tau^*(i)}}^{a^\top x_i}\int_{c=a^\top x_i - b}^{2}\sigma\left(\frac{a^\top x_i - b}{c}\right)\mathrm{d}c\,\mathrm{d}b \\
&+\int_{b=-1}^{a^\top y_{\tau^*(i)}}\int_{c=a^\top y_{\tau^*(i)} - b}^{2}\left|\sigma(\frac{a^\top x_i - b}{c}) - \sigma(\frac{a^\top y_{\tau^*(i)} - b}{c})\right|\mathrm{d}c\,\mathrm{d}b \\
\leq&\frac{\kappa_2}{4}\int_{b=a^\top y_{\tau^*(i)}}^{a^\top x_i}(a^\top x_i - b)\ln\left(\frac{2}{a^\top x_i - b}\right)\mathrm{d}b
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\kappa_2}{4}\left|a^\top(x_i-y_{\tau^*(i)})\right|\int_{b=-1}^{a^\top y_{\tau^*(i)}}\ln\left(\frac{2}{a^\top y_{\tau^*(i)}-b}\right)\,\mathrm{d}b\\
\leq&\frac{\kappa_2}{4}\left(\int_0^{a^\top(x_i-y_{\tau^*(i)})} z\ln\left(\frac{2}{z}\right)\,\mathrm{d}z+\left|a^\top(x_i-y_{\tau^*(i)})\right|\int_0^{a^\top y_{\tau^*(i)}+1}\ln\left(\frac{2}{z}\right)\,\mathrm{d}z\right)\\
\leq&\frac{\kappa_2}{4}\left(\left|a^\top(x_i-y_{\tau^*(i)})\right|+2\left|a^\top(x_i-y_{\tau^*(i)})\right|\right)<\kappa_2\left|a^\top(x_i-y_{\tau^*(i)})\right|\\
\leq&\kappa_2\|x_i-y_{\tau^*(i)}\|,\ \text{by Cauchy Schwarz}
\end{aligned}
$$

Thus taking one more expectation over $a\sim\mathcal{S}^{d-1}$, and summing over all $i\in[M]$, we get:

$$
\mathbb{E}_{a,b,c}\sum_{i=1}^{M}\left|\sigma(\frac{a^\top x_i-b}{c})-\sigma(\frac{a^\top y_{\tau^*(i)}-b}{c})\right|\leq\kappa_2\sum_{i=1}^{M}\|x_i-y_{\tau^*(i)}\|
$$

Thus, combining our computations for "isolated" y_j's and "coupled" $(x_i,y_{\tau^*(i)})$'s we get that:

$$
\mathbb{E}_{a,b,c}|F(S)-F(T)|\leq\frac{\kappa_2}{4}\sum_{j\neq\tau^*(i)}\|y_j-z\|+\kappa_2\sum_{i=1}^{M}\|x_i-y_{\tau^*(i)}\|\leq\kappa_2\mathbb{W}^{(k)}(S,T)
$$

And our proof of F being upper Lipschitz is complete. □

8.4 Universal approximators with MAS functions

Theorem 14 (Universality)**.** *Let V be a finite ground state, and let $F:\mathcal{P}_{\leq k}(V)\to\mathbb{R}^m$ be a MAS function. Then for every multiset-to-vector monotone function $f:\mathcal{P}_{\leq k}(V)\to\mathbb{R}^s$, there exists a vector-to-vector monotone function $M:\mathbb{R}^m\to\mathbb{R}^s$ such that $F(S)=M\circ f(S)$.*

Proof. A MAS function is in particular invertible, and therefore we can write $f(S)=f\circ F^{-1}\circ F(S)$. On the image of F we can define the function $M=f\circ F^{-1}$ which is a vector-to-vector functions, and it is monotone on the image of F, because if $v=F(S),u=F(T)$ for some sets S,T, and if $v\leq u$, then by separability $S\subseteq T$, and therefore

$$
M(v)=f\circ F^{-1}(v)=f(S)\leq f(T)=f\circ F^{-1}(u)=M(u)
$$

Finally, we need to show that M can be extended to a montone function on all of $\mathbb{R}^m$. we accomplish this by defining

$$
M(v)=\max\{M(u)|\quad u\leq v,u\in\mathrm{Image}(F)\}.
$$

□

9 Details about our model

9.1 Architecture details:

In all 4 of MASNET models, we use an elementwise neural network $NN_\theta:\mathbb{R}^d\to\mathbb{R}^m$, where the ground set $V\subseteq\mathbb{R}^d$ and the output in in $\mathbb{R}^m$. We follow that with a sum aggregation, followed by a monotone vector-to-vector neural netwrok $M_{2,\phi}:\mathbb{R}^m\to\mathbb{R}^m$. For the embedding Neural Network, we use a 1 hidden layer NN with ReLU activation in the hidden layers and ReLU or Hat activation in the output, depending on whether we're using MASNET-ReLU or Hat activation based MASNET. To make $M_{2,\phi}$ monotone, we use non-negative weights by taking the absolute value of parameters before applying the linear transformation, and use all monotonic activations like ReLU in the intermediate layers. Now, as discussed before, we give details of our model MASNET-INT.

9.2 MASNET-INT

In this appendix we design universal approximators for the class of Hat functions to learn σ. We parametrize For that, we derive equivalent conditions on the derivatives of σ (per embedding dimension) and use the techniques of learning functions by modelling the derivatives using neural networks and then using a numerical integration as in Wehenkel and Louppe [35].

Lemma 22. *If $\sigma:\mathbb{R}\to\mathbb{R}_{\geq 0}$ belongs to the hat activation class , then σ' satisfies the following conditions for some positive constants $c,C>0$ and $\alpha\in\mathbb{R},\beta>0,\gamma\in(0,1)$:*

1. *supp* $(\sigma') \subseteq [\alpha, \alpha+\beta]$
2. $\int_{\alpha}^{\alpha+\beta} \sigma'(x)\,\mathrm{d}x = 0$
3. $\sigma'(x) \leq C, \forall x \in \mathbb{R}$
4. $\sigma'(x) \geq c, \forall x \in (\alpha, \alpha+\gamma\cdot\beta)$

The above 4 conditions, along with $\sigma(\alpha)=0$ are the necessary-sufficienet conditions that characterizes the hat activation class.

Proof. We first prove that, if σ belongs to the hat activation class as defined in 8, then conditions 1-4 are satisfied. Note that, if σ is a Hat activation, then σ has compact support in some $[\alpha, \alpha+\beta]$ by deifnition, and it's piecewise continuously differentiable. Thus, outside $[\alpha, \alpha+\beta]$, we have $\sigma \equiv 0 \implies \sigma' \equiv 0$. Hence, supp$(\sigma') \subseteq [\alpha, \alpha+\beta]$. We also have that, $0 = \sigma(\alpha) = \sigma(\alpha+\beta) \implies \sigma(\alpha+\beta) - \sigma(\alpha) = 0 \implies \int_{\alpha}^{\alpha+\beta} \sigma'(t)\,\mathrm{d}t = 0$. Finally, σ being Lipschitz is equivalent to $|\sigma'(t)| \leq C$ for some $C > 0$ and by continuity of σ' in $(\alpha, \alpha+\beta)$ we must have that $2c := \lim_{t\to\alpha^+} \sigma'(t) > 0 \implies \sigma'(t) \geq c, \forall t \in (\alpha, \alpha+\gamma\cdot\beta)$ for some $\gamma \in (0,1)$. Thus, conditions 1-4 are implied by σ belonging to the Hat activation class.

Moreover, if conditions 1-4 are satisfied, and $\sigma(\alpha)=0$, then condition-2 implies $\sigma(\alpha+\beta)=0$. This, along with condition-1 implies that $\sigma(t)=0, \forall t \notin [\alpha, \alpha+\beta]$. Condition-4 immediately implies that $\sigma(\alpha+\gamma\cdot\beta) \geq c\cdot\gamma\beta > 0$, thus $\sigma \not\equiv 0$. On the other hand, we get that $\sigma' \leq C$ implies by LMVT that, $|\sigma(x)-\sigma(y)| \leq C\cdot|x-y|$. Hence, σ belongs to the class of Hat activations following the definition from 8. □

Neural Parmetrization of MASNET-INT: Consider $h^1_\theta(.), h^2_\phi(.)$ to be non-negative + bounded and bounded fully connected Neural Networks respectively of one hidden layer each. Also consider the trainable support parameters to be $\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^m_+, \boldsymbol{\gamma} \in (0,1)^m$ where m is the output dimension and $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are the support parameters, aggregated for all output dimensions. Now, we define an integral based model of parametric hat functions. Let $\Theta = (\theta, \phi, \alpha, \beta, m)$ be the parameter space.

$$\sigma_\Theta(x) = \int_{\alpha}^{x} h^1_\theta(z)\mathbf{1}_{\{\alpha\leq z\leq\alpha+\gamma\beta\}}\,\mathrm{d}z - \left[\frac{\int_{\alpha}^{\alpha+\gamma\beta} h^1_\theta(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h^2_\phi(z)dz}\right]\left(\int_{\alpha+\gamma\beta}^{x} h^2_\phi(z)\mathbf{1}_{\{\alpha+\gamma\beta\leq z\leq\alpha+\beta\}}dz\right) \tag{15}$$

In the above formulation, in addition to the support parameters we also learn the function itself rather than performing a linear interpolation, this makes the above class a universal approximator of hat functions.

Lemma 23 (Universal Hat approximator)**.** *The family σ_Θ defined in Equation* (15) *is an universal approximator of the class of Hat functions having support in $[\alpha, \alpha+\beta]$ with $\sigma'(t) > 0, \forall t \in (\alpha, \alpha+\gamma\beta)$*

Proof. We show that the given parametric model is an universal approximator of Hat functions using their equivalent formulation in Lemma 22. Let σ_1 be σ restricted to the interval $[\alpha, \alpha+\gamma\beta]$ and σ_2 be σ bestricted to the interval $[\alpha+\gamma\beta, \alpha+\beta]$. Using Theorem-1 from Lu et al. [36], we can get ReLU neural network $h^1_\theta : \mathbb{R} \to \mathbb{R}$ and $h^2_\phi : \mathbb{R} \to \mathbb{R}$ of width 5 each such that approximates the derivatives of the restricted functions σ_1' and σ_2'. Thus, the following hold for any $\epsilon > 0$ and any constant $c \in \mathbb{R}$:

$$\int_{\alpha}^{\alpha+\gamma\beta} \left|\sigma_1'(z) - h^1_\theta(z)\right|\,\mathrm{d}z \leq \epsilon \text{ and } \int_{\alpha+\gamma\beta}^{\alpha+\beta} \left|\sigma_2'(t) + c\cdot h^2_\phi(t)\right|\,\mathrm{d}t \leq \epsilon \tag{16}$$

We now observe that, if $x \in (-\infty, \alpha)$, both the indicator functions: $\mathbf{1}_{\{\alpha\leq z\leq\alpha+\gamma\beta\}}$ and $\mathbf{1}_{\{\alpha+\gamma\beta\leq z\leq\alpha+\beta\}}$ as in the integrands of Equation (15) evaluate to 0, thus σ_Θ exactly coincides with σ. On the other hand, if $x \in (\alpha+\beta, \infty)$ then we would have the first integral of 15 evaluate to $\int_{\alpha}^{\alpha+\gamma\beta} h^1_\theta(z)\,\mathrm{d}z$ (due to presence of the indicator $\mathbf{1}_{\{\alpha\leq z\leq\alpha+\gamma\beta\}}$ and the second expression evaluates to $-\left[\frac{\int_{\alpha}^{\alpha+\gamma\beta} h^1_\theta(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h^2_\phi(z)dz}\right]\left(\int_{\alpha+\gamma\beta}^{\alpha+\beta} h^2_\phi(z)dz\right) = -\int_{\alpha}^{\alpha+\gamma\beta} h^1_\theta(z)dz$. Thus, both the expressions sum to 0 and σ_Θ coincides with σ for $x \in (\alpha+\beta, \infty)$ as well.

Now we shall show that, for any $x \in (\alpha, \alpha+\beta)$ we must also have $|\sigma_\Theta(x) - \sigma(x)| \leq \epsilon$, implying convergence in $\sup$ norm. We consider the following cases:

- **Case-I:** When $x \in (\alpha, \alpha + \gamma\beta)$, we have that:

$$\begin{aligned}
&|\sigma(x) - \sigma_\Theta(x)| \\
=&\left| \int_\alpha^x \sigma_1'(z)\,\mathrm{d}z - \int_\alpha^x h_\theta^1(z)\,\mathrm{d}z \right| \\
\leq& \int_\alpha^x \left|\sigma_1'(z) - h_\theta^1(z)\right|\,\mathrm{d}z, \text{by traingle ineq.} \\
\leq& \int_\alpha^{\alpha+\beta\gamma} \left|\sigma_1'(z) - h_\theta^1(z)\right|\,\mathrm{d}z, \text{by non-negativity of integrand} \\
\leq& \epsilon, \text{by Equation (16)}
\end{aligned}$$

- **Case-II:** When $x \in (\alpha + \gamma\beta, \alpha + \beta)$ we have:

$$\begin{aligned}
&|\sigma(x) - \sigma_\Theta(x)| \\
=&\left| \int_\alpha^x \sigma'(z)\,\mathrm{d}z - \int_\alpha^{\alpha+\gamma\beta} h_\theta^1(z)\,\mathrm{d}z - \int_{\alpha+\gamma\beta}^x h_\phi^2(z)\,\mathrm{d}z \right| \\
=&\left| \int_\alpha^{\alpha+\gamma\beta} \left(\sigma_1'(z) - h_\theta^1(z)\right)\,\mathrm{d}z + \int_{\alpha+\gamma\beta}^x \left(\sigma_2'(z) + \left[\frac{\int_\alpha^{\alpha+\gamma\beta} h_\theta^1(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h_\phi^2(z)dz}\right] h_\phi^2(z)\right)\,\mathrm{d}z \right| \\
\leq& \int_\alpha^{\alpha+\gamma\beta} \left|\sigma_1'(z) - h_\theta^1(z)\right|\,\mathrm{d}z + \int_{\alpha+\gamma\beta}^x \left|\sigma_2'(z) + \left[\frac{\int_\alpha^{\alpha+\gamma\beta} h_\theta^1(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h_\phi^2(z)\,\mathrm{d}z}\right] h_\phi^2(z)\right|\,\mathrm{d}z, \text{tri. ineq.} \\
\leq& \epsilon + \int_{\alpha+\gamma\beta}^{\alpha+\beta} \left|\sigma_2'(z) + \left[\frac{\int_\alpha^{\alpha+\gamma\beta} h_\theta^1(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h_\phi^2(z)\,\mathrm{d}z}\right] h_\phi^2(z)\right|\,\mathrm{d}z, \text{by 16 on 1st term \& tri. ineq. on 2nd} \\
\leq& 2\epsilon
\end{aligned}$$

  where last inequality is obtained $c = \frac{\int_\alpha^{\alpha+\gamma\beta} h_\theta^1(z)dz}{\int_{\alpha+\gamma\beta}^{\alpha+\beta} h_\phi^2(z)\,\mathrm{d}z}$ in the universal approximator for σ_2' in Equation (16). Thus, just we see that, for any $\epsilon > 0$ and $\alpha \in \mathbb{R}, \beta > 0, \gamma \in (0, 1)$ we can find a universal approximator for the Hat function class

□

Parametrizing α, β, γ We have provided an universal approximator for the Hat function class given α, β, γ but we'd also want to have α, β, γ as learnable parameters. This is valid for both MASNET-Hat and MASNET-INT, as in both cases we seek to learn the support parameters. Now, there is no constraint on α, so we can directly initialize $\boldsymbol{\alpha} \in \mathbb{R}^{3m}$ and learn it through gradient descent. However, we have a constraint on β that $\beta > 0$. For this, we first initialize $\boldsymbol{\beta_0} \in \mathbb{R}^m$ and we obtain $\boldsymbol{\beta} \in \mathbb{R}^m$ by appling co-ordinatewise positive transformation on $\boldsymbol{\beta_0}$. This we do in two ways, depending on which works better on a given task: **(I)** by simply taking $\boldsymbol{\beta_0} \mapsto |\boldsymbol{\beta_0}|$, where the absolute value is taken pointwise. **(II)**by taking the pointwise transformation $\boldsymbol{\beta_0} \mapsto \text{ELU}\left(\boldsymbol{\beta_0}; \upsilon\right) + \upsilon$, where $\text{ELU}\left(\cdot; \upsilon\right)$ is the Exponential Linear Unit with hyperparameter $\upsilon > 0$. Note that, the ELU function is given by: $\text{ELU}\left(x; \upsilon\right) = \begin{cases} x, \text{if } x > 0 \\ \upsilon\left(e^x - 1\right), \text{if } x \leq 0 \end{cases}$. For parametrizing $\gamma \in (0, 1)$, we similarly initialize $\boldsymbol{\gamma_0} \in \mathbb{R}^m$ and then apply a pointwise transfprmation that takes each co-ordinate to $(0, 1)$. We use tempered sigmoid function: $\boldsymbol{\gamma_0} \mapsto \text{Sigmoid}\left(\tau \cdot \boldsymbol{\gamma_0}\right)$ where $\tau > 0$ is a hyperparameter

Designing soft indicator functions: In the integrands for MASNET-INT, we have some indicator functions of the form $\mathbf{1}_{\{a \leq \cdot \leq b\}}(x) = \begin{cases} 1, \text{if } x \in [a, b] \\ 0, \text{o.w} \end{cases}$. We want these indicator functions to have non zero gradients as we intend to learn the "support parameters" a, b of such indicator function. If we chose binary indicators, then that is not differentiable everywhere and have gradient 0 in most places, making it difficult to learn a, b. Thus, we design a "soft" indicator function with the help of tempered sigmoid with hyperparameter τ_S as follows:

$$\mathbf{1}_{\{a \leq \cdot \leq b\}}(x) = \text{Sigmoid}\left(\tau_S \cdot (x - a)\right) \cdot \text{Sigmoid}\left(\tau_S \cdot (b - x)\right)$$

9.3 Making MASNET Mini Batch Consistent

There has been a chain of works to make set-based neural models to be mini batch consistent in [37, 38] which have applications in large scale machine learning. The goal of these Mini Batch Consistent (MBC) works is to design a set function F, which, if applied on any subsets from partitions of a set S and pooled through an activation g, should return $F(S)$. Thus, at a high level, MBC functions analyse and design functions for the following batch consistency condition: $g(F(S_1), \cdots, F(S_n)) = F(S)$. MBC methods achieve this by imposing certain restrictions on attention-based set architectures. Similar to this, our work also puts restrictions on the inner embedding transformation and the outer vector-to-vector transformations. It is thus a natural question to ask, if MASNET can be made minibatch consistent. Since set containment is the major theme of our work, with potential applications in retrieval and recommendation systems, processing huge sets at once can be a key bottleneck, and it would be nice to make MASNET process over mini-batches and then design some aggregation mechanism to make the aggregated embedding equal to the embedding of the entire set.

If we consider a partition of a given set $S = \bigcup_{i=1}^{k} S_i$, where S_i's are all disjoint, then $\text{MASNET}(S_i) \leq \text{MASNET}(S)$, by monotonicity. But, we can try to find an appropriate pooling mechanism over the subset embeddings to make the aggregated embedding same as that of the entire set. For example, if the outer transformation M_{θ_2}, as in Equation (7) was invertible, and then we may define a particular permutation-invariant pooling function g such that, $g(X_1, \cdots, X_k) = M_{\theta_2}\left(\sum_{i=1}^{k} M_{\theta_2}^{-1}(X_i)\right)$, then we'd have: $g\left(\text{MASNET}(S_1), \cdots, \text{MASNET}(S_k)\right) = M_{\theta_2}\left(M_{\theta_2}^{-1}\left(\text{MASNET}(S_i)\right)\right) = M_{\theta_2}\left(\sum_{i=1}^{k}\sum_{x\in S_i}\sigma \circ M_{\theta_1}(x)\right) = M_{\theta_2}\left(\sum_{x\in S}\sigma \circ M_{\theta_1}(x)\right) = \text{MASNET}(S)$. his is irrespective of how we partition S into $S_1, \cdots S_k$. And this would enable us to process the subsets independently and aggregate them accordingly.

For our purposes, we have needed the outer transformation M_{θ_2} to be a vector-to-vector monotone function in Equation (7). But that doesn't guarantee injectivity or invertibility. Thus, we need to restrict the choice of M_{θ_2} to be monotone and invertible for MASNET to be MBC consistent. One possible choice is to use ideas from relevant works like [39, 35] to design monotone and invertible neural networks, and analyse the expressibility properties of the resulting set model. This forms an interesting direction for future work.

10 Additional details about experiments and more experiments

10.1 Hardware details:

All experiments were performed in a compute server, running the OS of GNU Linux Version 12, equipped with a 16 core Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz CPU architecture and equipped with a cluster of 6 NVIDIA RTX A6000 GPUs with a memory of 49GB each.

10.2 Dataset details and more experiments

We now give an account of preparation details of each dataset and accounts on more experiments on each.

Synthetic datasets and related experiments:

We begin with a controlled synthetic setting where we generate pairs (S, T) such that $k_S < k_T$.

Each pair is labeled either positive or negative. For positive examples, we first sample k_T vectors in $\mathbb{R}^d$ from $\mathcal{N}(0, \mathbb{I}_d)$, which form the set T. A subset of size k_S is then uniformly sampled from T to obtain S. For negative examples, we again sample T as above, but then independently draw k_S vectors from $\mathbb{R}^d$ to form S .

To simulate a real world scenerio of information loss, we inject gaussian noise into the vectors of S to get S' but keep the boolean label unchanged.

On this, we provide two more plots: Figs. 8 and 9 that compares the accuracies of MASNET vs baselines, by varying the gap between target set size $|T|$ and query set size $|S|$. Note that the more this gap is, the more it becomes challenging to separate two non-subsets. In all cases, the embedding MLP M_{θ_1} from Equation (7) is comprising of ≥ 2 layers.

Text-based datasets and related experiments:

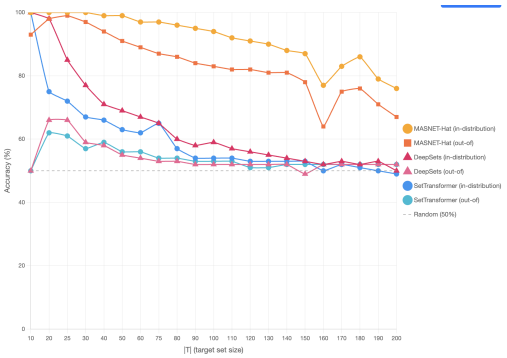

Figure 8: Acc vs $|T|$ for all models $\geq$ 2 layer MLP

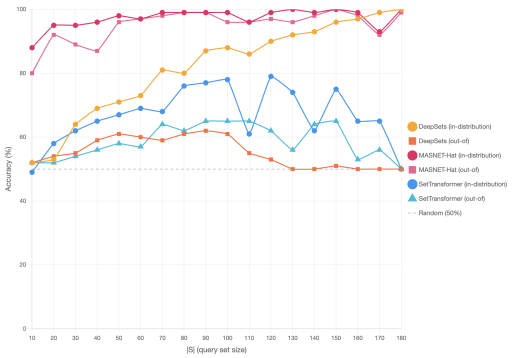

Figure 9: Acc vs $|S|$ for all models $\geq$ 2 layer MLP

We evaluate our models on three text based real-world datasets: MSWEB, MSNBC, and the Amazon Baby Registry datasets, all of which reflect naturally occurring set containment structures coming out of user behavior in websites and recommendation engines.

MSWEB and MSNBC: These datasets consist of user activity logs from *www.microsoft.com* and *www.msnbc.com*, respectively. Each user session is treated as a bag of page identifiers, which are embedded into 768-dimensional vectors using a pre-trained BERT model. We construct query-target pairs (S, T) by sampling S from a session V, and letting $T = V \setminus S$. A pair is labeled 1 if $S \subseteq T$, and 0 otherwise. To simulate noisy real-world conditions, Gaussian noise is added to each element in S while preserving the label. The positive-to-negative ratio was taken to be 0.1, which is the number of labels 1 to labels 0 in the entire dataset.

Amazon Baby Registry: This dataset contains subsets of products selected by customers, each tagged by category (e.g., "toys" or "feeding"). Product descriptions are embedded using BERT. We filter out sets of size < 2 or > 30, and for each valid subset S, we generate $T \supseteq S$ of size 30 by sampling additional items from the same category. For negative samples, T is sampled randomly and verified to satisfy $S \not\subseteq T$.

Noise is added to S in both cases. This models scenarios such as predicting whether a set of displayed products T includes a customer's interest S. On these datasets, we provide three sets of additional experiments. In the first table, we perform the same set-containment task as done in the main text, but with a different negative-to-positive ratio of $90:10$.

Model	Bedding	Feeding	MSWEB	MSNBC
DeepSets	0.91	0.90	0.93	**0.97**
SetTransformer	0.91	0.90	0.94	0.96
FlexSubNet	0.90	0.91	0.94	0.94
Neural SFE	0.91	0.91	0.91	0.93
MASNET-ReLU	**0.95**	**0.93**	**0.97**	**0.97**
MASNET-Hat	0.91	0.92	0.95	**0.97**

Table 10: Set containment in text datasets with 90:10 ratio

Note that, due to the inductive bias of monotonicity, all the positive examples are correctly classified by design is case of MASNET. Thus, it only has to learn to identify the to separate the negative examples. Thus, if the test set has a higher proportion of negative examples, then it throws a model a toughter task to learn. This is shown in the following Table 10. Other than the modified class ratio, we provide two additional tables: in the first one, we compare shallow(1 layer) vs deep ($\geq$ 2 layers) embedding MLP M_{θ_1} in MASNET, which is shown in Table 11. White noise of std 0.1 added for inexact set containment.

Ablation studies on text datasets: From the definition of MASNET in Equation (7), we see that, the formulation os quite similar to Deepsets, which is of the form $F(S) = \mathrm{M}_2\left(\sum_{x\in S} \mathrm{ReLU}(a^\top \mathrm{M}_1(x) + b)\right)$. As shown earlier, using ReLU in the last layer of an elementwise function with a universally approximating M_1 is an instance of MASNET.

Model	bedding	feeding	MSWEB	MSNBC
Shallow-ReLU	0.69	0.36	0.92	0.93
Shallow-MASNET-TRI	0.83	0.76	0.94	0.95
Shallow-MASNET-Hat	0.89	0.77	*0.97*	0.96
Shallow-MASNET-INT	0.90	0.91	0.92	0.95
(Deep)MASNET-ReLU	**0.95**	**0.93**	*0.97*	**0.97**
Deep-MASNET-TRI	0.92	*0.93*	**0.98**	0.96
Deep-MASNET-Hat	0.91	0.92	0.95	*0.97*
Deep-MASNET-INT	*0.93*	0.92	0.96	*0.97*

Table 11: Comparison of shallow vs. deep variants of MASNET across datasets.

But it is different from DeepSets in a key points, namely: • For set containment tasks, we don't have an outer M_2, which DeepSets have; for universal approximation tasks, we use a monotonically increasing M_2 that we enforce by taking positive weights and increasing activation functions. Also, for Hat function based MASNET models, • we use a re-parametrization in which we perform a division based scaling. We now show the effect of the outer monotonic M_2 on DeepSets and division based re-parametrization on MASNET-Hat in the ablation study in table 12. White noise of std 0.1 added for inexact containment.

Pointcloud datasets and related experiments

ModelNet40 [33] is a benchmark dataset of 12,311 CAD models across 40 object categories, with each object represented as a 3D point cloud. We frame our task as checking if a given pointcloud S is a segmenet of a target pointcloud T.

Model	**bedding**	**feeding**	**MSWEB**	**MSNBC**
DeepSets	*0.91*	0.90	0.93	*0.97*
DeepSets, monotone M_2	0.90	0.91	0.94	0.93
MASNET-ReLU	**0.95**	**0.93**	**0.97**	**0.97**
MASNET-Hat	0.91	*0.92*	*0.95*	*0.97*
MASNET-Hat (No division)	0.87	0.92	0.92	0.94

Table 12: Ablation study

Firstly, we choose an object from an object category C_1, and randomly sample 1024 points from that object to get the target point cloud T. To obtain a positive sample (i.e true subset) from T, we first sample a random center point from T, and then extract S using a hybrid approach: selecting the nearest point to the center, a few local neighbors via k-NN, and the rest via importance-weighted sampling (inverse-distance from center with noise). This makes sure that we're selecting a true local region from the point cloud, which is in-line with the actual task of detecting whether a given segment of an object is contained in a target object. For a negative sample(non-subset), we sample S from an object of a different category C_2.

$\lvert S\rvert \rightarrow$	128	256	512
DeepSets	0.90	0.90	0.91
SetTransformer	0.90	0.91	0.91
FlexSubNet	0.84	0.88	0.92
Neural SFE	0.89	0.89	0.90
MASNET-ReLU	**0.91**	**0.93**	**0.98**
MASNET-Hat	0.87	0.91	0.94

Table 13: Performance on Point cloud for different values of $|S|$, for 90:10 class ratio

We first give the accuracy table for PointNet encoder, but with the modified negative-to-positive class ratio of 90:10 to make the task harder for MASNET, which correctly classifies the actual subsets by the induictive bias of monotonicity. The numbers with the modified class-ratio are given in the following table of Tab Table 13:

We also give the accuracy numbers for DGCNN encoder, in Table 14

$\lvert S\rvert \rightarrow$	128	256	512
DeepSets	0.89	0.90	0.90
SetTransformer	0.90	0.91	0.91
MASNET-ReLU	0.87	**0.91**	**0.93**
MASNET-Hat	0.85	0.91	0.89
MASNET-TRI	0.83	0.93	0.87
MASNET-INT	0.87	0.90	0.92

Table 14: Point cloud with DGCNN for different values of $|S|$.

Details on Linear assignment problem In this problem, we are given a positive matrix $M \in \mathbb{R}^{n\times m}, n \leq m$, where $M_{i,j}$ represents the salary that a 'worker' will be paid to do a 'job' j. The goal of the task is to maximize the average salary obtained by all workers. This is done by finding the optimal assignment $\pi \in S_{n,m}$ which maps a worker $i \in [n]$ to a 'job' $\pi(i) \in [m]$, where by construction each worker can be mapped to at most one job, so $\pi(i) \in \{0, 1\}$. This gives us the following maximization problem:

$$F(M) = \frac{1}{n} \max_{\pi\in S_{n,m}} \sum_{i=1}^{n} M_{i,\pi(i)}.$$

Thinking of the matrix M as a set of m columns $M = [M_1, \ldots, M_m]$, we see that the function F is permutation invariant and monotone. Accordingly, our goal will be to evaluate our MASNET model and baselines.