MEASURE-THEORETIC SET REPRESENTATION LEARNING

A PREPRINT

Michael Boratko, Dhruvesh Patel, Shib Sankar Dasgupta, and Andrew McCallum
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
{mboratko,dhruveshpate,ssdasgupta,mccallum}@cs.umass.edu

July 1, 2022

Keywords Representation Learning · Embeddings · Box Embeddings · Fuzzy Sets

1 Introduction

Sets are fundamental to human knowledge representation and reasoning, most explicitly as the axiomatic cornerstone on which all of modern mathematics is built, but also more primitively in the organization of thought into *concepts* [Peacocke, 1992]. Much of machine learning is also about identifying concepts, not just in natural language where interpreting human communication may obviously require modeling conceptual structure, but in ways which span across modalities - image classification, video recognition, everything from protein folding to music generation requires models which can effectively represent, decompose, and combine concepts.

Many machine learning systems involve matching user queries to internal representations, and sets are a better way to handle both ends of this interaction. Preference expression through set operations is very natural for humans. For example, a user may express a preference to watch "comedies, but not romantic comedies," or perhaps search for a research paper about "non-embedding-based approaches to open-domain question answering over knowledge graphs." Models with explicit set representations also provide better interpretability and opportunities for constraint injection, as set-theoretic rules can be extracted or injected in a form amenable to human interpretation. Much of the success in lightly-supervised training is the result of finding ways to provide constraints that efficiently guide the model away from providing bad outputs, and such constraints are often expressed most easily using set operations.

Naïve set theory is not enough to capture *everything* we care about, of course. Machine learning is also inherently about modeling *uncertainty*. Modeling drug-disease datasets, for example, would require not only set-theoretic representations but also *uncertainty* over membership in a given set, as the effect of a drug for treating a particular disease is not deterministic. Probabilistic reasoning can be seen as the extension of set theoretic operations to settings with uncertainty, and thus it is necessary to have not only the option of flexible membership representations but also valid probabilistic semantics.

In this paper we explore, from first principles, the task of learning differentiable representations of sets. Motivated by the use-cases suggested above, we outline specific requirements these representations must adhere to. We derive a family of representations with valid set-theoretic and probabilistic semantics, capable of graded membership, complex queries, and capturing dense representations of dependencies.

2 Notation

We will use the shorthand $[\![n]\!] \coloneqq \{1, \ldots, n\}$. Given any product of sets $X_1 \times \cdots \times X_n = \prod_{i=1}^{n} X_i$, we let $\pi_i \colon \prod_{i=1}^{n} X_n \to X$ denote projection to the i^{th} coordinate, i.e. $\pi_i(x_1, \ldots, x_n) = x_i$. The set of functions from X to Y can be written, with decreasing verbosity, as

$$\{f \mid f \colon X \to Y\} = \{X \to Y\} = Y^X.$$

Given some set $\mathcal{U}$, we denote the *powerset* (set containing all subsets) of $\mathcal{U}$ as $\mathcal{P}(\mathcal{U}) = \{S \mid S \subseteq \mathcal{U}\}$. When considering some subset $S \subseteq \mathcal{U}$ we will often refer to $\mathcal{U}$ as the *universe*. For a fixed universe $\mathcal{U}$, subsets $S \subseteq \mathcal{U}$ can be represented via their *characteristic function*,

$$\mathbb{1}_S \colon \mathcal{U} \to \{0,1\} \quad \text{where} \quad \mathbb{1}_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Thus, we think of the powerset $\mathcal{P}(\mathcal{U})$ as equivalent to the set of functions $\{0,1\}^{\mathcal{U}}$, commonly shortened to just $2^{\mathcal{U}}$.

The standard notations $\cap$ for intersection and $\cup$ for union will be used to denote their traditional naïve set-theoretic operations, and Δ denotes symmetric difference, i.e. $A \Delta B = (A \cup B) \setminus (A \cap B)$. While typically written using infix notation, we may also use prefix notation when emphasizing the interpretation of these operations as functions from $2^{\mathcal{U}} \times 2^{\mathcal{U}}$ to $2^{\mathcal{U}}$, eg. $\cap(A, B) = A \cap B$. In addition, given some universe $\mathcal{U}$ we will denote the set-theoretic complement of $A \subseteq \mathcal{U}$ as $A^c = \{x \in \mathcal{U} \mid x \notin A\}$, and use $\complement : 2^{\mathcal{U}} \to 2^{\mathcal{U}}$ when referring to the complement as a function, i.e. $\complement(A) = A^c$. We say a collection of sets $\mathcal{F}$ is *closed under intersection* if, for all $A, B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$, a similarly use the terms *closed under union* and *closed under complement*.

3 Set-Theoretic Embedding

In this section we formally define the notion of a *set-theoretic embedding* as an injective map which preserves set-theoretic operations such as intersection, union, and complement. Such a function is known, abstractly, as a *homomorphism of Boolean algebras*[1], however this abstraction is better suited to purely mathematical investigations, whereas our motivation is to practically learn such a function via gradient descent, leading to a slightly different line of inquiry.

3.1 Motivating Example - Market Basket

As a running example, consider a market-basket task, where $\mathcal{U}_{\mathbf{M}} \coloneqq \{B_i\}_{i=1}^n$ is an indexed set[2] of "baskets", themselves sets of products $p \in P$. We are interested in modeling the sets of baskets which contain a particular product, i.e. for each $p \in P$ we consider

$$\text{baskets}(p) \coloneqq \{B_i \in \mathcal{U}_{\mathbf{M}} \mid p \in B_i\}, \quad \text{and} \quad \mathbf{M} \coloneqq \{\text{baskets}(p) \mid p \in P\}.$$

For purposes of illustration, suppose we only have 3 baskets:

$$B_1 = \{\texttt{bread}, \texttt{butter}\}, \quad B_2 = \{\texttt{bread}, \texttt{milk}\}, \quad B_3 = \{\texttt{butter}, \texttt{milk}\}.$$

Then

$$\mathbf{M} = \big\{ \text{baskets}(\texttt{bread}) = \{B_1, B_2\}, \quad \text{baskets}(\texttt{butter}) = \{B_1, B_3\}, \quad \text{baskets}(\texttt{milk}) = \{B_2, B_3\} \big\}.$$

There are various tasks which would benefit from the set-theoretic information in $\mathbf{M}$. Recommendation, for example, can be viewed as an explicit set-theoretic query (eg. "what items are purchased when also purchasing `bread` or `milk`?"), or a probabilistic query (eg. "what is the probability of purchasing `butter`, if one has also purchased `bread`?"). While such queries can be performed symbolically, these computations can become intractable with large data, motivating the use of a compact representation capable of encoding such set-theoretic and probabilistic information. In addition, the data may also be sparse, and we may wish to learn to generalize appropriately from both latent features and explicit hierarchies - for example, if we observe a cart which contains `vegan_butter`, we may wish to suggest `coconut_milk`. This sort of generalization is typically solved by embedding the discrete data in some dense space, thus we are after a representation which can encode both set-theoretic and probabilistic information.

Many existing tasks can be naturally formulated in a probabilistic or set-theoretic way, however even tasks which may not admit such a formulation may benefit from being able to encode this sort of information in a latent space. To do so in the context of a larger model trained end-to-end via gradient descent, it is imperative to have a differentiable representation of these sets. This representation will be accomplished via an *embedding*.

[1] Boolean algebras are also equivalent to an earlier formalism known as the "algebra of concepts", proposed by Leibniz under the guiding principle of constructing a system for combining a small number of simple ideas which form the "alphabet of human thought". [Geiger and Rudzka-Ostyn, 1993, Zalta, 2000, Leibniz and R., 1966]

[2] An indexed set allows for the existence of distinct baskets which contain the same items. This can be formalized by saying that baskets are represented as tuples of indices and sets of products, eg. $B_1 = (1, \{\texttt{bread}, \texttt{butter}\})$, however it is common to abuse notation slightly and drop the explicit tuple notation.

3.2 Background - Embedding as Morphism

An *embedding* is an injective structure-preserving map $f : \mathbf{X} \to \mathbf{Y}$, formalized in category theory as a *morphism*. For our purposes, it is sufficient to consider the *structure* as some n-ary relation(s) which may be required to satisfy a set of axioms. A common morphism ubiquitous in machine learning is a linear transformation, which is a morphism on the category of vector spaces. Here the structure includes the operations of vector addition and scalar multiplication together with the familiar axioms they must satisfy. A linear transformation $L : \mathbb{R}^m \to \mathbb{R}^n$ preserves this structure, in the sense that

$$L(a\vec{x} + b\vec{y}) = aL\vec{x} + bL\vec{y} \quad \text{for all } a, b \in \mathbb{R}, \vec{x}, \vec{y} \in \mathbb{R}^m.$$

The notation $f : \mathbf{X} \hookrightarrow \mathbf{Y}$ is used as shorthand to emphasize that f is an embedding, and is intended to evoke the impression that $\mathbf{X}$ can be identified with the subset $f(\mathbf{X}) \subseteq \mathbf{Y}$ in a structure-preserving way.

3.3 Background - Algebras

In this work our focus will be on preserving set-theoretic structure, which is captured abstractly by the concept of an *algebra*.

Given a universe $\mathcal{U}$, a collection of subsets $\mathcal{F} \subseteq 2^{\mathcal{U}}$ is called an *algebra over* $\mathcal{U}$ if it:

1. Contains the empty set ($\emptyset \in \mathcal{F}$).
2. Is closed under complements (for all $A \in \mathcal{F}$, $A^c \in \mathcal{F}$).
3. Is closed under binary intersections (for all $A, B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$).

Combining property 2 and 3, De Morgan's laws imply that algebras are also closed under binary unions. By repetition this also implies that algebras are closed under any *finite* combination of intersection, union, and complement operations. Given an arbitrary set $\mathcal{G} \subseteq 2^{\mathcal{U}}$, we define $\mathcal{A}(\mathcal{G})$ to be the smallest algebra containing $\mathcal{G}$.

Definition 1 (Homomorphism of Algebras)**.** Given some algebras $\mathcal{F} \subseteq 2^{\mathcal{U}_{\mathcal{F}}}$ and $\mathcal{G} \subseteq 2^{\mathcal{U}_{\mathcal{G}}}$, a **homomorphism of algebras** from $(\mathcal{F}, \cap, \cup, \complement)$ to $(\mathcal{G}, \cap, \cup, \complement)$ is an injective function $f \colon \mathcal{F} \hookrightarrow \mathcal{G}$ such that for all $A, B \in \mathcal{F}$:

1. $f(A \cap B) = f(A) \cap f(B)$
2. $f(A \cup B) = f(A) \cup f(B)$
3. $f(\emptyset) = \emptyset$
4. $f(\mathcal{U}_{\mathbf{X}}) = \mathcal{U}_{\mathbf{Y}}$

We obtain, as a consequence, that $f(\complement(A)) = \complement(f(A))$ for all $A \in \mathcal{F}$ as well.

Remark 1 (Boolean Algebras)**.** The 4-tuple $(\mathcal{F}, \cap, \cup, \complement)$ is a specific instance of a more general algebraic structure known as a *Boolean algebra*, which axiomatically encodes set-theoretic logic, characterizing 4-tuples of arbitrary sets and operations $(A, \vee, \wedge, \neg)$. This implies that the formalisms which we will discuss going forward can be used to learn representations of arbitrary Boolean algebras, however working at this level of generalization can cloud the issue. As our interest at this time exclusively involves sets, we will not have a need to abstract the structure in this way, and thus restrict our treatment to the setting of algbras.

Finally, given any set $\mathcal{G} \subseteq 2^{\mathcal{U}}$ and any functions $\gamma_1 : \mathcal{X} \to \mathcal{A}(\mathcal{G}), \gamma_2 : \mathcal{Y} \to \mathcal{A}(\mathcal{G})$ where $\mathcal{X}, \mathcal{Y}$ are any sets, we define the functions

$$\begin{aligned}
\gamma_1 \cap \gamma_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{A}(\mathcal{G}), &\quad \gamma_1 \cap \gamma_2(X, Y) = \gamma_1(X) \cap \gamma_2(Y), \\
\gamma_1 \cup \gamma_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{A}(\mathcal{G}), &\quad \gamma_1 \cup \gamma_2(X, Y) = \gamma_1(X) \cup \gamma_2(Y), \\
\complement\gamma_1 = \gamma_1^c : \mathcal{X} \to \mathcal{A}(\mathcal{G}), &\quad (\complement\gamma_1)(X) = \gamma_1^c(X) = \gamma_1(X)^c.
\end{aligned}$$

We define the collection of all finite combinations of set operations on $\mathcal{G}$ as

$$\Gamma_{\mathcal{G}} \coloneqq \{\gamma \mid \gamma = \iota_{\mathcal{G}} \text{ or } \gamma = \gamma_1 \cap \gamma_2 \text{ or } \gamma = \gamma_1 \cup \gamma_2 \text{ or } \gamma = \complement\gamma_1 \text{ for some } \gamma_1, \gamma_2 \in \Gamma_{\mathcal{G}}\}$$

where $\iota_{\mathcal{G}} \colon \mathcal{G} \to \mathcal{A}(\mathcal{G})$ is the inclusion map, $\iota_{\mathcal{G}}(S) = S$. For example, the function $\gamma \colon \mathcal{G}^3 \to \mathcal{A}(\mathcal{G})$ which is defined as $\gamma(A, B, C) = (A^c \cap B) \cup C$ is in $\Gamma_{\mathcal{G}}$. Note that, given any $S \in \mathcal{A}(\mathcal{G})$, there exists some $S_1, \ldots, S_n \in \mathcal{G}$ and $\gamma \in \Gamma_{\mathcal{G}}$ such that $S = \gamma(S_1, \ldots, S_n)$. For any collections of sets $\mathcal{F}, \mathcal{G}$ there is a natural bijection $J_{\mathcal{F} \to \mathcal{G}} \colon \Gamma_{\mathcal{F}} \to \Gamma_{\mathcal{G}}$ where for $J_{\mathcal{F} \to \mathcal{G}}(\gamma)$ is obtained by replacing $\iota_{\mathcal{F}}$ with $\iota_{\mathcal{G}}$.

3.4 Set-Theoretic Embeddings

In this section, we will formally define the notion of a set-theoretic embedding. It may seem as though this has already been accomplished by Definition 1, however as mentioned at the start of this section our eventual goal to *learn* such an embedding has practical considerations which need to be taken into account. One such consideration is that parameterizing the set of valid homomorphisms between given algebras is challenging, as the size of the algebra can be exponentially large and the structure preserved in Definition 1 implies there are many complicated dependencies between the mappings of these elements.

Typically the task itself has a natural set of subsets which generate the algebra for the input space. For example, in the market basket setting we identified the sets $\mathbf{M}$, and our example queries involved elements of $\mathcal{A}(\mathbf{M})$. In this case, $\mathcal{A}(\mathbf{M}) = 2^{\mathcal{U}_\mathbf{M}}$, an example of the setting where we would need to define a map on exponentially many elements, and furthermore ensure that our map was, in fact, a valid homomorphism of algebras, adhering to all the various set-theoretic dependencies that implies. A simpler approach is to determine if such an embedding can be defined on some subset of the algebra with minimal dependencies.

For any collections of sets $\mathbf{X}, \mathbf{Y}$, any homomorphism $F\colon \mathcal{A}(\mathbf{X}) \hookrightarrow \mathcal{A}(\mathbf{Y})$ is uniquely defined by it's values on $\mathbf{X}$. Namely, for any $A \in \mathcal{A}(\mathbf{X})$ there exists some $A_1, \ldots, A_n \in \mathbf{X}$ and an operation $\gamma_\mathbf{X} \in \Gamma_\mathbf{X}$ such that $A = \gamma_\mathbf{X}(A_1, \ldots, A_n)$. As F is a homomorphism, we have that

$$F(A) = F(\gamma_\mathbf{X}(A_1, \ldots, A_n)) = \gamma_\mathbf{Y}(F(A_1), \ldots, F(A_n)). \tag{2}$$

where $\gamma_\mathbf{Y} = J_{\mathbf{X}\to\mathbf{Y}}(\gamma_\mathbf{X})$ is the operation which corresponds to $\gamma_\mathbf{X}$ by replacing $\mathrm{id}_\mathbf{X}$ with $\mathrm{id}_\mathbf{Y}$. This motivates the following definition of a set-theoretic embedding.

Definition 2 (Set-Theoretic Embedding)**.** Given sets $\mathbf{X} \subseteq 2^{\mathcal{U}_\mathbf{X}}$ and $\mathbf{Y} \subseteq 2^{\mathcal{U}_\mathbf{Y}}$, a **set-theoretic embedding of $\mathbf{X}$ in $\mathbf{Y}$** is an injective function $f\colon \mathbf{X} \hookrightarrow \mathbf{Y}$ which can be extended to a homomorphism of algebras. That is, there exists a homomorphism $F\colon \mathcal{A}(\mathbf{X}) \hookrightarrow \mathcal{A}(\mathbf{Y})$ such that $F\big|_\mathbf{X} = f$.

Equation (2) also implies the following necessary (but not sufficient) condition on set-theoretic embeddings.

Proposition 1. *Let $f\colon \mathbf{X} \hookrightarrow \mathbf{Y}$ be a set-theoretic embedding. For all $A \in \mathbf{X}$ such that $A = \gamma_\mathbf{X}(A_1, \ldots, A_n)$ for some $A_1, \ldots, A_n \in \mathbf{X}$ and $\gamma_\mathbf{X} \in \Gamma_\mathbf{X}$, we have*

$$f(A) = \gamma_\mathbf{Y}(f(A_1), \ldots, f(A_n)),$$

where $\gamma_\mathbf{Y} = J_{\mathbf{X}\to\mathbf{Y}}(\gamma_\mathbf{X})$.

The function f will ultimately be learned, and the implied homomorphism F will be used when querying arbitrary elements of $\mathcal{A}(\mathbf{X})$. Proposition 1 highlights the dependencies f must capture between elements of $\mathbf{X}$, and thus a simple way to satisfy such dependencies is to simply remove any such A from $\mathbf{X}$. Given a fixed algebra $\mathcal{F}$ which we wish to embed, it would make sense to select $\mathbf{X} \subseteq \mathcal{F}$ to be as small as possible, while still ensuring that $\mathcal{A}(\mathbf{X}) = \mathcal{F}$. [3]

4 Set Representation Learning

In the previous section we defined the notion of a set-theoretic embedding, and provided some guidance on selecting the subset of a given algebra to define the embedding on. In the rest of this work, we will discuss the various options for output spaces. We start with $\mathcal{U}_\mathbf{Y} = \mathbb{R}$, but this will easily be generalized and extended to $\mathbb{R}^d$ and beyond.

4.1 Background - Representation Learning

Representation learning is, fundamentally, the task of finding an embedding for a given input set $\mathbf{X}$ by observing some evidence of the structure in $\mathbf{X}$. In the case of word embedding, for example, the structure of $\mathbf{X}$ is a latent "similarity", observed only indirectly through cooccurrance, and the set $\mathbf{Y}$ is often chosen to be a *real inner product space*, with the goal that the inner product should capture the similarity, i.e. for all $a, b \in \mathbf{X}$,

$$a \text{ is "similar to" } b \quad \text{if and only if} \quad \langle f(a), f(b)\rangle \text{ is "sufficiently large".}$$

In order to *learn* f we need a way of scoring the various functions in $\{f : \mathbf{X} \to \mathbf{Y}\} = \mathbf{Y}^\mathbf{X}$, typically called a loss function, $L\colon \mathbf{Y}^\mathbf{X} \to \mathbb{R}_{\geq 0}$, such that our preferred $f = \arg\min L$. In order to train this function via gradient descent,

[3] This is related to the notion of *generators* for a *free Boolean algebra*, but these require much more, and are not needed for our current purposes, which, practically speaking, simply state that we can learn a function on a subset of the algebra, and that removing any dependencies in $\mathbf{X}$ will reduce the possibility of a conflict.

it is necessary that we somehow parameterize the function space $\mathbf{Y}^{\mathbf{X}}$ via some open set $W \subseteq \mathbb{R}^m$, that is we need a surjection $\psi : W \to \mathbf{Y}^{\mathbf{X}}$ which allows us to perform gradient descent on $L \circ \psi \colon W \to \mathbb{R}_{\geq 0}$, assuming this function is differentiable.

If $\mathbf{X}$ is finite, there is a natural bijection $\rho \colon \mathbf{Y}^{\mathbf{X}} \to \mathbf{Y}^{|\mathbf{X}|}$. Namely, fix an ordering $\mathbf{X} = \{x_1, \ldots, x_{|\mathbf{X}|}\}$, then let $\rho(f) = (f(x_1), \ldots, f(x_{|\mathbf{X}|})) \in \mathbf{Y}^{|\mathbf{X}|}$. Conversely, any vector $(y_1, \ldots, y_{|\mathbf{X}|}) \in \mathbf{Y}^{|\mathbf{X}|}$ uniquely defines an $f \colon \mathbf{X} \to \mathbf{Y}$ such that $f(x_i) = y_i$, which is precisely ρ^{-1}. In particular, this implies that if $\mathbf{Y} = \mathbb{R}^d$ then we can obtain the necessary $\psi = \rho^{-1} \colon \mathbb{R}^{d|\mathbf{X}|} \to \mathbf{Y}^{\mathbf{X}}$.

If $\mathbf{Y}$ is not already $\mathbb{R}^d$ (or some subset), however, this shows that $\mathbf{Y}$ itself must be capable of parameterization.[4] Given some surjection $\varphi \colon \Omega \to \mathbf{Y}$ for $\Omega \subseteq \mathbb{R}^d$, we can define

$$g \colon \Omega^{|\mathbf{X}|} \to \mathbf{Y}^{|\mathbf{X}|} \quad \text{where} \quad g_i(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{|\mathbf{X}|}) = \varphi(\boldsymbol{\theta}_i),$$

and thereby obtain $\psi = \rho^{-1} \circ g \colon \Omega^{|\mathbf{X}|} \to \mathbf{Y}^{\mathbf{X}}$ which provides the necessary parameterization of $\mathbf{Y}^{\mathbf{X}}$ to allow $L \circ \psi$ to be trained. Such a parameterization is also useful if $\mathbf{X}$ is infinite, in which case the space $\mathbf{Y}^{\mathbf{X}}$ has greater cardinality than $\mathbb{R}^d$, and thus no such ψ exists. In both the finite and infinite case, we can (and routinely do) consider subsets of $\mathbf{Y}^{\mathbf{X}}$ given by parameterized familes of functions (eg. neural networks) which are composed with the parameterization $\varphi \colon \Omega \to \mathbf{Y}$, allowing for these parameterized familes of functions to output elements of $\mathbf{Y}$.

4.2 Parameterizablity

The discussion in the preceeding section provides our first requirement of the embedding space.

Desiderata 1 (Parameterizable)**.** The space $\mathbf{Y}$ should be capable of being parameterized via a surjective map $\varphi \colon \Omega \to \mathbf{Y}$ for some open set $\Omega \subseteq \mathbb{R}^d$.

Such a parameterization always exists as long as the cardinality of $\mathbf{Y}$ is less than the cardinality of $\mathbb{R}$,[5] however the choice of parameterization may have implications as to the ease of computation and differentiability of various set operations.

Example 1 (Rays)**.** Let $\mathbf{Y} = \mathbf{R} \coloneqq \{[x, \infty) \mid x \in \mathbb{R}\}$ be a set of rays in $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}$, which can be parameterized as $\varphi \colon \mathbb{R} \to \mathbf{R}$, with $\varphi(x) = [x, \infty)$. Note that

$$[x, \infty) \cap [y, \infty) = [\max(x, y), \infty) \quad \text{and} \quad [x, \infty) \cup [y, \infty) = [\min(x, y), \infty).$$

Therefore, the intersection and union can be lifted to maps on the parameters, $\widetilde{\cap} \colon \mathbb{R}^2 \to \mathbb{R}$ and $\widetilde{\cup} \colon \mathbb{R}^2 \to \mathbb{R}$, with

$$\widetilde{\cap}(x, y) = \max(x, y) \quad \text{and} \quad \widetilde{\cup}(x, y) = \min(x, y),$$

which are differentiable almost everywhere.

Notably, the complement operation $\complement$ is absent from the set-theoretic operations above, and for good reason - $\mathbf{R}$ is not even closed with respect to $\complement$.

Example 2 (Intervals)**.** Let $\mathbf{Y} = \mathbf{I} \coloneqq \{[a, b] \mid a, b \in \mathbb{R}\}$ be the set of closed and bounded intervals in $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}$.[6] This can be parameterized with $\varphi : \Omega = \mathbb{R}^2 \to \mathbf{I}$ defined by $\varphi(a, b) = [a, b]$. We have

$$[a, b] \cap [c, d] = [\max(a, c), \min(b, d)],$$

and thus we can lift the intersection to a map on the parameters $\widetilde{\cap} : \mathbb{R}^4 \to \mathbb{R}^2$ given by

$$\widetilde{\cap}(a, b, c, d) = (\max(a, c), \min(b, d)).$$

In this case, we *only* have closure with respect to intersection. While it certainly would be desirable to have closure with respect to all basic set-theoretic operations, thus requiring that $\mathbf{Y}$ be an algebra, this competes with the pragmatic requirement of parameterizing the elements of $\mathbf{Y}$ in a way which is amenable to learning such representations.[7] The

[4] Proof contained in Appendix A.

[5] This is not a vacuous requirement, however. Note that Cantor's theorem implies that the powerset $2^{\mathbb{R}}$ has cardinality strictly greater than $|\mathbb{R}|$. This also excludes $\mathbf{Y} = \mathcal{M}$, the Lebesgue measurable sets on $\mathbb{R}$ (see Section 4.4), which can be shown to have cardinality $2^{|\mathbb{R}|}$.

[6] Recall that $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, thus in particular if $a > b$, $[a, b] = \emptyset$.

[7] The technical reasons for this deserve more attention. It can be shown that if $|\mathbf{Y}| \leq |\mathbb{R}|$ then $|\mathcal{A}(\mathbf{Y})| \leq |\mathbb{R}|$ (assuming the axiom of choice). Thus it is possible to parameterize $\mathcal{A}(\mathbf{R})$ using $\mathbb{R}$, however such parameterizations are almost certainly not able to yield trainable representations as very "similar" sets may have very different parameters (for example). This will be discussed more technically in Section 4.6.

desiderata of the following sections will formalize this, and moreover reveal that it is not actually necessary for $\mathbf{Y}$ to be an algebra, but rather that the measure of sets in $\mathcal{A}(\mathbf{Y})$ are differentiable functions of the parameterization. By leveraging inclusion-exclusion, we will find that closure under intersection is sufficient for this purpose, as long as the intersection operation is differentiable.

4.3 Loss Functions for Set-Theoretic Operations

Given a parameterization φ we can start to consider methods of *learning* the function f from data, which amounts to performing gradient descent on some loss function L. As is often the case, the loss will be a sum over training examples. Typically, our training data does not take the form of explicit equality constraints, such as $A \cap B = C$ for some $A, B, C \in \mathbf{X}$, but rather as evidence of various set-theoretic relationships.

For example, in the market basket example (Section 3.1), our training data may be the set of baskets, $\mathcal{U}_{\mathbf{M}}$. Upon observing $B_1 = \{\texttt{bread}, \texttt{butter}\}$, this informs us that

$$\text{baskets}(\texttt{bread}) \cap \text{baskets}(\texttt{butter}) \neq \emptyset \tag{3}$$

Therefore, for this training example, we would like to consider a loss function which encourages this intersection to be non-empty. If $\mathbf{Y} = \mathbf{I}$ this can be accomplished by manually inspecting the parameterization, and noting that $\text{baskets}(\texttt{bread}) = [a, b]$ intersects $\text{baskets}(\texttt{butter}) = [c, d]$ if and only if $c < b < d$ or $c < a < d$, which might suggest the hand-crafted loss function

$$h(a, b, c, d) = \min(\max(d - \varepsilon - b, b - (c + \varepsilon), 0), \max(d - \varepsilon - a, a - (c + \varepsilon), 0)),$$

where ε is some adaptive margin (eg. $\varepsilon = \frac{d-c}{100}$) to ensure it is possible to satisfy this constraint. This is unappealing for a number of reasons, but most problematic is the fact that extending this approach to other potential training examples is nontrivial - for example, if we encountered a basket such as $B_4 = \{\texttt{bread}, \texttt{butter}, \texttt{milk}\}$, for which we would like a similar function to encourage

$$\text{baskets}(\texttt{bread}) \cap \text{baskets}(\texttt{butter}) \cap \text{baskets}(\texttt{milk}) \neq \emptyset. \tag{4}$$

An alternative approach is to note that we can define the length of an interval $\ell : \mathbf{I} \to [0, \infty)$ as $\ell([a, b]) = \max(b - a, 0)$. The intersection of intervals is given by

$$[a, b] \cap [c, d] = [\max(a, c), \min(b, d)],$$

thus all that is really required is to ensure the size of this interval is positive, which can be accomplished with a max-margin loss,

$$h(a, b, c, d) = \max(\varepsilon - \ell([a, b] \cap [c, d]), 0) = \max(\varepsilon - \max(\min(b, d) - \max(a, c), 0), 0),$$

where here $\varepsilon > 0$ is a global hyperparameter. The benefit of this approach is that, since $\mathbf{I}$ is closed under intersection, it can be extended to n-ary intersections. For example, to handle Equation (4), if $\text{baskets}(\texttt{milk}) = [e, f]$ we would have

$$\begin{aligned} h(a, b, c, d, e, f) &= \max(\varepsilon - \ell([a, b] \cap [c, d] \cap [e, f]), 0) \\ &= \max(\varepsilon - \max(\min(b, d, f) - \max(a, c, e), 0), 0). \end{aligned}$$

We could also handle loss functions involving unions, by observing that the union of two intervals is (at most) two disjoint intervals, for which the total length can be calculated using inclusion-exclusion as

$$\ell([a, b]) + \ell([c, d]) - \ell([a, b] \cap [c, d]).$$

This is still a rather customized loss function, however, in that it seems to depend heavily on the use of $\mathbf{I}$, and does not seem to extend to arbitrary objectives which may involve more complicated elements of $\mathcal{A}(\mathbf{Y})$ - for example, it does not allow for complements to be trained. These objections can be handled by observing that the fundamental aspect which made this approach possible is the function ℓ, which assigned a size to a set.

4.4 Background - Measure Theory

The seemingly simple objective of assigning a notion of "size" to a set has a myriad of nuances and subtleties which have been handled rigorously in *measure theory*, which also forms the foundational framework of integration and probability theory. The difficulties stem from ensuring a set can be assigned a consistent size, regardless of the way in which the set is constructed via set-theoretic operations. We will give a very brief introduction to measure theory here.[8]

Emboldened by the success above measuring the length of intervals using $\ell \colon \mathbf{I} \to \mathbb{R}_{\geq 0}$, we wish to extend ℓ to a function $\lambda \colon 2^{\mathbb{R}} \to [0, \infty]$ which behaves intuitively as a "size" of the set. That is, we would like λ to have the following properties:

[8]The introduction we provide is an abbreviated version of the notes `https://math.unl.edu/~gmeisters1/papers/Measure/measure.pdf`. For a more complete introduction, see Folland [1999].

1. **Extension of Length:** $\lambda([a,b]) = \ell([a,b]) = \max(b-a, 0)$.
2. **Monotonicity:** If $A \subseteq B \subseteq \mathbb{R}$ then $\lambda(A) \leq \lambda(B)$.
3. **Translation Invariance:** For each $A \subseteq \mathbb{R}$ and $x_0 \in \mathbb{R}$, $\lambda(\{a + x_0 \mid a \in A\}) = \lambda(A)$.
4. **Countable Additivity:** If $\{A_i\}_{i=1}^{\infty}$ is a set of disjoint subsets of $\mathbb{R}$, then $\lambda(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$.

Unfortunately, no such μ exists, as there are pathological subsets of $\mathbb{R}$ which lead to contradictory assignments of measure. The largest family of subsets $\mathcal{M} \subseteq \mathcal{P}(\mathcal{U})$ for which we can define $\lambda\colon \mathcal{M} \to [0,\infty]$ such that (1)–(4) hold are called *Lebesgue measurable*, and λ is called *Lebesgue measure*. Obviously $\mathbf{I} \subseteq \mathcal{M}$, and it can be shown that $\mathcal{M}$ is closed under complements and finite intersections, and therefore $\mathcal{M}$ is an algebra. In fact, $\mathcal{M}$ is closed under *countably infinite* [9] intersections.

An algebra $\mathcal{F}$ which is also closed under countably infinite intersections (or, equivalently, unions) is known as a *σ-algebra*. Given an abitrary subset $\mathcal{G} \subseteq 2^{\mathcal{U}}$, we denote $\sigma(\mathcal{G})$ the smallest σ-algebra containing $\mathcal{G}$. Note that $\mathcal{G} \subseteq \mathcal{A}(\mathcal{G}) \subseteq \sigma(\mathcal{G}) \subseteq 2^{\mathcal{U}}$. The usefulness of $\sigma(\mathcal{G})$ is that it excludes some sets which are pathological, in the sense that their size cannot be defined consistently, and thus allows us to define a measure.

Definition 3 (Measure)**.** Given a σ-algebra $\mathcal{F}$, a *measure* is a countably-additive function $\mu\colon \mathcal{F} \to [0,\infty]$ for which $\mu(\emptyset) = 0$.

4.5 Measure-Theoretic Loss Functions

The concept of a measure allows us to not only extend beyond just intersection, but also to generalize to arbitrary embedding spaces $\mathbf{Y}$, as long as such a measure exists.

Desiderata 2 (Measurable)**.** The set $\mathbf{Y}$ should admit a measure $\nu\colon \sigma(\mathbf{Y}) \to [0,\infty]$.

Lebesgue measure allows us to generalize from $\mathbf{Y} = \mathbf{I}$ and satisfy Desiderata 2 for any $\mathbf{Y} \subseteq \mathcal{M}$, however this may be unsatisfying in various respects. For example, $\mathbf{R} \subseteq \mathcal{M}$, but $\lambda([x,\infty)) = \infty$ for any x, so we would not be able to make meaningful comparisons between the elements of $\mathbf{R}$. A simple solution is to use a Lebesgue-integrable function on $\mathbb{R}$ which (at least) assigns finite measures to elements in $\mathbf{Y}$.

Example 3 (Finite Measure of Rays)**.** The function $\nu(U) = \int_U e^{-x}\, d\lambda(x)$ is a measure on $\sigma(\mathbf{R})$ which assigns finite measure to elements in $\mathbf{R}$.[10] In particular, $\nu([x,\infty)) = e^{-x}$.

While this solves the problem when comparing elements in $\mathbf{Y}$, there may be set-theoretic queries we are interested in which return elements of $\mathcal{A}(\mathbf{Y})$ which have infinite volume. For example, if $\nu(\mathcal{U}_{\mathbf{Y}}) = \infty$ then for any $U \in \mathbf{Y}$ we have $\nu(U) = \infty$ or $\nu(U^c) = \infty$. If we are interested in such queries, we must further require ν to be a *finite measure*, where $\nu(\mathcal{U}_{\mathbf{Y}}) < \infty$.

Example 4 (Finite Measure on $\mathbb{R}$)**.** The function $\nu(U) = \int_U e^{-|x|}\, d\lambda(x)$, is a finite measure on $\sigma(\mathbf{R})$. In particular,

$$\nu([x,\infty)) = \begin{cases} e^{-x} & \text{if } x \geq 0, \\ 2 - e^{x} & \text{otherwise.} \end{cases}$$

Remark 2 (Probabilistic Modeling)**.** There are some settings where the training data allows us to calculate a measure μ on $\sigma(\mathbf{X})$ as well. For example, in the market basket example, for any $S \in 2^{\mathcal{U}_{\mathbf{M}}}$ we can define $\mu(S) = |S|$. If $\mathcal{U}_{\mathbf{M}}$ is finite, we can normalize μ by dividing by $|\mathcal{U}_{\mathbf{M}}|$, allowing us to interpret the measure probabilistically. For example,

$$P(\texttt{milk} \in B \,|\, \texttt{bread} \in B) = \frac{\mu(\text{baskets}(\texttt{milk}) \cap \text{baskets}(\texttt{bread}))}{\mu(\text{baskets}(\texttt{bread}))}.$$

We may choose to normalize any finite measure ν on $\sigma(\mathbf{Y})$ such that $\nu(\mathcal{U}_{\mathbf{Y}}) = 1$, which yields a probabilistic interpretation of the elements of $\mathbf{Y}$ as event sets, and train the set-theoretic embedding by using a probabilistic loss function, for example KL-divergence.

Requiring a finite measure is still not enough to completely escape the problem, as Desiderata 2 can be satisfied trivially for any $\mathbf{Y}$ by assigning all sets in $\sigma(\mathbf{Y})$ zero measure, for example. These issues will be addressed in the following section, which stipulate how the parameterization and measure should interact.

[9] A countably infinite set is one which is in bijection with the natural numbers. Thus, a countably infinite intersection is one which can be written as $A = \bigcap_{i=1}^{\infty} A_i$ for some sets A_i.

[10] Lai and Hockenmaier [2017] used the product measure of ν^d on the space of cones in the positive orthant, $\mathbf{R}^d \cap \mathbb{R}^d_{\geq 0}$, allowing a probabilistic interpretation of the volume. Their model fits into the framework we are developing. For a discussion on product spaces see Section 6.3, and for a discussion on their model see Section 7.

4.6 Differentiability of Training Objectives

With a measure in hand, our training objectives can now target any element of the algebra $\mathcal{A}(\mathbf{X})$. Our loss function is typically a sum over a set of training objectives,

$$L(f) = \sum_{h \in H} h(f)$$

where each $h(f)$ is some objective based on a single training example, and H is the set of all such objectives. For example, to handle (3) from our market basket task,

$$h(f) = \min\left(\varepsilon - \nu\left(f(\text{baskets}(\texttt{bread})) \cap f(\text{baskets}(\texttt{butter}))\right), 0\right), \tag{5}$$

and the triple intersection from (4) is handled similarly as

$$h(f) = \min\left(\varepsilon - \nu\left(f(\text{baskets}(\texttt{bread})) \cap f(\text{baskets}(\texttt{butter})) \cap f(\text{baskets}(\texttt{milk}))\right), 0\right).$$

This measure-theoretic objective is not limited to targeting intersections, of course. As it is a measure, we can also directly target the volume of arbitrary elements of $\mathcal{A}(\mathbf{X})$, as for any $A_1, \ldots, A_n \in \mathbf{X}$ and $\gamma_{\mathbf{X}} \in \Gamma_{\mathbf{X}}$ we can calculate $\nu(\gamma_{\mathbf{Y}}(f(A_1), \ldots, f(A_n)))$.

We can use this to encode things beyond mere disjointness or intersection, for example

$$h(f) = \nu(f(A) \cap f(B)) - \nu(f(A))$$

encodes set containment, since

$$f(A) \subseteq f(B) \quad \text{if and only if} \quad \nu(f(A) \cap f(B)) = \nu(f(A)).$$

If, as in Remark 2, we were training using KL-divergence of conditional probability distributions, our loss objectives would involve terms such as

$$\frac{\nu\left(f(\text{baskets}(\texttt{milk})) \cap f(\text{baskets}(\texttt{bread}))\right)}{\nu\left(f(\text{baskets}(\texttt{bread}))\right)}$$

(or, more typically, the $\log$ of such an expression).

As defined in Section 4.1 L is a function from $\mathbf{Y}^{\mathbf{X}}$ to $\mathbb{R}_{\geq 0}$, and so is each $h \in H$, however in order to optimize L we plan to perform gradient descent on $L \circ \psi \colon W \to \mathbb{R}_{\geq 0}$ for some open set $W \subseteq \mathbb{R}^m$ for some m.[11] We can thus consider the functions $\widetilde{h} = h \circ \psi \colon W \to \mathbb{R}$, and denote $\widetilde{H}$ the set of all such functions. Ideally we might hope that the functions in $\widetilde{H}$ are convex, but in general this may be too stringent. At the very least, we would like $\widetilde{h} \in \widetilde{H}$ to be differentiable, with nonzero gradient almost everywhere[12] which is not a global minimum.

Desiderata 3 (Differentiability)**.** The set of training functions on parameters $\widetilde{H}$ have nonzero gradient almost everywhere they are not at a global minimum.

This requirement links the parameterization and measure. For the pairwise intersection case in (3), this would mean $\nu(\varphi(a) \cap \varphi(b))$ should have nonzero gradient for all settings of parameters for which the measure of intersection is $> \varepsilon$. Other training instances may present evidence that sets be disjoint, and thus in general we would like $(a, b) \mapsto \nu(\varphi(a) \cap \varphi(b))$ to have nonzero derivative almost everywhere in Ω^2.

Example 5 (Differentiability of Intersection Volume for Rays)**.** If $\mathbf{Y} = \mathbf{R}$, $\varphi(a) = [a, \infty)$, and $\nu(U) = \int_U e^{-x}\, d\lambda(x)$ then

$$\nabla\, \nu(\varphi(a) \cap \varphi(b)) = \nabla \int_{[\max(a,b)]}^{\infty} e^{-x}\, du = \begin{cases} (-e^{-a}, 0) & \text{if} \quad a > b, \\ (0, -e^{-b}) & \text{if} \quad a < b. \end{cases}$$

and thus $\nu(\varphi(a), \varphi(b))$ is differentiable with nonzero derivative almost everywhere in $\mathbb{R}^2$.

Example 6 (Zero Gradient for Volume of Intersection of Intervals)**.** If $\mathbf{Y} = \mathbf{I}$, $\varphi(a, b) = [a, b]$, and $\nu = \lambda$ (Lebesgue measure), then

$$\nabla\, \nu(\varphi(a, b) \cap \varphi(c, d)) = (0, 0, 0, 0) \quad \text{if} \quad b < c \quad \text{or} \quad d < a,$$

in other words the gradient of the volume for two disjoint intervals is zero. These issues will be addressed in Section 6.2.

[11] Here ψ parameterizes $\mathbf{Y}^{\mathbf{X}}$ by way of the parameterization $\varphi \colon \Omega \to \mathbf{Y}$, as described at the end of Section 4.1.

[12] Here, *almost everywhere* is meant in the technical sense, i.e. the set of parameters for which $\widetilde{h}$ is not at a global minimum but the gradient is zero or does not exist has Lebesgue measure zero.

Remark 3. Another common set of functions in $\widetilde{H}$ include pairwise joint and conditional probabilities, which we can model if ν is a finite measure. In this case we often use KL-divergence as our loss function, and thus the functions in $\widetilde{H}$ would be constant multiples of terms such as

$$\log\big(\nu\big(\varphi(x)\cap\varphi(y)\big)\big),\quad \log\big(\nu\big(\varphi(x)\cap\varphi(y)^c\big)\big),\quad \log\big(\nu\big(\varphi(x)^c\cap\varphi(y)^c\big)\big)$$

or

$$\log\big(\nu\big(\varphi(x)\cap\varphi(y)\big)\big)-\log\big(\nu\big(\varphi(x)\big)\big),\quad \log\big(\nu\big(\varphi(x)^c\cap\varphi(y)\big)\big)-\log\big(\nu\big(\varphi(x)\big)\big).$$

Even in situations where the gradient is nonzero almost everywhere, it may present other issues for training. For instance, in Example 5 the gradient is discontinuous, and in each region where it is continuous only one of the parameters receives an update. In general, the discrete assignment of elements to a set presents challenges for gradient descent based learning, which will be addressed in the following section.

5 Fuzzy Set Representation Learning

What has been described thus far relies on "crisp" sets. Although we may not have explicit access to the sets in $\mathbf{X}$, Definition 2 requires that $\mathbf{X}$ is a set of subsets in some universe $\mathcal{U}_{\mathbf{X}}$. On the other hand, concepts rarely adhere to such crisp boundaries in the real-world. For example, the "set" of punk rock albums is not really a set, as the classification of musical genres is not exact and somewhat subjective. A reasonable approach might be to label each album with the percentage of people who consider it to be punk rock. More generally, the primary objective in many machine learning tasks is to generalize from the training data through some soft notion of "similarity", and the "set" of items "similar" to another is typically not crisp, and must be expressed using graded membership. These examples precisely embody the notion of a membership function of a fuzzy set.

5.1 Background - Fuzzy Sets

Fuzzy sets, introduced in Zadeh [1965], are the mathematical "softening" of sets, and come with a corresponding generalization of standard set-theoretic operations.[13] As mentioned in Section 2, given some universe $\mathcal{U}$, the set of subsets $\mathcal{P}(\mathcal{U})$ is in bijection with the set of functions from $\mathcal{U}$ to $\{0,1\}$ via their characteristic function,

$$\mathbb{1}_S\colon \mathcal{U}\to\{0,1\}\quad\text{where}\quad \mathbb{1}_S(x)=\begin{cases}1 & \text{if } x\in S,\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

The definition of fuzzy sets simply broadens this to the set of functions from $\mathcal{U}$ to $[0,1]$, which we denote $[0,1]^{\mathcal{U}}$. Formally, therefore, a *fuzzy set* is a tuple $(m,\mathcal{U})$, with some universe set $\mathcal{U}$ and a *membership function* $m:\mathcal{U}\to[0,1]$. We will also often refer to the membership function itself as a fuzzy set when the universe is clear from context.

Example 7 (Random Intervals)**.** Let X and Y be real-valued random variables, then

$$m_{[X,Y]}(x) \coloneqq P(x>X)P(x<Y)$$

is a fuzzy set.

Since $\{0,1\}^{\mathcal{U}}\subset[0,1]^{\mathcal{U}}$, traditional sets are also fuzzy sets, precisely those whose membership function has range $\{0,1\}$. When this distinction is important, we will refer to $\{0,1\}^{\mathcal{U}}$ as *crisp sets*.

5.1.1 Fuzzy Set Operations

Note that, for crisp sets, the set-theoretic complement can be lifted to a function on $\{0,1\}^{\mathcal{U}}$, for example

$$\mathbb{1}_{A^c}(x)=1-\mathbb{1}_A(x).$$

This can be extended to a function on $[0,1]^{\mathcal{U}}$, which provides a notion of *fuzzy intersection*. Similarly, the set-theoretic intersection and union can be given as

$$\mathbb{1}_{A\cap B}(x)=\min(\mathbb{1}_A(x),\mathbb{1}_B(x)),\quad\text{and}\quad \mathbb{1}_{A\cup B}(x)=\max(\mathbb{1}_A(x),\mathbb{1}_B(x)).$$

These can also be generalized to binary operations on $[0,1]^{\mathcal{U}}$ to give a notion of fuzzy intersection and fuzzy union, and furthermore these generalizations obey De Morgan's law, however they are not the only operations to do so.

Potential complement operations can be given by a *negator* $\eta:[0,1]\to[0,1]$ for which, at a minimum, we require:

[13] We present only a cursory overview of the relevant properties of fuzzy sets, for a more thorough treatment see Bede [2013].

Axiom η1. (Boundary Condition) $\eta(0) = 1, \eta(1) = 0$

Axiom η2. (Non-Decreasing) If $a \leq b$ then $\eta(a) \geq \eta(b)$

In this case, the fuzzy set complement of $A \in [0,1]^{\mathcal{U}}$ is $\eta \circ A \in [0,1]^{\mathcal{U}}$. For example, the *Yager class of fuzzy complements* is given by $\eta(x) = (1 - x^w)^{1/w}$ for $w \in (0, \infty)$. When $w = 1$, we obtain the *standard negation*.

Intersections can be generalized through the notion of a *t-norm*, which is a function $\top : [0,1]^2 \to [0,1]$ with the following properties:

Axiom $\top$1. (Boundary Condition) $\top(a, 1) = a$

Axiom $\top$2. (Non-Decreasing) If $b \leq c$, $\top(a, b) \leq \top(a, c)$

Axiom $\top$3. (Commutative) $\top(a, b) = \top(b, a)$

Axiom $\top$4. (Associative) $\top(a, \top(b, c)) = \top(\top(a, b), c)$

In this case, the intersection of fuzzy sets $A, B \in [0,1]^{\mathcal{U}}$ is given by $x \mapsto \top(A(x), B(x))$. Classic examples of t-norms include $\top(a, b) = \min(a, b)$ (which induces an intersection on fuzzy sets known as the *standard intersection*) and the *product t-norm* $\top(a, b) = ab$.

A function $\perp : [0,1]^2 \to [0,1]$ which obeys axioms $\top$2 - $\top$4 but has boundary condition $\perp(a, 0) = a$ is called a *t-conorm*. Given a t-conorm $\perp$, the union of fuzzy sets $A, B \in [0,1]^{\mathcal{U}}$ is given by $x \mapsto \perp(A(x), B(x))$. Examples include $\perp(a, b) = \max(a, b)$ (which induces a union known as the *standard union*) and the *probabilistic sum t-conorm*, $\perp(a, b) = a + b - ab$.

The operations $(\top, \perp, \eta)$ induce fuzzy set operations

$$\begin{aligned} \cap : [0,1]^{\mathcal{U}} \times [0,1]^{\mathcal{U}} \to [0,1]^{\mathcal{U}}, \quad (A \cap B)(x) &= \top(A(x), B(x)), \\ \cup : [0,1]^{\mathcal{U}} \times [0,1]^{\mathcal{U}} \to [0,1]^{\mathcal{U}}, \quad (A \cup B)(x) &= \perp(A(x), B(x)), \\ \complement : [0,1]^{\mathcal{U}} \to [0,1]^{\mathcal{U}}, \quad \complement(A)(x) &= \eta(A(x)). \end{aligned}$$

Ideally, we would prefer to select $(\top, \perp, \eta)$ such that the corresponding fuzzy set operations obey De Morgan's law. This is achieved if and only if $\perp(a, b) = \eta(\top(\eta(a), \eta(b)))$ for all $a, b \in [0, 1]$, in which case $\top$ and $\perp$ are called *dual with respect to* η, and we call $(\top, \perp, \eta)$ a *De Morgan triplet*.

The following choices of t-norm and t-conorms form a De Morgan triplet when used with the standard negation, $\eta(x) = 1 - x$.

Example 8 (Zadeh [1965])**.** Known as *Gödel's t-norm and t-conorm*, $\top(a, b) = \min(a, b)$ and $\perp(a, b) = \max(a, b)$ induce the standard intersection and union as originally proposed by Zadeh.

Example 9 (Fodor and Roubens [1994])**.** The product and probabilistic sum, $\top(a, b) = ab$ and $\perp(a, b) = a + b - ab$, are also known as *Goguen's t-norm and t-conorm*.

Example 10 (Hájek [2013])**.** The *Łukasiewicz t-norm and t-conorm*, $\top(a, b) = \max(a + b - 1, 0)$ and $\perp(a, b) = \min(a + b, 1)$, are also often used in fuzzy logic.

There are many alternative De Morgan triplets which could be considered. While the general framework we propose makes no restriction as to which negator, t-norm, or t-conorms are used, in the examples to come we will make use of the product and probabilistic sum, as the lack of hard max and min functions make them most amenable to training via gradient descent.

5.1.2 Fuzzy Algebras

We also generalize the notion of an algebra and homomorphisms between algebras to fuzzy sets. For convenience, we denote the "empty fuzzy set" using $\mathbb{1}_\emptyset$ and the "universe fuzzy set" with $\mathbb{1}_{\mathcal{U}}$. Given a universe $\mathcal{U}$ and some fuzzy set operations $(\cap, \cup, \complement)$, we call a collection of fuzzy sets $\mathcal{F} \subseteq [0,1]^{\mathcal{U}}$ a *fuzzy algebra over $\mathcal{U}$ with respect to* $(\cap, \cup, \complement)$ if it:

1. Contains the empty set ($\mathbb{1}_\emptyset \in \mathcal{F}$).
2. Is closed under complements (for all $A \in \mathcal{F}, \complement(A) \in \mathcal{F}$).
3. Is closed under binary intersections (for all $A, B \in \mathcal{F}, A \cap B \in \mathcal{F}$).
4. Is closed under binary unions (for all $A, B \in \mathcal{F}, A \cup B \in \mathcal{F}$).

Note that the boundary condition for negators means that 1 and 2 imply that $\mathbb{1}_{\mathcal{U}} \in \mathcal{F}$. Furthermore, if the fuzzy set operations $(\cap, \cup, \complement)$ were induced from a De Morgan triplet, then 4 is implied by 2 and 3. As with crisp sets, repetition implies that $\mathcal{F}$ is closed with respect to any finite combination of these operations, and we denote $\mathcal{A}(\mathcal{F})$ to be the smallest fuzzy algebra containing $\mathcal{F}$. If $\mathcal{F}$ contains only crisp sets, these definitions are equivalent to those given in Section 3.3, regardless of the choice of fuzzy set operations.

Definition 4 (Homomorphism of Fuzzy Algebras)**.** Let $\mathcal{G} \subseteq [0,1]^{\mathcal{U}_\mathcal{G}}$ be a fuzzy algebra with respect to $(\cap_\mathcal{G}, \cup_\mathcal{G}, \complement_\mathcal{G})$, and $\mathcal{F} \subseteq [0,1]^{\mathcal{U}_\mathcal{F}}$ a fuzzy algebra with respect to $(\cap_\mathcal{F}, \cup_\mathcal{F}, \complement_\mathcal{F})$. A **homomorphism of fuzzy algebras** from $(\mathcal{G}, \cap_\mathcal{G}, \cup_\mathcal{G}, \complement_\mathcal{G})$ to $(\mathcal{F}, \cap_\mathcal{F}, \cup_\mathcal{F}, \complement_\mathcal{F})$ is an injective function $f : \mathcal{G} \hookrightarrow \mathcal{F}$ such that for all $A, B \in \mathcal{G}$:

1. $f(A \cap_\mathcal{G} B) = f(A) \cap_\mathcal{F} f(B)$
2. $f(A \cup_\mathcal{G} B) = f(A) \cup_\mathcal{F} f(B)$
3. $f(\mathbb{1}_\emptyset) = \mathbb{1}_\emptyset$
4. $f(\mathbb{1}_{\mathcal{U}_\mathcal{G}}) = \mathbb{1}_{\mathcal{U}_\mathcal{F}}$
5. $f(\complement_\mathcal{G}(A)) = \complement_\mathcal{F}(f(A))$

This is quite similar to Definition 1, with the exception that 5 is not implied by 1-4.

5.1.3 Measure of Fuzzy Sets

Now that we have a notion of a fuzzy set, we need to define the notion of the *measure* of a fuzzy set, which can be accomplished whenever the universe has an associated measure.

Definition 5. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and suppose (m, Ω) is a fuzzy set. If m is μ-measurable, then we say the fuzzy set (m, Ω) is measurable, with measure given by $\int_\Omega m(x)\, d\mu$..

Example 11 (Measure of Random Interval)**.** If X and Y are real-valued random variables then the fuzzy set $m_{[X,Y]}$ as defined in Example 7 is Lebesgue measurable with measure

$$\int_\mathbb{R} m_{[X,Y]}(u)\, d\lambda(u) = \int_\mathbb{R} P(u > X)P(u < Y)\, du = \mathbb{E}[P(u > X)P(u < Y)].$$

Given any function $f : \mathcal{G} \to \mathcal{F}$ between fuzzy algebras, we can measure how far f is from being a valid homomorphism of fuzzy algebras by considering the measure of the error. For example, if Δ is the fuzzy symmetric difference derived from the corresponding fuzzy set operations, consider

$$E_1 = \{f(A \cap_\mathcal{G} B)\Delta_\mathcal{F}(f(A) \cap_f f(B)) : A, B \in \mathcal{G}\}$$

and likewise consider $E_2, \ldots, E_5$ for each requirement in Definition 4. The set $\{\nu(S) \mid S \in E_i\}$ contains the measure of the errors. If, as requested in Desiderata 3, these values are differentiable with respect to the parameterization of $\mathbf{Y}$ then we could train a function which approximates a homomorphism of fuzzy sets. As in the crisp set case, however, these sets may be arbitrarily large, and thus it is again of interest to consider learning a function defined on a subset of elements which can be extended to the rest of the algebra via fuzzy set operations.

5.2 Fuzzy Set Embeddings

We now come to the main result which will allow us to satisfy the various desiderata discussed in Section 4 by generalizing Definition 2 from crisp sets to fuzzy sets.

Definition 6 (Fuzzy Set Embedding)**.** Suppose $\mathbf{X}$ is a set of fuzzy sets in $\mathcal{U}_\mathbf{X}$ with fuzzy set operations $(\cap_\mathbf{X}, \cup_\mathbf{X}, \complement_\mathbf{X})$, and $\mathbf{Y}$ is a set of fuzzy sets in $\mathcal{U}_\mathbf{Y}$ with fuzzy operations $(\cap_\mathbf{Y}, \cup_\mathbf{Y}, \complement_\mathbf{Y})$. Then a **fuzzy set embedding of $\mathbf{X}$ in $\mathbf{Y}$ (with respect to $(\cap_\mathbf{X}, \cup_\mathbf{X}, \complement_\mathbf{X})$ and $(\cap_\mathbf{Y}, \cup_\mathbf{Y}, \complement_\mathbf{Y})$)** is an injective function $f : \mathbf{X} \hookrightarrow \mathbf{Y}$ which can be extended to a homomorphism of fuzzy algebras. That is, there exists a homomorphism $F : \mathcal{A}(\mathbf{X}) \hookrightarrow \mathcal{A}(\mathbf{Y})$ such that $F \mid_\mathbf{X} = f$.

As fuzzy sets are generalizations of crisp sets, all previous examples still apply to this definition. This definition also includes settings where $\mathbf{X}$ is a set of crisp sets, and $\mathbf{Y}$ is fuzzy. There are benefits to using fuzzy sets for representation, even if the original space is crisp, as the soft membership can result in loss functions with smoother gradients, as we will demonstrate in the following section.

6 Gumbel Fuzzy Sets

In this section we will describe several collections of fuzzy sets $\mathbf{Y}$ that preserve all the previous desiderata, adapted to fuzzy sets:

1. **Parameterizable**: The set $\mathbf{Y}$ should be capable of being parameterized via a surjective map $\varphi : \Omega \to \mathbf{Y}$ for some open set $\Omega \subseteq \mathbb{R}^d$.
2. **Measurability**: There exists a measure space $(\mathcal{U}_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}}, \nu)$ such that all fuzzy sets in $\mathbf{Y}$ are ν-measurable.
3. **Differentiability of Set Operations**: A reasonable set of training functions on parameters $\widetilde{H}$ have nonzero gradient almost everywhere they are not at a global minimum.

The random intervals described in Example 7 give us some idea of how to proceed, however the requirement that the random variables of the endpoints be parameterizable in such a way as to preserve differentiability of relevant set operations is non-trivial, particularly when we take into account the need to actually *compute* the derivatives. The use of Gumbel variables, as proposed in Dasgupta et al. [2020], provides a solution due to their min/max stability properties.

6.1 Background - Gumbel Distributions

The Gumbel distribution comes in min and max variants, with probability density functions

$$f_{\max}(x;\mu,\beta) = \tfrac{1}{\beta}\exp\Big(-\tfrac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}\Big), \qquad \text{and} \qquad f_{\min}(x;\mu,\beta) = \tfrac{1}{\beta}\exp\Big(\tfrac{x-\mu}{\beta} - e^{\frac{x-\mu}{\beta}}\Big). \tag{7}$$

We denote LogSumExp with temperature β by

$$\mathrm{LSE}(x_1,\ldots,x_n;\beta) \coloneqq \beta\log\left(\sum_{i=1}^{n}\exp\left(\frac{x_i}{\beta}\right)\right)$$

Min (resp. Max) Gumbel distributions are called min (resp. max) stable due to the following:

Proposition 2.

- *If* $X \sim \mathrm{GumbelMin}(\mu_X,\beta), Y \sim \mathrm{GumbelMin}(\mu_Y,\beta)$ *then*
$$\min(X,Y) \sim \mathrm{GumbelMin}(\mathrm{LSE}(\mu_X,\mu_Y;-\beta)),\beta).$$
- *If* $X \sim \mathrm{GumbelMax}(\mu_X,\beta), Y \sim \mathrm{GumbelMax}(\mu_Y,\beta)$ *then*
$$\max(X,Y) \sim \mathrm{GumbelMax}(\mathrm{LSE}(\mu_X,\mu_Y);\beta),\beta).$$

This stability property has implications for the analytic tractability of computing the measure of random intervals with Gumbel endpoints, as we describe in the next section.

6.2 Gumbel Intervals

We begin, as before, in the simplest setting, where $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}$, and thus $\mathbf{Y} \subseteq [0,1]^{\mathbb{R}}$.

For a given $\beta \in \mathbb{R}_+$ and $x^-,^+ \in \mathbb{R}$, let

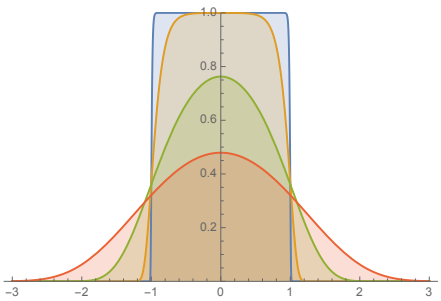
$$X^-(x^-,\beta) \sim \mathrm{GumbelMax}(x^-,\beta) \quad \text{and} \quad X^+(x^+,\beta) \sim \mathrm{GumbelMin}(x^+,\beta).$$

(For notational simplicity, we will drop the explicit dependence on parameters x^-, x^+, β when it is clear from context, i.e. $X^- = X^-(x^-,\beta)$.) We then consider $\mathbf{Y}$ to be the fuzzy sets on $\mathbb{R}$ given by

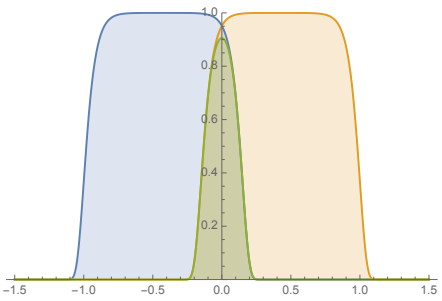
$$m(x;x^-,x^+,\beta) = P(x > X^-)P(x < X^+). \tag{8}$$

For a fixed $\beta > 0$, we can parameterize $\mathbf{Y}$ with $\mathbb{R}^2$ using $\varphi_\beta(x^-,x^+) = m(x;x^-,x^+,\beta)$. If we set $\nu = \lambda$, Lebesgue measure, Dasgupta et al. [2020] proves that we can approximate

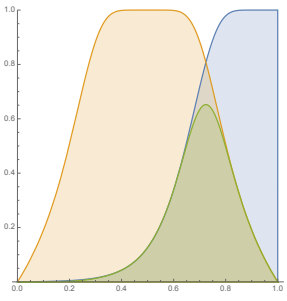
$$\int_{\mathbb{R}} m(x;x^-,x^+,\beta)\,d\lambda(x) = 2\beta K_0\left(2\exp\left(-\frac{x^+-x^-}{2\beta}\right)\right) \approx \beta\log\left(1+\exp\left(\frac{x^+-x^-}{\beta}-2\gamma\right)\right) \tag{9}$$



(a) Gumbel intervals $m(x; -1, 1, \beta)$ for $\beta = 0.01, 0.1, 0.5, 1$. Note that as $\beta \to 0$, we have that $m(x; a^-, a^+, \beta) \to \mathbb{1}_{[a^-, a^+]}(x)$, i.e. the Gumbel interval approaches the crisp interval.



(b) Fuzzy intersection of Gumbel intervals $m(x; -1, 0.15, 0.05)$ and $m(x; -0.15, 1, 0.05)$ is approximately $m(x; -0.15, 0.15, 0.05)$.



(c) Fuzzy intersection of bounded Gumbel intervals $m(x; 4, 14, 1)$ and $m(x; -3, 3, 1)$ with "universe" box $m(x; -5, 5, 1)$.

Figure 1: Membership functions for various Gumbel intervals

where K_0 is the modified Bessel function of the second kind, and γ is the Euler-Mascheroni constant. Moreover, with product as fuzzy intersection, Proposition 2 implies

$$\begin{aligned} m(x; a^-, a^+, \beta) m(x; b^-, b^+, \beta) &= P(x > A^-) P(x < A^+) P(x > B^-) P(x < B^+) \\ &= P(x > \max(A^-, B^-)) P(x < \min(A^+, B^+)) \\ &= m(x; \mathrm{LSE}(a^-, b^-; \beta), \mathrm{LSE}(a^+, b^+; -\beta), \beta). \end{aligned}$$

Thus, if $\widetilde{H}$ includes measures of pairwise intersections (as it often does), we can approximate this as

$$\nu(\varphi_\beta(a^-, a^+) \cap \varphi_\beta(b^-, b^+)) \approx \beta \log\left(1 + \exp\left(\frac{\mathrm{LSE}(a^+, b^+; -\beta) - \mathrm{LSE}(a^-, b^-; \beta)}{\beta} - 2\gamma\right)\right), \tag{10}$$

a smooth function with nonzero gradient everywhere in $\mathbb{R}^4$ (see Appendix B). The same holds for the other training functions we have discussed, as well as the measure of n-ary fuzzy intersections, unions, and arbitrary combinations thereof.

Thus, Gumbel intervals are a natural choice for training fuzzy set representations in $\mathbb{R}$. As mentioned previously, even if $\mathbf{X}$ is comprised of crisp sets, the fact that the training objective for pairwise interactions is smooth with nonzero gradient everywhere is a great training benefit. Furthermore, as depicted in Section 6.2, as $\beta \to 0$ the fuzzy sets approach the indicator function for crisp intervals (see Appendix C). For this reason, given $\beta \in \mathbb{R}_{\geq 0}$ we denote the set of Gumbel intervals as

$$\mathbf{G}(\beta) := \left\{ (m, \mathbb{R}) : m(x; x^-, x^+, \beta) = \begin{cases} P(x > X^-) P(x < X^+) & \text{if } \beta > 0, \\ \mathbb{1}_{[x^-, x^+]}(x) & \text{if } \beta = 0. \end{cases} \right\} \tag{11}$$

In particular, $\mathbf{G}(0) = \mathbf{I}$, and therefore Gumbel intervals are a generalization of intervals.

In addition, Boratko et al. [2021a] proves that if we desire a finite measure (so that the measure can be interpreted probabilistically), we can take $\nu(U) = \int_U m(x; u^-, u^+, \beta)\, d\lambda$ for some fixed "universe" Gumbel interval $m(\,\cdot\,; u^-, u^+, \beta)$, in which case we have that the measure of a fuzzy set in $m(\,\cdot\,; x^-, x^+, \beta) \in \mathbf{G}(\beta)$ is given by

$$\int_{\mathbb{R}} m(x; x^-, x^+, \beta) m(x; u^-, u^+, \beta)\, d\lambda(x),$$

and once again we will find that finite intersections, unions, negations, and arbitrary combinations thereof can be approximated by a smooth function with nonzero gradient everywhere.

6.3 Gumbel Boxes

We have thus far assumed that $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}$, primarily for simplicity. We started with rays, which provided the simplest example of a set-based representation, moved to intervals due to their increased representational capacity, and then generalized these to fuzzy sets by modeling the endpoints of intervals with Gumbel random variables. In this section we will demonstrate one way to easily increase representational capacity further using a product space. We will apply

this process to Gumbel intervals, resulting in a representation we call *Gumbel boxes*, an explicit parameterization of fuzzy sets in $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}^d$.

In general, given some universes $\{\mathcal{U}_{\mathbf{Y}_i}\}_{i=1}^d$, we can consider the product space

$$\mathcal{U}_{\mathbf{Y}} = \prod_{i=1}^{d} \mathcal{U}_{\mathbf{Y}\,i} = \{(u_1, \ldots, u_d) : u_i \in \mathcal{U}_{\mathbf{Y}\,i}\}.$$

Given some associated collection of fuzzy sets $\{\mathbf{Y}_i\}_{i=1}^d$, we can take $\mathbf{Y}$ to be the fuzzy sets

$$\mathbf{Y} = \{(m, \mathcal{U}_{\mathbf{Y}}) : \exists m_i \in \mathbf{Y}_i \quad \text{such that} \quad m(\mathbf{u}) = \prod_{i=1}^{d} m_i(u_i)\}.$$

If each $\mathbf{Y}_i$ can be parameterized with $\varphi_i : \Omega_i \to \mathbf{Y}_i$, then we can parameterize $\mathbf{Y}$ using $\varphi : \prod_{i=1}^d \Omega_i \to \mathbf{Y}$ defined as

$$\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_d)(u) = \prod_{i=1}^{d} \varphi_i(\mathbf{x}_i)(u_i).$$

Finally, if we have a measure space $(\mathcal{U}_{\mathbf{Y}_i}, \mathcal{F}_i, \nu_i)$ for each i such that all $m \in \mathbf{Y}_i$ are ν_i-measurable, then we can use a product measure $\nu : \mathcal{F} \to [0, \infty]$, where

$$\mathcal{F} = \bigotimes_{i=1}^{d} \mathcal{F}_i = \sigma\left(\{A_1 \times \cdots \times A_d : A_i \in \mathcal{F}_i\}\right)$$

and ν is some measure such that

$$\nu(A_1 \times \cdots \times A_d) = \left(\prod_{i=1}^{d} v_i\right)(A_1 \times \cdots \times A_d) = \prod_{i=1}^{d} v_i(A_i). \tag{12}$$

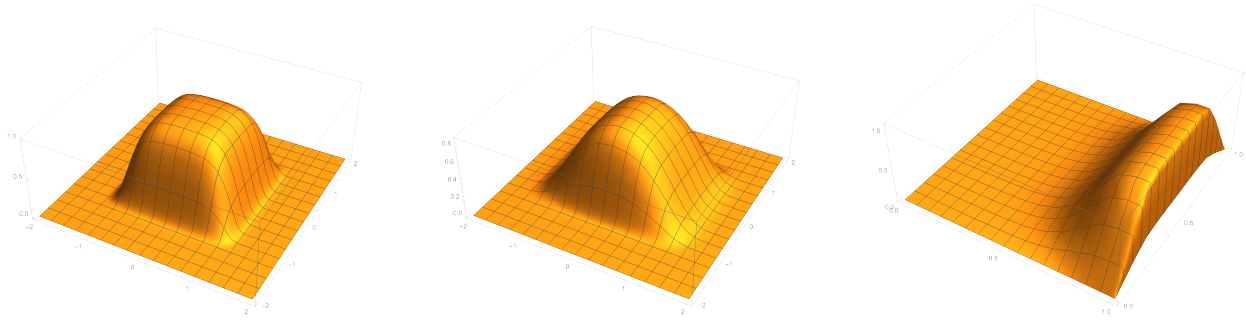
If ν_i are σ-finite[14] then (12) defines ν uniquely, and any $m \in \mathbf{Y}$ is ν-measurable, with measure

$$\int_{\mathcal{U}_{\mathbf{Y}}} m \, d\,\nu = \prod_{i=1}^{d} \int_{\mathcal{U}_{\mathbf{Y}_i}} m_i \, d\,\nu_i \,.$$

Applying this construction to Gumbel intervals $\mathbf{G}(\beta)$, given $\mathbf{x}^-, \mathbf{x}^+ \in \mathbb{R}^d$, and $\boldsymbol{\beta} \in \mathbb{R}^d_{\geq 0}$ we find the membership function of a Gumbel box $M_d : \mathbb{R}^d \to [0, 1]$ to be

$$M_d(\mathbf{x}; \mathbf{x}^-, \mathbf{x}^+, \boldsymbol{\beta}) \coloneqq \prod_{i=1}^{d} m(x_i; x_i^-, x_i^+, \beta_i) = \prod_{i=1}^{d} P(x_i > X_i^-) P(x_i < X_i^+).$$

[14]Given a measure space (ω, Σ, μ) we say μ is $sigma$-finite if we can write $\Omega = \bigcup_{i=1}^{\infty} U_i$ such that $\mu(U_i) < \infty$ for all i. Lebesgue measure on $\mathbb{R}$ is σ-finite, as $\mathbb{R} = \bigcup_{i=1}^{\infty} [-i, i]$, and $\lambda([-i, i]) = 2i < \infty$.



(a) Membership function for Gumbel box in $\mathbb{R}^2$:

$M_2(\mathbf{x}; (-1,-1),(1,1),(0.15,0.15))$

(b) Membership function for Gumbel box in $\mathbb{R}^2$ with different β_i:

$M_2(\mathbf{x}; (-1,-1),(1,1),(0.35,0.15))$

(c) Gumbel box with membership function

$M_2(\mathbf{x}; (3,-4),(9,4),(1,1))$

using universe box

$M_2(\mathbf{x}; (-5,-5),(5,5),(1,1))$

Figure 2: Membership functions for Gumbel boxes in $\mathbb{R}^2$

We denote the collection of all such fuzzy sets as $\mathbf{G}_d(\boldsymbol{\beta})$, which includes the Gumbel intervals as $\mathbf{G}_1(\beta) = \mathbf{G}(\beta)$. If we use $\nu = \lambda$, Lebesgue measure, then we can calculate the measure of a Gumbel box as

$$\int_{\mathbb{R}^d} M_2(\mathbf{x}; \mathbf{x}^-, \mathbf{x}^+, \boldsymbol{\beta})\, d\lambda^d(\mathbf{x}) = \prod_{i=1}^{d} \int_{\mathbb{R}} m(x_i; x_i^-, x_i^+, \beta_i)\, d\lambda(x_i).$$

The measure of pairwise intersections becomes a product of terms like Equation (10), and thus is smooth with nonzero derivative almost everywhere. This also holds if ν_i is given by some fixed universe box, in which case the product measure is also a valid probability measure,and thus the resulting model can be interpreted probabilistically [Boratko et al., 2021b].

As promised, we can prove that Gumbel boxes in $\mathbb{R}^{d+1}$ can be shown to have strictly more representational capacity than Gumbel boxes in $\mathbb{R}^d$.

Proposition 3. *For any $\boldsymbol{\beta} \in \mathbb{R}^d_{\geq 0}$, $\mathbf{G}_d(\boldsymbol{\beta})$ has strictly less representational capacity than $\mathbf{G}_{d+1}(\boldsymbol{\beta}, 0)$.*

A complete proof can be found in Appendix D, but we provide a quick sketch here. A mapping from $\mathbf{G}_d(\boldsymbol{\beta})$ to $\mathbf{G}_{d+1}(\boldsymbol{\beta}, 0)$ can be provided by fixing the width of the Gumbel boxes in the new dimension to be equal (say, $[0,1]$). To show that no such mapping works in reverse, we use the fact that this new dimension allows for additional disjointness, and create an example which entwines this disjointness pattern in such a way which prevents it from being possible to embed in the original space.[15]

Gumbel boxes have been employed in a number of tasks, including fine-grained entity typing [Onoe et al., 2021], uncertain knowledge graph embedding [Chen et al., 2021], and probabilistic modeling [Boratko et al., 2021b]. As expected, they have proven to be far easier to train than previous methods, even in situations where $\mathbf{X}$ has no inherent "fuzzyness". The $\boldsymbol{\beta}$ parameter can be chosen as a hyperparameter, or trained via gradient descent [Boratko et al., 2021a].

6.4 Gumbel Box Mixtures

Another simple way to increase representational capacity is to consider a disjoint union of universes

$$\mathcal{U}_{\mathbf{Y}} = \bigsqcup_{i=1}^{n} \mathcal{U}_{\mathbf{Y}_i} = \{(u,i) : u \in \mathcal{U}_{\mathbf{Y}_i}, i \in \{1,\dots,n\}\}.$$

Given some set of fuzzy sets $\{(m_i, \mathcal{U}_{\mathbf{Y}_i})\}_{i=1}^n$ we can form the fuzzy set $(m, \mathcal{U}_{\mathbf{Y}})$ with $m(u,i) = m_i(u)$, thus we can define

$$\mathbf{Y} = \left\{ (m, \mathcal{U}_{\mathbf{Y}}) : m : \mathcal{U}_{\mathbf{Y}} \to [0,1] \quad \text{such that} \quad \forall i \in \{1,\dots,n\}, \quad m\big|_{\mathcal{U}_{\mathbf{Y}_i} \times \{i\}} \in \mathbf{Y}_i \right\}. \tag{13}$$

[15] This construction is motivated by classic results on *boxicity*, which is a well-studied graph characteristic [Roberts, 1969].

Note that we can identify each fuzzy set $m \in \mathbf{Y}$ with the n-tuple $(m_i)_{i=1}^n$ of fuzzy sets such that $m_i = m\big|_{\mathcal{U}_{\mathbf{Y}_i} \times \{i\}}$. Letting $g = (g_i)_{i=1}^n$ and $h = (h_i)_{i=1}^n$ be fuzzy sets in $\mathbf{Y}$, we have that the product is given by $gh : \mathcal{U}_{\mathbf{Y}} \to [0,1]$ where

$$g(u,i)h(u,i) = g_i(u)h_i(u), \quad \text{thus} \quad gh = (g_i h_i)_{i=1}^n.$$

Assuming each $\mathbf{Y}_i$ comes with a parameterization $\varphi_i : \Omega_i \to \mathbf{Y}_i$ we can parameterize $\mathbf{Y}$ by $\varphi : \prod_{i=1}^n \Omega_i \to \mathbf{Y}$ where

$$\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_n)(u,i) = \varphi_i(\mathbf{x}_i)(u),$$

in other words the function $\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_n) : \mathcal{U}_{\mathbf{Y}} \to [0,1]$ is identified with $(\varphi_i(\mathbf{x}_i))_{i=1}^n$, where $\varphi_i(\mathbf{x}_i) : \mathcal{U}_{\mathbf{Y}_i} \to [0,1]$. If each $\mathbf{Y}_i$ comes with a measure space $(\mathcal{U}_{\mathbf{Y}_i}, \mathcal{F}_i, \nu_i)$ for which all $m \in \mathbf{Y}_i$ are ν_i-measurable, we can create the sigma algebra

$$\mathcal{F} = \{S \subseteq \mathcal{U}_{\mathbf{Y}} : \forall i \in \{1, \ldots, n\}, \quad \{u : (u,i) \in S\} \in \mathcal{F}_i\},$$

and define the measure $\nu : \mathcal{F} \to [0, \infty]$ as

$$\nu(S) = \sum_{i=1}^n \nu_i \left(\{u : (u,i) \in S\}\right).$$

Then any $m \in \mathbf{Y}$ will be ν-measurable, with measure

$$\int_{\mathcal{U}_{\mathbf{Y}}} m \, d\, \nu = \sum_{i=1}^n \int_{\mathcal{U}_{\mathbf{Y}_i}} m(u,i) \, d\, \nu_i(u).$$

In particular, we can take n copies of Gumbel Boxes in $\mathbb{R}^d$, which we denote as $\mathbf{G}_d^n(\boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{dn}$. We will index $\boldsymbol{\beta}$ using lower and upper indices, i.e. β_i^j is the temperature in the i^{th} dimension of the j^{th} universe. We will also denote $\boldsymbol{\beta}^j = (\beta_1^j, \ldots, \beta_d^j)$. As each $\mathcal{U}_{\mathbf{Y}_i} = \mathbb{R}^d$, we have that $\mathcal{U}_{\mathbf{Y}} = \mathbb{R}^d \times \{1, \ldots, n\}$.[16]

We can observe the following immediate result regarding representational capacity.

Proposition 4. *Given $\boldsymbol{\beta} \in \mathbb{R}_{\geq 0}^{dn}$, for any $j \in \{1, \ldots, n\}$ we have that $\mathcal{R}(\mathbf{G}_d^n(\boldsymbol{\beta})) \subseteq \mathcal{R}(\mathbf{G}_d^{n+1}(\boldsymbol{\beta}, \boldsymbol{\beta}^j))$, with strict containment if $\beta_i^j = 0$ for any $i \in \{1, \ldots, d\}$.*

Proof. Let $f : \mathbf{G}_d^n(\boldsymbol{\beta}) \hookrightarrow \mathbf{G}_d^n(\boldsymbol{\beta}, \boldsymbol{\beta}^j)$ be defined as

$$f((m_1, \ldots, m_n)) = (m_1, \ldots, m_n, m_j).$$

Then given two fuzzy sets $g = (g_i)_{i=1}^n, h = (h_i)_{i=1}^n \in \mathbf{G}_d^n(\boldsymbol{\beta})$ we have

$$\begin{aligned} f(gh) &= f((g_1 h_1, \ldots, g_n, h_n)) = (g_1 h_1, \ldots, g_n h_n, g_j h_j) \\ &= (g_1, \ldots, g_n, g_j)(h_1, \ldots, h_n, h_j) \\ &= f(g) f(h) \end{aligned}$$

as desired. As $\mathbf{G}_d^n$ does not contain unions or negations for any elements, this is enough to prove that f is a fuzzy set representation. □

In fact, for any set of non-negative weights $w_i \in \mathbb{R}_{\geq 0}$, the function

$$\nu(S; \mathbf{w}) = \sum_{i=1}^n w_i \, \nu_i \left(\{u : (u,i) \in S\}\right)$$

is a valid measure[17] and any $m \in \mathbf{Y}$ is ν-measurable, with measure

$$\int_{\mathcal{U}_{\mathbf{Y}}} m \, d\, \nu = \sum_{i=1}^n w_i \int_{\mathcal{U}_{\mathbf{Y}_i}} m(u,i) \, d\, \nu_i(u).$$

[16] Note that, in general, we could also consider settings where the dimension of each universe is not equal.

[17] This is easily observed by considering the measure ν_i replaced by $\nu_i' = w_i \, \nu_i$.

7 Related Work

Most often, representation learning uses vector representations in $\mathbb{R}^d$ capturing relevant information using vector-based operations (eg. dot-product, distance) in the training objective. Thus, the data is mapped into a real vector space regardless of the structure which this representation is attempting to preserve. This may be due, in part, to the computational efficiency of performing such vector-based operations, allowing such architectures to scale to billions of parameters and train on massive datasets. Furthermore, it is well known that, with enough parameters, these vector-based architectures can perfectly represent their training data [Yun et al., 2019], and thus impressive positive empirical results have been obtained by increasing representational capacity and the amount of training data [Chowdhery et al., 2022]. Training on very large data can ensure that "unseen" queries are in the convex hull and even relatively close to some training data, and changes to the model architectures can capitalize on invariants in the data making the coverage even denser [Fedus et al., 2021], however ultimately the learned representation will embed the data into a real vector space, and be subject to the available operations and limitations of such a space.

For example, we may wish to preserve some partial order $\preceq_{\mathbf{X}}$, in which case we must embed the data in some partially ordered set (poset) [18] $(\mathbf{Y}, \preceq_{\mathbf{Y}})$ such that for all $a, b \in \mathbf{X}$,

$$a \preceq_{\mathbf{X}} b \quad \text{if and only if} \quad f(a) \preceq_{\mathbf{Y}} f(b).$$

A classic example of such a scenario is that of *hypernymy*, i.e. the `IsA` relation, where we wish to capture structure such as "`dog IsA mammal` ". The notion of a partial order embedding was introduced to the machine learning community in this context by Vendrov et al. [2016], in which they proposed the order embedding (OE) model which used $\mathbf{Y} = \mathbb{R}^N_+$ and $\preceq_{\mathbf{Y}} = \preceq_{\mathbb{R}^N_+}$ was the *reversed product order* on $\mathbb{R}^N_+$: for all $u, v \in \mathbb{R}^N_+$,

$$u \preceq_{\mathbf{Y}} v \quad \text{if and only if} \quad \forall i \in \{1, \ldots, N\}, u_i \geq v_i.$$

This is far from being the only choice of partial order on $\mathbb{R}^N_+$ (lexicographic order being another prominent example) however it does have a rather pleasing geometric and set-theoretic interpretation: for a point $u \in \mathbb{R}^N_+$, the set of all points $v \in \mathbb{R}^N_+$ for which $v \preceq_{\mathbb{R}^N_+} u$ forms a *convex cone* [19] which we will denote as $\text{Cone}(u) \coloneqq \{v : v \preceq_{\mathbb{R}^N_+} u\}$. Letting $\mathbf{C} \coloneqq \{\text{Cone}(u) : u \in \mathbb{R}^N_+\}$ be the set of all such cones, then Cone is an *isomorphism* between $(\mathbb{R}^N_+, \preceq_{\mathbb{R}^N_+})$ and $(\mathbf{C}, \subseteq)$, as we have that

$$u \preceq_{\mathbb{R}^N_+} v \quad \text{if and only if} \quad \text{Cone}(u) \subseteq \text{Cone}(v).$$

In the case of hypernymy, this means the cone for `dog` should be contained within the cone for `mammal`, which also aligns with a set-theoretic construction attempting to capture denotational semantics, where every instance of `dog` is also an instance of `mammal`. By composition, one can view any embedding using reversed product order on $\mathbb{R}^N_+$ as an embedding into $(\mathbf{C}, \subseteq)$, however as in the vector case we also have additional structure on $\mathbf{C}$ beyond the $\subseteq$ operation. For example, $\mathbf{C}$ is a π-system, which is a nonempty set of subsets closed under (finite) intersection - for all $A, B \in \mathbf{C}$ we have that $A \cap B \in \mathbf{C}$. As with vector addition, while the intersection operation may yield sensible results in some cases (eg. $\text{Cone}(f(\texttt{dog})) \cap \text{Cone}(f(\texttt{mammal})) = \text{Cone}(f(\texttt{dog}))$) it also exposes incongruencies in the embedded representation (eg. *all* cones have nonempty intersection, even $\text{Cone}(f(\texttt{mammal})) \cap \text{Cone}(f(\texttt{plant})) \neq \emptyset$).

Another example of structure which is often encountered is *probabilistic* structure - where the input space $\mathbf{X}$ is actually a subset of the σ-algebra for some probability measure μ. What is therefore desired is that $\mathbf{Y}$ is a (subset of a) σ-algebra as well, with probability measure ν, and that for all $a \in \mathbf{X}$, $\nu(f(a)) = \mu(a)$. This was first introduced in the probabilistic order embedding (POE) model of Lai and Hockenmaier [2017], which used $\mathbf{Y} = \mathbf{C}$ and $\nu(u) = \int_u e^{-z}\, dz$, the negative exponential measure.[20]

Any probability model also implictly captures a partial order via conditional probability - given some threshold τ, the relation $A \preceq B$ if and only if $P(B \mid A) > \tau$ is a partial order, for instance. Thus, probabilistic models can

[18] A **partially ordered set** or **poset** is a set $\mathbf{S}$ with binary operation $\preceq_{\mathbf{S}}$ which is:

1. **reflexive**: $\forall a \in \mathbf{S}, a \preceq_{\mathbf{S}} a$
2. **antisymmetric**: $\forall a, b \in \mathbf{S}$, if $a \preceq_{\mathbf{S}} b$ and $b \preceq_{\mathbf{S}} a$ then $a = b$
3. **transitive**: $\forall a, b, c \in \mathbf{S}$, if $a \preceq_{\mathbf{S}} b$ and $b \preceq_{\mathbf{S}} c$ then $a \preceq_{\mathbf{S}} c$

[19] A subset C of a real vector space is called a **cone** if it is closed under multiplication by positive reals, i.e. $rC \subseteq C$ for all $r \in \mathbb{R}_+$. A **convex cone** is also closed with respect to vector addition, i.e. $C + C \subseteq C$.

[20] The space $\mathbf{Y}$ for POE is actually a π-system, and the σ-algebra generated by $\mathbf{Y}$ is actually the Borel σ-algebra, i.e. $\sigma(\mathbf{Y}) = \mathcal{B}(\mathbb{R}^N_+)$, which justifies the use of the Lebesgue-Stieltjes measure ν.

also be used to embed partial orders, with the intuitive semantic correspondence that `dog IsA mammal` would be modeled such that $P(\texttt{mammal} \mid \texttt{dog}) = 1$. As with OE, this amounts to the cone for `dog` being contained within the cone for `mammal`, however the limitations from OE now manifest themselves probabilistically in that POE cannot model negative correlation. In fact, conditioning with any additional element will cause the probability to increase, eg. $P(\texttt{dog}) < P(\texttt{dog} \mid \texttt{plant})$.

The representational limitations of OE and POE were addressed by the probabilistic box embedding model (PBE), introduced in Vilnis et al. [2018], which greatly expanded the representational capacity while preserving the probabilistic structure. Denoting the set of intervals in $[0, 1]$ as $\mathbf{I} = \{[a, b] : 0 \leq a < b \leq 1\}$, Vilnis et al. [2018] set $\mathbf{Y} = \mathbf{I}^d =: \mathbf{B}$. That is, the objects in $\mathbf{Y}$ are Cartesian products of intervals, or hyperrectangles, and thus a subset of the Lebesgue σ-algebra on $[0, 1]^d$. Vilnis et al. [2018] use the standard Lebesgue measure on $[0, 1]^d$, in which the measure of each box is simply the product of its side-lengths, thus obtaining an embedding capable of preserving probabilistic structure with greater representational capacity and computational simplicity.

Subsequent work has improved box representation learning, first by approximating a Gaussian convolution of box indicator functions (SMOOTHBOX, Li et al. [2019]) and then by considering a latent random variable approach (GUMBELBOX, Dasgupta et al. [2020]) which smoothed the loss landscape further. These training improvements temporarily dispensed with the probabilistic structure, however this was regained in [Boratko et al., 2021b] which uses $\mathbf{Y} = \mathbb{R}^d$ with a specific probability measure compatible with the GUMBELBOX representations. Box embeddings have also been leveraged to preserve graph structure [Boratko et al., 2021a], and found many applications from modeling uncertain knowledge graphs [Chen et al., 2021], fine-grained entity typing [Onoe et al., 2021], and multi-label classification [Patel et al., 2021].

The primary takeaways from this brief historical tour of structural representation learning is that it is beneficial to pick a representation with structure which matches the space you wish to represent.

- **Too little structure makes it impossible to faithfully capture and generalize from the patterns of structure in the training data**
- **Unnecessary structure in the representation space presents itself as implicit bias or representational limitations.**

References

Christopher Peacocke. *A study of concepts.* The MIT Press, 1992.

R.A. Geiger and B. Rudzka-Ostyn. *Conceptualizations and Mental Processing in Language*. Cognitive linguistics research. Mouton de Gruyter, 1993. ISBN 9783110127140. URL `https://books.google.com/books?id=0O_I9hvI_6QC`.

Edward N Zalta. *A (Leibnizian) theory of concepts*. Citeseer, 2000.

Gottfried Wilhelm Leibniz and Parkinson G H R. *A Study in the Calculus of Real Addition (1690)*. Oxford, 1966.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Alice Lai and Julia Hockenmaier. Learning to predict denotational probabilities for modeling entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 721–730, 2017.

L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(65)90241-X. URL `https://www.sciencedirect.com/science/article/pii/S001999586590241X`.

Barnabas Bede. *Mathematics of Fuzzy Sets and Fuzzy Logic*. Springer Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-35221-8.

Janos C Fodor and MR Roubens. *Fuzzy preference modelling and multicriteria decision support*, volume 14. Springer Science & Business Media, 1994.

Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 2013.

Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. *NeurIPS*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/01c9d2c5b3ff5cbba349ec39a570b5e3-Abstract.html`.

Michael Boratko, Dongxu Zhang, Nicholas Monath, Luke Vilnis, Kenneth Clarkson, and Andrew McCallum. Capacity and bias of learned geometric embeddings for directed graphs. *NeurIPS*, 2021a.

Michael Boratko, Javier Burroni, Shib Sankar Dasgupta, and Andrew McCallum. Min/max stability and box distributions. *UAI*, pages 2146–2155, 2021b.

Fred S Roberts. On the boxicity and cubicity of a graph. *Recent progress in combinatorics*, 1(1):301–310, 1969.

Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. Modeling fine-grained entity types with box embeddings. *ACL*, 2021.

Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. Probabilistic box embeddings for uncertain knowledge graph reasoning. *NAACL*, 2021.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways, 2022. URL `https://arxiv.org/abs/2204.02311`.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv*, 2021.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *ICLR*, 2016.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. *ACL*, 2018.

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. *ICLR*, 2019.

Dhruvesh Patel, Pavitra Dangati, Jay-Yoon Lee, Michael Boratko, and Andrew McCallum. Modeling label space interactions in multi-label classification using box embeddings. *ICLR*, 2021.

A Necessity of Parameterization

Proposition 5. *If $\mathbf{X}$ is finite, there exists a surjection $\psi : W \to \mathbf{Y}^{\mathbf{X}}$ for some $W \subseteq \mathbb{R}^m$ if and only if there exists a surjection $\varphi : \Omega \to \mathbf{Y}$ for some $\Omega \subseteq \mathbb{R}^d$.*

Proof. Suppose $\psi : W \to \mathbf{Y}^{\mathbf{X}}$ is surjective. Recall that when $\mathbf{X}$ is finite, there is a natural bijection $\rho : \mathbf{Y}^{\mathbf{X}} \to \mathbf{Y}^{|\mathbf{X}|}$. Thus, the function

$$\varphi = \pi_1 \circ \rho \circ \psi : W \to \mathbf{Y}$$

is surjective, with $\Omega = W$.

The opposite implication is already mentioned in Section 4.1, but we repeat it here for convenience. Given a surjective function $\varphi : \Omega \to \mathbf{Y}$ where $\Omega \subseteq \mathbb{R}^d$, let

$$g : \Omega^{|\mathbf{X}|} \to \mathbf{Y}^{|\mathbf{X}|} \quad \text{where} \quad g_i(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{|\mathbf{X}|}) = \varphi(\boldsymbol{\theta}_i),$$

then $\psi = \rho^{-1} \circ g : \Omega^{|\mathbf{X}|} \to \mathbf{Y}^{|\mathbf{X}|}$ is surjective, with $W = \Omega^{|\mathbf{X}|} \subseteq \mathbb{R}^{d|\mathbf{X}|}$. □

B Pairwise Intersection Gradient Calculation

The pairwise intersection volume in Equation (10) can be simplified to

$$\begin{aligned}\nu(\varphi_\beta(a^-, a^+) \cap \varphi_\beta(b^-, b^+)) &\approx \beta \log\left(1 + \exp\left(\frac{\mathrm{LSE}(a^+, b^+; -\beta) - \mathrm{LSE}(a^-, b^-; \beta)}{\beta} - 2\gamma\right)\right) \\ &= \beta \log\left(1 + \frac{1}{e^{2\gamma}\left(e^{\frac{a^-}{\beta}} + e^{\frac{b^-}{\beta}}\right)\left(e^{-\frac{a^+}{\beta}} + e^{-\frac{b^+}{\beta}}\right)}\right) =: g(a^-, a^+, b^-, b^+),\end{aligned}$$

for which

$$\begin{aligned}\nabla g(a^-, a^+, b^-, b^+) = \Bigg(&-\frac{e^{\frac{a^-}{\beta}}}{\left(e^{\frac{a^-}{\beta}} + e^{\frac{b^-}{\beta}}\right)\left(\left(e^{\frac{a^-}{\beta}} + e^{\frac{b^-}{\beta}}\right)\left(e^{\frac{a^+}{\beta}} + e^{\frac{b^+}{\beta}}\right)e^{-\frac{a^+ + b^+ - 2\beta\gamma}{\beta}} + 1\right)}, \\ &\frac{e^{\frac{a^+ + 2b^+}{\beta}}}{\left(e^{\frac{a^+}{\beta}} + e^{\frac{b^+}{\beta}}\right)\left(e^{\frac{a^+ + b^- + 2\beta\gamma}{\beta}} + e^{\frac{a^- + b^+ + 2\beta\gamma}{\beta}} + e^{\frac{a^+ + b^+}{\beta}} + e^{\frac{a^- + a^+ + 2\beta\gamma}{\beta}} + e^{\frac{b^- + b^+ + 2\beta\gamma}{\beta}}\right)}, \\ &-\frac{e^{\frac{b^-}{\beta}}}{\left(e^{\frac{a^-}{\beta}} + e^{\frac{b^-}{\beta}}\right)\left(\left(e^{\frac{a^-}{\beta}} + e^{\frac{b^-}{\beta}}\right)\left(e^{\frac{a^+}{\beta}} + e^{\frac{b^+}{\beta}}\right)e^{-\frac{a^+ + b^+ - 2\beta\gamma}{\beta}} + 1\right)}, \\ &\frac{e^{\frac{2a^+ + b^+}{\beta}}}{\left(e^{\frac{a^+}{\beta}} + e^{\frac{b^+}{\beta}}\right)\left(e^{\frac{a^+ + b^- + 2\beta\gamma}{\beta}} + e^{\frac{a^- + b^+ + 2\beta\gamma}{\beta}} + e^{\frac{a^+ + b^+}{\beta}} + e^{\frac{a^- + a^+ + 2\beta\gamma}{\beta}} + e^{\frac{b^- + b^+ + 2\beta\gamma}{\beta}}\right)}\Bigg).\end{aligned}$$

Noting that the numerators and denominators are strictly positive, we see that the gradient is nonzero everywhere.

C Gumbel Intervals Approximate Crisp Intervals

The cumulative distribution functions for Gumbel distributions are

$$F_{\max}(x; \mu, \beta) = \exp(-e^{-(x-\mu)/\beta}) \quad \text{and} \quad F_{\min}(x; \mu, \beta) = 1 - \exp(-e^{(x-\mu)/\beta}).$$

Thus, the membership function for a Gumbel interval, where

$$X^- \sim \mathrm{GumbelMax}(x^-, \beta) \quad \text{and} \quad X^+ \sim \mathrm{GumbelMin}(x^+, \beta)$$

is

$$\begin{aligned} m(x; x^-, x^+, \beta) &= P(x > X^-)P(x < X^+) \\ &= F_{\max}(x; x^-, \beta)(1 - F_{\min}(x; x^+, \beta)) \\ &= \exp(-e^{-(x-x^-)/\beta}) \exp(-e^{(x-x^+)/\beta}) \\ &= \exp(-e^{-(x-x^-)/\beta} - e^{(x-x^+)/\beta}). \end{aligned}$$

Now,

$$e^{-(x-x^-)/\beta} \to \begin{cases} \infty & \text{if } x < x^-, \\ 1 & \text{if } x = x^-, \\ 0 & \text{if } x > x^-, \end{cases} \quad \text{and} \quad e^{(x-x^+)/\beta} \to \begin{cases} 0 & \text{if } x < x^+, \\ 1 & \text{if } x = x^+, \\ \infty & \text{if } x > x^+, \end{cases}$$

as $\beta \to 0^+$. If $x^- < x^+$, then we have the following:

$$\lim_{\beta \to 0^+} m(x; x^-, x^+, \beta) = \begin{cases} 0 & \text{if } x < x^- \text{ or } x > x^+, \\ e^{-1} & \text{if } x = x^- \text{ or } x = x^+, \\ 1 & \text{if } x^- < x < x^+, \end{cases}$$

which is equal to $\mathbb{1}_{[x^-, x^+]}(x)$ almost everywhere.

It is also worth investigating the volume approximation (9) in Section 6.2, to ensure that interpretation of this pointwise limit of membership functions is not lost due to the approximation. Recall

$$\int_{\mathbb{R}} m(x; x^-, x^+, \beta)\, d\lambda(x) \approx \beta \log\left(1 + \exp\left(\frac{x^+ - x^-}{\beta} - 2\gamma\right)\right),$$

so we wish to calculate

$$\lim_{\beta \to 0^+} \beta \log(1 + Ce^{\frac{x}{\beta}})$$

where $C = e^{-2\gamma}$ and $x = x^+ - x^-$. If $x \leq 0$ then this limit is simply 0, but for $x > 0$ this is an indeterminate form $(0 \cdot \infty)$. We convert this to

$$\lim_{\beta \to 0^+} \frac{\log(1 + Ce^{\frac{x}{\beta}})}{\beta^{-1}} = \lim_{\beta \to 0^+} \frac{xCe^{\frac{x}{\beta}}}{1 + Ce^{\frac{x}{\beta}}} = \lim_{\beta \to 0^+} \frac{xCe^{\frac{x}{\beta}}(\frac{-x}{\beta^2})}{Ce^{\frac{x}{\beta}}(\frac{-x}{\beta^2})} = x$$

where the first two equalities follow from an application of L'Hôpital's rule. Hence, we find that

$$\lim_{\beta \to 0^+} \beta \log\left(1 + \exp\left(\frac{x^+ - x^-}{\beta} - 2\gamma\right)\right) = \max(x^+ - x^-, 0),$$

i.e. the size of the interval $[x^-, x^+]$, as desired.

D Proof of Increased Representational Capacity for Gumbel Boxes

Lemma 1. *Suppose A, B, C, D are intervals such that A and B are disjoint, but*

$$C \cap A \neq \emptyset, \quad C \cap B \neq \emptyset, \quad D \cap A \neq \emptyset, \quad \textit{and} \quad D \cap B \neq \emptyset.$$

Then $C \cap D \neq \emptyset$.

Proof. If A and B are disjoint, then either $a^+ < b^-$ or $b^+ < a^-$. WLOG, suppose $a^+ < b^-$. Then due to the intersection requirements for C and D,

$$c^- \leq a^+, \quad c^+ \geq b^-, \quad d^- \leq a^+, \quad \text{and} \quad d^+ \geq b^-,$$

thus

$$[c^-, c^+] \cap [d^-, d^+] \supseteq [a^+, b^-] \neq \emptyset$$

□

Lemma 2. *If $\{A_i, B_i\}_{i=1}^n$ are a set of Gumbel boxes in $\mathbf{G}_d(\boldsymbol{\beta})$ such that*

$$A_i \cap B_j = \emptyset \iff i = j, \quad \textit{and} \quad A_i \cap A_j \neq \emptyset, \quad B_i \cap B_j \neq \emptyset,$$

then the temperature vector $\boldsymbol{\beta}$ must have at least n zeros, i.e. $|\{i : \beta_i = 0\}| \geq n$.

Proof. Since $A_i \cap B_i = \emptyset$ there must be some dimension j for which $\beta_j = 0$. In this dimension, the membership functions are simply characteristic functions of intervals, and thus Lemma 1 implies that for all $k \neq i$, A_k and B_k must intersect in this dimension. But this argument applied to each pair A_i, B_i, and therefore there must be n distinct dimensions with temperature 0 to allow for the intersection pattern described. □

Proposition 3. *For any $\boldsymbol{\beta} \in \mathbb{R}^d_{\geq 0}$, $\mathbf{G}_d(\boldsymbol{\beta})$ has strictly less representational capacity than $\mathbf{G}_{d+1}(\boldsymbol{\beta}, 0)$.*

Proof. We can prove $\mathbf{G}_{d+1}(\boldsymbol{\beta}, 0)$ is as expressive as $\mathbf{G}_d(\boldsymbol{\beta})$ by explicitly constructing the fuzzy set representation $f : \mathbf{G}_d(\boldsymbol{\beta}) \hookrightarrow \mathbf{G}_{d+1}(\boldsymbol{\beta}, 0)$ given by

$$f(M(\mathbf{x}; \mathbf{x}^-, \mathbf{x}^+, \boldsymbol{\beta})) = M(\mathbf{z}; (\mathbf{x}^-, 0), (\mathbf{x}^+, 1), (\boldsymbol{\beta}, 0)),$$

where this latter function is a Gumbel box in $\mathbb{R}^{d+1}$. The fuzzy intersection of Gumbel boxes is the product of their membership functions, and since we can group this product per-dimension it is enough to note:

$$\begin{aligned}
f(M(\mathbf{x}; \mathbf{a}^-, \mathbf{a}^+, \boldsymbol{\beta}))f(M(\mathbf{x}; \mathbf{b}^-, \mathbf{b}^+, \boldsymbol{\beta})) &= M(\mathbf{z}; (\mathbf{a}^-, 0), (\mathbf{a}^+, 1), (\boldsymbol{\beta}, 0))M(\mathbf{z}; (\mathbf{b}^-, 0), (\mathbf{b}^+, 1), (\boldsymbol{\beta}, 0)) \\
&= \left(\prod_{i=1}^{d} m(z_i; a_i^-, a_i^+, \beta_i)m(z_i; b_i^-, b_i^+, \beta_i)\right) m(z_{d+1}; 0, 1, 0)^2 \\
&= \left(\prod_{i=1}^{d} m(z_i; \text{LSE}(a_i^-, b_i^-; -\beta_i), \text{LSE}(a_i^+, b_i^+; \beta_i), \beta_i)\right) m(z_{d+1}; 0, 1, 0) \\
&= f(M(\mathbf{x}; \mathbf{a}^-, \mathbf{a}^+, \boldsymbol{\beta})M(\mathbf{x}; \mathbf{b}^-, \mathbf{b}^+, \boldsymbol{\beta})),
\end{aligned}$$

as desired.

In order to prove that it is strictly more expressive, we consider boxes $A_i, B_i \in \mathbf{G}_{d+1}((\boldsymbol{\beta}, 0))$ with parameters as follows:

$$\forall j \neq i, \quad a_j^- = b_j^- = -2 \quad \text{and} \quad a_j^+ = b_j^+ = 2$$
$$a_i^- = -2, \quad a_i^+ = -1, \quad b_i^- = 1, \quad b_i^+ = 2.$$

Let $I = \{i : \beta_i = 0\} \cup \{d + 1\}$. Suppose there exists a fuzzy set representation $f : \mathbf{G}_{d+1}((\boldsymbol{\beta}, 0)) \hookrightarrow \mathbf{G}_d(\boldsymbol{\beta})$, and consider the set $S = \{f(A_i), f(B_i) : i \in I\}$. The set S is a set of Gumbel boxes which satisfy the condition of Lemma 2, thus implying that $\boldsymbol{\beta}$ has $|I|$ zeros, however by construction $|I|$ is one greater than the number of zeros in $\boldsymbol{\beta}$, which is a contradiction. □