Min/Max Stability and Box Distributions

Michael Boratko[1] **Javier Burroni**[1] **Shib Sankar Dasgupta**[1] **Andrew McCallum**[1]

[1]Computer Science Dept., University of Massachusetts, Amherst

Abstract

In representation learning, capturing correlations between the represented elements is paramount. A recent line of work introduces the notion of learning region-based representations, with the objective of being able to better capture these correlations as set interactions. Box models use regions which are products of intervals on $[0, 1]$ (i.e., "boxes"), representing joint probability distributions via Lebesgue measure. To mitigate issues with training, a recent work models the endpoints of these intervals using Gumbel distributions, chosen due to their min/max-stability. In this work we analyze min/max-stability on a bounded domain and provide a specific family of such distributions which, replacing Gumbel, allow for stochastic boxes embedded in a finite measure space. This allows for a latent noise model which is a probability measure. Furthermore, we demonstrate an equivalence between this region-based representation and a density representation, where intersection is given by products of densities. We compare our model to previous region-based probability models, and demonstrate it is capable of being trained effectively to modeling correlations.

1 INTRODUCTION

Two of the most fundamental tasks in machine learning are the ability to compactly represent joint probability distributions and the ability to learn representations of data which facilitate performance in downstream tasks. While historically disparate, several recent lines of work have explored methods which combine these two objectives, learning a representation which is simultaneously capable of encoding a joint probability distribution. Vilnis and McCallum [2015] introduce the idea of learning representations with probabilistic semantics by representing words using Gaussian densities, an approach which has been successfully extended to mixtures [Athiwaratkun and Wilson, 2017] and graph embeddings [He et al., 2015, Bojchevski and Günnemann, 2018].

An alternative line of work takes a region-based approach, representing elements using "cones" [Vendrov et al., 2016] or "boxes" [Subramanian and Chakrabarti, 2018] in $\mathbb{R}^n$. Lai and Hockenmaier [2017] extends the cone representations with probabilistic semantics by integrating their volume under the negative exponential measure on $\mathbb{R}_+$. Probabilistic box embeddings, introduced in Vilnis et al. [2018], constrain the space to the unit hypercube, wherein Lebesgue measure becomes a probability measure. These models provide geometric representations which, given a finite measure on the space, compactly represent a joint probability distribution over binary random variables.

The problem we are interested in is as follows: given a probability space over a finite set $(S, \mathcal{P}(S), P_S)$, we seek a distributed embedding representation which is capable of modeling P_S on $\mathcal{P}(S)$. For example, we consider the set S of outcomes from n (not necessarily independent) coin flips, $S = \{H, T\}^n$, and seek to model the full joint distribution over head and tail outcomes. The goal is to choose a representation which can encode this distribution with fewer than $2^n - 1$ parameters while maintaining sufficient flexibility to represent the sort of distributions we typically encounter.

If $\pi_i \colon S \to \{H, T\}$ is projection to the ith coordinate, then $\pi_i^{-1}(H)$ is the set of outcomes where the ith coin is heads. We define I to be the set of intervals $[a, b] \subseteq [0, 1]$, along with the empty set. A (one dimensional) *box embedding* is a random variable $B \colon [0, 1] \to S$ such that

$$B^{-1} \circ \pi_i^{-1}(H) = [x_i^-, x_i^+] \in I. \qquad (1)$$

Note that this implicitly defines x_i^- and x_i^+ as the endpoints of an interval, and specifying these endpoints for each i also fully defines the box embedding B.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

We denote the pushforward measure as $Q = \lambda \circ B^{-1}$. In practice, these parameters $\{x_i^\pm\}_{i=1}^n$ are trained via gradient-descent on a cross-entropy loss between P_S and Q. Learning such parameters via gradient descent is problematic, as many regions of the loss landscape are flat. This issue was addressed in Li et al. [2019] by convolving the indicator functions of the boxes with a Gaussian kernel. Dasgupta et al. [2020] improved on this further by introducing latent noise on the parameters to improve learning. Specifically, they model the endpoints $\{x_i^\pm\}$ as random variables $X = \{X_i^\pm\}$. If the X_i^-, X_i^+ are constrained to $[0,1]$, given X we can define a probability measure $\widetilde{Q}$ on $\mathcal{P}(S)$ where, for $R \subseteq S$,

$$\widetilde{Q}(R) = \mathbb{E}_X\big[Q(R \,|\, X)\big]. \tag{2}$$

In Dasgupta et al. [2020], the authors make use of Gumbel distributions, which come in two variants,

$$f_{\max}(x;\mu,\beta) = \tfrac{1}{\beta}\exp\left(-\tfrac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}\right), \quad \text{and} \tag{3}$$

$$f_{\min}(x;\mu,\beta) = \tfrac{1}{\beta}\exp\left(\tfrac{x-\mu}{\beta} - e^{\frac{x-\mu}{\beta}}\right). \tag{4}$$

The motivating reason for choosing these distributions is that they are max- and min-stable, respectively, which facilitates the ability to tractably compute (2). Gumbel random variables are unbounded, and thus (2) is not a probability measure. However, the authors exclusively evaluate on tasks which require modeling conditional probabilities of the form

$$P_S(\pi_i^{-1}(H) \,|\, \pi_j^{-1}(H)), \tag{5}$$

and thus consider the ratio of expectations

$$\frac{\mathbb{E}_X[Q(\pi_i^{-1}(H) \cap \pi_j^{-1}(H) \,|\, X)]}{\mathbb{E}_X[Q(\pi_j^{-1}(H) \,|\, X)]} \tag{6}$$

as an approximation, on which sets $\widetilde{Q}$ is finite. The practical impact of these approximations is not entirely clear, however a model which eschews these difficulties by introducing bounded min- and max-stable random variables would not only have the benefit of probabilistic soundness but also provide the capability for training and evaluating on joint probabilities, a major benefit of the original probabilistic box embeddings model which was lost as a consequence of introducing latent noise.

In this work, we provide a theoretical treatment of min- and max-stable distributions on a bounded domain. Using this framework, we derive a particular set of bounded min- and max-stable distributions which allow for tractable computation of (2). An alternative approach, as taken by probabilistic order embeddings, would be to learn a finite measure on the space $\mathbb{R}$, however we demonstrate that these two perspectives are actually one and the same, connected via the unique form of the integrand which implies that the expected volume of box intersections are actually given by products of their unnormalized densities. We demonstrate that this entire enterprise can be viewed as learning representations in the form of a particular class of probability distributions, thus tying the probabilistic box embedding model tighter to previous work on density representation learning. Finally, we demonstrate empirically that this approach has advantages when regressing to joint probabilities.

2 BACKGROUND

Probabilistic Box Embeddings, introduced in Vilnis et al. [2018], are an embedding method which represents entities with a Cartesian product of intervals, or "box",

$$\begin{aligned}\operatorname{Box}(\mathbf{x}) \coloneqq \prod_{\ell=1}^{d}[x_\ell^-, x_\ell^+] &= [x_1^-, x_1^+] \times \cdots \times [x_d^-, x_d^+] \\ &\subseteq \Omega_{\text{Box}} \subseteq \mathbb{R}^d\end{aligned}$$

where $x_\ell^- \leq x_\ell^+$, $\mathbf{x} = (x_1^-, \ldots, x_d^-, x_1^+, \ldots, x_d^+) \in \mathbb{R}^{2d}$. We adopt the convention that if any $x_\ell^- > x_\ell^+$, $\operatorname{Box}(\mathbf{x}) = \emptyset$, and define the collection of all boxes as $I(\Omega_{\text{Box}})$. Note that this is closed under (set) intersection, that is if $X, Y \in I(\Omega_{\text{Box}})$, $X \cap Y \in I(\Omega_{\text{Box}})$. If this intersection is nonempty we have

$$\operatorname{Box}(\mathbf{x}) \cap \operatorname{Box}(\mathbf{y}) = \prod_{\ell=1}^{d}[\max(x_\ell^-, y_\ell^-), \min(x_\ell^+, y_\ell^+)]. \tag{7}$$

A *box embedding* of a probability space $(S, \mathcal{P}(S), P_S)$ into some measure space $(\Omega_{\text{Box}}, \sigma(I(\Omega_{\text{Box}})), \mu_{\text{Box}})$ is a measurable function $B\colon \Omega_{\text{Box}} \to S$, such that $B^{-1} \circ \pi_i^{-1}(H) \in I(\Omega_{\text{Box}})$, where π_i is the ith projection. If μ_{Box} is a probability measure, the pushforward measure $Q = \mu_{\text{Box}} \circ B^{-1}$ is a probability measure on S, in which case we call B a *probabilistic box embedding*.

Example 1. Given some parameters $\mathbf{x}_i \in \mathbb{R}^{2d}$ associated with each $\pi_i^{-1}(H)$, we define the random variable $B(x)$ such that

$$\pi_i(B(x)) = \begin{cases} H & \text{if } x \in \operatorname{Box}(\mathbf{x}_i), \\ T & \text{otherwise.} \end{cases} \tag{8}$$

If $\Omega_{\text{Box}} = [0,1]^d$, $\mu_{\text{Box}} = \lambda$ is Lebesgue measure, then B is a probabilistic box embedding, and

$$\mu_{\text{Box}}(\operatorname{Box}(\mathbf{x})) = \prod_{\ell=1}^{d}\max(0, x_\ell^+ - x_\ell^-). \tag{9}$$

We train box embeddings using gradient descent, learning the parameters $\{\mathbf{x}_i\}_{i=1}^n$ which minimize a cross-entropy loss between Q and P_S. We will only consider situations where Ω_{Box} is a product space, and μ_{Box} a product measure. Thus, it is enough to consider one-dimensional box embeddings.

Following Dasgupta et al. [2020], we consider a latent noise model where the box parameters $\{x_i^\pm\}$ are modeled using $X = \{X_i^\pm\}$, where X_i^-, X_i^+ are independent random variables taking values in Ω_{Box}. Our goal will be to choose a distribution for the variables in X such that we can compute (or at least reasonably approximate) $\widetilde{Q}$, defined for $R \subseteq S$ as

$$\widetilde{Q}(R) = \mathbb{E}_X\big[Q(R \,|\, X)\big]. \tag{10}$$

As previously mentioned, in the case of $\Omega_{\mathrm{Box}} = [0, 1]$ with μ_{Box} Lebesgue measure, $\widetilde{Q}$ is a probability measure. More generally, if $\mu_{\mathrm{Box}}(\Omega_{\mathrm{Box}})$ is finite, we can normalize $\widetilde{Q}$ to a probability measure on $\mathcal{P}(S)$. Furthermore, the restriction of $\tilde{Q}$ to $\sigma(\bigcup_{i=1}^n \pi_i^{-1}(H)) = \sigma(S \backslash \{T^n\})$ is always finite for each X, and hence the expectation is a finite measure. (This follows from sub-additivity of the measure by observing that $\widetilde{Q}(\pi_i^{-1}(H)) < \infty$, provided $\mathbb{E}[X]$ is finite).

We are thus left with two possible solutions:

1. Use bounded random variables for X, and $\mu_{\mathrm{Box}} = \lambda$.
2. Allow X to be unbounded, but find a measure μ_{Box} for which $\mu_{\mathrm{Box}}(\mathbb{R}) < \infty$.

In both cases, we further require that (10) is able to be calculated explicitly, at least for the subsets $R \subseteq S$ of interest, and thus we first investigate the difficulty of calculating this expectation in general.

3 BOX MEASURES

We ended the previous section by mentioning that we need to be able to compute (10) for "sets of interest", which we will refine and explore in this section. In particular, the most primitive set of interest we might inquire about are boxes given by $\pi_i^{-1}(H)$. If μ_{Box} is Lebesgue, this means we must be able to calculate

$$\mathbb{E}_X\left[Q(\pi_i^{-1}(H) \,|\, X)\right] = \mathbb{E}_X\left[\max(0, X_i^+ - X_i^-)\right]. \tag{11}$$

We adopt the following conventional notation: for a random variable X we denote its cdf, potentially parameterized by θ, by $F_X(x; \theta)$, and, if X is a.c., we denote its pdf as $f_X(x; \theta)$.

We recall the following lemmas:

Lemma 1. *If X is a real-valued random variable with finite mean then*

$$\lim_{x \to -\infty} xF(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} x(1 - F(x)) = 0$$

Lemma 2. *Let X, Y be independent random variables a.c. with respect to the Lebesgue measure. Then*

$$\mathbb{E}[\max(X, Y)] = \int_{-\infty}^{\infty} z\Big(f_Y(z)F_X(z) + f_x(z)F_Y(z)\Big)\,dz.$$

For completeness, we include elementary proofs of these results in Appendix A.

Leveraging these results, we prove the following general result for computing (11):

Lemma 3. *Let X, Y be independent real-valued random variables for which $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are finite. Then*

$$\mathbb{E}[\max(0, Y - X)] = \int_{\mathbb{R}} [1 - F_Y(z)]\, F_X(z)\,dz. \tag{12}$$

Proof. Since $\mathbb{E}[X]$ is finite, we have

$$\begin{aligned}
\mathbb{E}[\max(0, Y - X)] &= \mathbb{E}[\max(X, Y)] - \mathbb{E}[X] && (13)\\
&= \int_{\mathbb{R}} z\Big(f_Y(z)F_X(z) + f_X(z)F_Y(z)\Big)\,dz - \int_{\mathbb{R}} z f_X(z)\,dz && (14)\\
&= \int_{\mathbb{R}} z\Big(f_Y(z)F_X(z) - f_X(z)[1 - F_Y(z)]\Big)\,dz. && (15)
\end{aligned}$$

Integrating by parts, we find this is equal to

$$-z[1 - F_Y(z)]F_X(z)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} [1 - F_Y(z)]F_X(z)\,dz. \tag{16}$$

The first term is 0 by Lemma 1, which completes the proof. □

Remark 1. Lemma 3 provides a much more straightforward calculation for the expected volume of a Gumbel box than that which was provided in Dasgupta et al. [2020]. (See Appendix B.)

The next simplest set of interest to consider are pairwise intersections, $\pi_i^{-1}(H) \cap \pi_j^{-1}(H)$ for $i \neq j$. Given X, this intersection becomes

$$\begin{aligned}
\pi_i^{-1}(H) \cap \pi_j^{-1}(H) &= [X_i^-, X_i^+] \cap [X_j^-, X_j^+] && (17)\\
&= [\max(X_i^-, X_j^-), \min(X_i^+, X_j^+)]. && (18)
\end{aligned}$$

Setting

$$Z^- = \max(X_i^-, X_j^-) \quad \text{and} \quad Z^+ = \min(X_i^+, X_j^+), \tag{19}$$

Lemma 3 implies,

$$\begin{aligned}
\widetilde{Q}(\pi_i^{-1}(H) \cap \pi_j^{-1}(H)) &= \mathbb{E}_X[\max(0, Z^+ - Z^-)] && (20)\\
&= \int_{\mathbb{R}} [1 - F_{Z^+}(z)]\, F_{Z^-}(z)\,dz && (21)\\
&= \int_{\mathbb{R}} \Big[1 - F_{X_i^+}(z)\Big] F_{X_i^-}(z) \Big[1 - F_{X_j^+}(z)\Big] F_{X_j^-}(z)\,dz. && (22)
\end{aligned}$$

where (22) follows due to independence. This extends, via induction, for any nonempty set of indices $J \subseteq \{1, \ldots, m\}$

as follows:

$$\widetilde{Q}(\cap_{j\in J}\pi_j^{-1}(H)) = \int_{\mathbb{R}} \prod_{j\in J}(1 - F_{X_j^+}(z))F_{X_j^-}(z)\,dz. \tag{23}$$

Remark 2. Since, for all $z \in \mathbb{R}$,

$$[1 - F_{Y^+}(z)]\, F_{Y^-}(z) \leq 1,$$

then

$$\mathbb{E}[\max(0, Z^+ - Z^-)] \leq \mathbb{E}[\max(0, X^+ - X^-)].$$

That is, the expected volume of the intersection of two boxes is bounded above by the minimum of the expected volume of each box.

Remark 3. Although motivated and presented as latent noise over the parameters of a region-based model, an alternative perspective which this analysis lays bare is that one is simply learning representations of unnormalized densities of the form

$$\left[1 - F_{X_i^+}(z_i)\right] F_{X_i^-}(z_i). \tag{24}$$

We will investigate this perspective further in Section 6.

Remark 4. If μ_{Box} is a probability measure, and therefore so is $\widetilde{Q}$, we can use inclusion-exclusion to calculate $\widetilde{Q}(R)$ for any $R \subseteq S$. Hence, if we can identify a bounded distribution such that (23) is tractable we can calculate $\widetilde{Q}(R)$ for any $R \subseteq S$.

Without additional assumptions about the random variables, finding all general distributions for which (23) is tractable for any nonempty $J \subseteq \{1, \ldots, m\}$ is difficult. If, however, we identify some parameterized families of distributions $\{F_\theta^-\}$, $\{F_\theta^+\}$ which are max- and min-stable respectively for which we can also calculate

$$\int_{\mathbb{R}} \left[1 - F_{\theta_1}^+(z)\right] F_{\theta_2}^-(z)\,dz \tag{25}$$

for any setting of parameters θ_1, θ_2 then by observing (19) and (21) we find the measure of any pairwise intersection $\widetilde{Q}(\pi_i^{-1}(H) \cap \pi_j^{-1}(H))$ is also tractable. This extends inductively to show that (23) can be calculated explicitly, and thus we seek a formal characterization of min- and max-stable distributions.

4 MIN/MAX-STABILITY

Motivated by our observation in Section 3 that deriving bounded min- and max-stable distributions for which (25) is calculable implies that $\widetilde{Q}$ is a probability measure on $\mathcal{P}(S)$ and allows us to calculate the probability $\widetilde{Q}(R)$ for any $R \subseteq S$, we formally define our notion of min- and max-stability.

Definition 1 (Min/Max-Stability)**.** We call a family of distributions on Ω parameterized by $\theta \in \Theta$, which we denote

$$\mathcal{F} = \mathcal{F}_{\Omega,\Theta} = \{F_\theta\colon \Omega \to [0,1]\}_{\theta\in\Theta}, \tag{26}$$

max-stable if there exists some $g\colon \Theta \times \Theta \to \Theta$, such that, for X and Y random variables with distributions F_{θ_X} and F_{θ_Y},

$$F_{\theta_X}, F_{\theta_y} \in \mathcal{F}_\Omega \implies F_{\max(X,Y)} = F_{g(\theta_X,\theta_Y)} \in \mathcal{F}_\Omega.$$

We call g the *max parameter transformation*. *Min-stability* and *min parameter transformation* are defined analogously.

As mentioned previously, Gumbel distributions come in min- and max-stable variants, and in fact the min variant $f_{\min}(x;\mu,\beta)$ in (4) is equivalent to $f_{\max}(-x;-\mu,\beta)$, as defined in (3). The following proposition shows that this is actually true more generally.

Proposition 1. *Let $S, T \subseteq \mathbb{R}$, $\mathcal{F}_{T,\Theta}$ be some min- or max-stable family of distributions, and let $h\colon T \to S$ be any measurable monotonic bijection. Then the family*

$$\mathcal{F}_{S,\Theta} \coloneqq \{F_\theta \circ h^{-1}\colon S \to [0,1] \quad \text{for} \quad F_\theta \in F_{T,\Theta}\}$$

has the same stability as $\mathcal{F}_{T,\Theta}$ if h is increasing and opposite stability (eg. min becomes max) if h is decreasing. In each case, the parameter transformation is preserved.

Proof. Suppose that $\mathcal{F}_{T,\Theta}$ is max-stable and h is monotonically increasing. Let X and Y be random variables with distributions $F_X, F_Y \in \mathcal{F}_{S,\Theta}$. By definition of $\mathcal{F}_{S,\Theta}$, there exist random variables $\tilde{X}$ and $\tilde{Y}$ with distributions $F_{\theta_{\tilde{X}}}, F_{\theta_{\tilde{Y}}} \in \mathcal{F}_{T,\Theta}$ such that $X = h \circ \tilde{X}$ and $Y = h \circ \tilde{Y}$. Then,

$$\max(X, Y) = \max(h \circ \tilde{X}, h \circ \tilde{Y})) = h \circ \max(\tilde{X}, \tilde{Y}).$$

Since $\mathcal{F}_{T,\Theta}$ is max-stable, there exists g, such that, $F_{\max(\tilde{X},\tilde{Y})} = F_{g(\theta_{\tilde{X}},\theta_{\tilde{Y}})} \in \mathcal{F}_{T,\Theta}$. Therefore, $\mathcal{F}_{S,\Theta}$ is max-stable with the same g as $\mathcal{F}_{S,\Theta}$.

Alternatively, suppose h is monotonically decreasing. Then,

$$\min(X, Y) = \min(h \circ \tilde{X}, h \circ \tilde{Y})) = h \circ \max(\tilde{X}, \tilde{Y}).$$

The proof for the case of $\mathcal{F}_{T,\Theta}$ min-stable is similar. □

This yields the following useful corollaries:

Corollary 1. *If $\mathcal{F} = \{F_\theta\colon \mathbb{R} \to \mathbb{R}\}$ is min- (resp. max-) stable then $\mathcal{F}' = \{F_\theta \circ (x \mapsto -x)\colon \mathbb{R} \to \mathbb{R}\}$ is max- (resp. min-) stable. That is, there is a bijection between max- and min-stable families on $\mathbb{R}$.*

Corollary 2. *There is a bijection between min/max-stable distributions on $\mathbb{R}$ and those on any open interval (a, b).*

This latter corollary arises from noting that (a, b) is homeomorphic to $\mathbb{R}$, and any continuous injective function $h\colon (a, b) \to \mathbb{R}$ must be (strictly) monotonic. Obviously, this gives us infinitely many choices for bounded min- and max-stable distributions, however we may reasonably pause at this point to consider if there are already some common min- and max-stable distributions on a bounded interval.

Example 2. Let H_θ be the cdf of the delta distribution δ_θ, then

$$\mathcal{F} = \{H_\theta : \theta \in [0, 1]\} \tag{27}$$

is a bounded family of distributions on $[0, 1]$ which are both min- and max-stable, with parameter transformation functions simply given by $\min$ and $\max$ respectively. Utilizing these distributions in our current framework leads precisely to the original probabilistic box embedding model introduced in Vilnis et al. [2018].

Delta distributions struggle with learning precisely because $\mathbb{E}[\max(0, Y - X)] = \max(0, \theta_Y - \theta_X)$ has zero gradient when $\theta_Y < \theta_X$, and this situation is exasperated when computing the expectation of the intersection of multiple elements. Dasgupta et al. [2020] introduce the Gumbel distribution to mitigate this issue.

Example 3. The Gumbel max distribution provides a source of infinitely many max-stable families on $\mathbb{R}$ (one for each scale parameter $\beta \in \mathbb{R}_+$) defined as:

$$\mathcal{F}(\mathbb{R}, \mathbb{R}; \beta) = \left\{ F_\mu(x) = \exp\left(-e^{-\frac{x-\mu}{\beta}}\right) \mid \mu \in \mathbb{R} \right\}. \tag{28}$$

The max parameter transformation is given by

$$g(\mu_x, \mu_y) = \beta \ln\left(e^{\frac{\mu_x}{\beta}} + e^{\frac{\mu_y}{\beta}}\right) \tag{29}$$

The fact that the Gumbel min distribution is min-stable is now a specific instance of Proposition 1.

Remark 5. Note that the parameter transformation g for the Gumbel distributions is smooth. Furthermore, the approximation

$$\mathbb{E}[\max(0, Y - X)] \approx \beta \log(1 + \exp(\tfrac{\mu_Y - \mu_X}{\beta} - 2\gamma)) \tag{30}$$

introduced in Dasgupta et al. [2020] is smooth with respect to the μ_Y, μ_X parameters, and thus the chain rule implies that if X is modeled using Gumbel min/max distributions, $\widetilde{Q}(R)$ is a smooth function of their location parameters, which assists with training.

5 GUMBEL BOX MEASURES

With the necessary machinery developed, we now derive specific min- and max-stable families of distributions which are bounded, for which we can approximate $\mathbb{E}[\max(0, Y - X)]$, and furthermore preserve the smoothness property mentioned in Remark 5.

Building off our previous work, we consider Y with distribution $F_Y \in \mathcal{F}_{\min}$ and X with distribution $F_X \in \mathcal{F}_{\max}$, where $\mathcal{F}_{\min}$ is min-stable and $\mathcal{F}_{\max}$ is max-stable on $\mathbb{R}$. Let $h\colon \mathbb{R} \to (0, 1)$ be a monotonically increasing C^1 bijection with C^1 inverse. Then by Lemma 3, we have that

$$\begin{aligned}
&\mathbb{E}_{h(X),h(Y)}[\max(0, h(Y) - h(X))] && (31)\\
&\quad = \int_0^1 [1 - F_{h(Y)}(z)] F_{h(X)}(z)\, dz && (32)\\
&\quad = \int_0^1 [1 - F_Y(h^{-1}(z))] F_X(h^{-1}(z))\, dz && (33)\\
&\quad = \int_{-\infty}^{\infty} [1 - F_Y(u)] F_X(u) \frac{d}{du} h(u)\, du. && (34)
\end{aligned}$$

Note that for any h as described, $\frac{d}{du} h(u) = g(u)$ is the pdf of some probability distribution, and furthermore any pdf $g(u)$ yields a unique $h(u)$. Thus, it simply remains to choose a probability distribution for which the above calculation is tractable, which depends on the possible F_Y, F_X distributions. The observation in Remark 3 suggests considering

$$g(u) \propto [1 - F_{U^+}(u)]\, F_{U^-}(u) \tag{35}$$

for any fixed $F_{U^+} \in \mathcal{F}_{\max}$, $F_{U^-} \in \mathcal{F}_{\min}$. The properties of min/max stability referenced in (25) imply, therefore, that (34) is computable if and only if $\int_{-\infty}^{\infty} [1 - F_Y(u)] F_X(u)\, du$ is. If we take $\mathcal{F}_{\min}, \mathcal{F}_{\max}$ as Gumbel min and max families, therefore, we can approximate this using (30), and further preserve the smoothness of $\widetilde{Q}(R)$ with respect to the Gumbel parameters.

5.1 ALTERNATIVE PERSPECTIVES

As described, the approach we took focused on deriving min/max stable distributions on $[0, 1]$, in which case we have $\Omega_{\text{Box}} = [0, 1]$ and $\mu_{\text{Box}} = \lambda$. Inspecting the form of (34), however, it is obvious that this is entirely equivalent to allowing $\Omega_{\text{Box}} = \mathbb{R}$ and simply using the measure $\mu_{\text{Box}} = g(u)\, du$.

Furthermore, a naïve approach to mitigating the fact that Gumbel distributions are not defined over a bounded domain would be to simply fix some Gumbel box, called the "universe box" U. Then, when asked to calculate a marginal probability $P(\pi_i^{-1}(H))$, for example, we would actually compute

$$\frac{\mathbb{E}_{X \times U}\left[\mu_{\text{Box}}(B^{-1}(\pi_i^{-1}(H)) \cap U)\right]}{\mathbb{E}_U\left[\mu_{\text{Box}}(U)\right]}. \tag{36}$$

If $\mu_{\text{Box}} = \lambda$, using Lemma 3 and (22), this is equivalent to

$$\frac{\int_{\mathbb{R}} \left[1 - F_{X_i^+}(u)\right] F_{X_i^-}(u) \left[1 - F_{U^+}(u)\right] F_{U^-}(u)\, du}{\int_{\mathbb{R}} \left[1 - F_{U^+}(u)\right] F_{U^-}(u)\, du}, \tag{37}$$

and hence this is equivalent to the aforementioned approach, with g as in (35).

Remark 6. Note that this also naturally proves that the ratio of expectations (6) as used in Dasgupta et al. [2020] is, for the $\pi_i^{-1}(H)$, a valid probability measure.

Thus, there are actually three entirely equivalent perspectives:

1. Transforming Gumbel distributions to min-/max-stable distributions on $(0, 1)$ via h^{-1}, where $\frac{d}{du} h(u) = g(u)$.
2. Defining μ_{Box} to be some a.c. finite measure on $\mathbb{R}$, in which case $g(u) = \frac{d\mu_{\text{Box}}}{d\lambda}$.
3. Intersection with some (normalized) "universe" box, corresponding to the density $g(u)$.

These alternative perspectives align geometric intuition and probabilistic formalism. The first perspective is most natural when attempting to rectify the lack of finite measure present in Dasgupta et al. [2020], and also opens the possibility of other transformation functions h. Perspective 2 facilitates analytic computation of intersection volumes, as described in section 6. When choosing parameters of the distributions depicted in Figures 2 and 3 (a realization of the first perspective) we can use geometric intuition based on perspective 3 to adjust position/scale. Perspective 3 is also most useful when implementing the model in practice.

6 GUMBEL BOX DENSITIES

We end our theoretical analysis by expanding on Remark 3, wherein we observed our region-based interpretation was equivalent to one in which elements were represented using unnormalized densities, and thus the box embedding model (particularly in cases with latent noise on the parameters) can be equivalently viewed as a density representation.

To recap, a box embedding with latent noise parameterizes the sets $\pi_i^{-1}(H)$ via some random variables X_i^-, X_i^+ which represent the endpoints of some interval in $[0, 1]$. Intersections $\pi_i^{-1}(H) \cap \pi_j^{-1}(H)$ correspond to intersections of these stochastic intervals. Thus, from this perspective we consider elements to be represented by regions (intervals), with unary marginal probability given by expected length of this interval and joint probabilities given by their intersections.

On the other hand, as noted in (23), for any nonempty set of

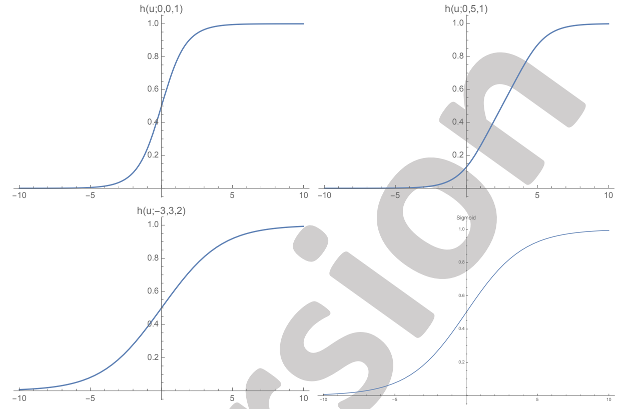

Figure 1: **Universe Box Transformation** h We plot the function h for different parameters of the universe box. Plots are labeled as $h(u^-, u^+, \beta)$. As one would expect, shifting the universe box to the right has the corresponding effect on h, as does stretching the width, both of which can be observed in the plot for $h(0, 5, 1)$ above. Increasing the β also has an intuitive effect, as lower values of β lead to steeper h functions. In the bottom right we plot sigmoid, noting it's extreme similarity with $h(-3, 3, 2)$.

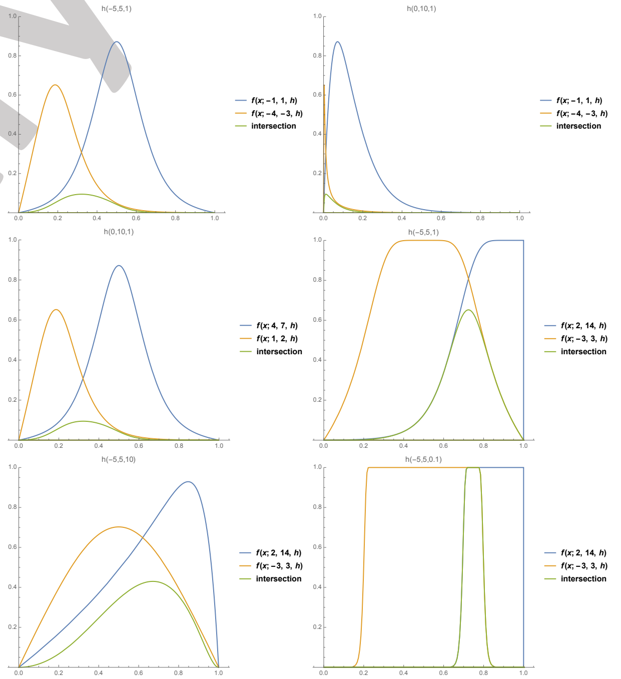

Figure 2: **One-Dimensional Gumbel Boxes** We plot the unnormalized densities representing the transformed Gumbel boxes in $[0, 1]$. The parameters for the universe box $h(u^-, u^+, \beta)$ are shown at the top for each plot, the legend includes labels for each density as $f(x; \mu^-, \mu^+, h)$, and in each plot we include the density which represents the intersection in green. From left to right, top to bottom: (a) arbitrary boxes in position, (b) translates the universe box from (a) by 5, (c) translates the box parameters by 5 as well, (d) shows the effect of using larger boxes, (e) demonstrates the effect of higher scale β, and (f) demonstrates low β.

indices $J \subseteq \{1, \dots, m\}$ we have

$$\widetilde{Q}(\cap_{j\in J}\pi_j^{-1}(H)) = \int_{\mathbb{R}} \prod_{j\in J} \left[1 - F_{X_j^+}(z)\right] F_{X_j^-}(z)\, dz. \tag{38}$$

Therefore, we may consider $\pi_i^{-1}(H)$ to be represented by the unnormalized density $[1 - F_{X_i^+}(z)]F_{X_i^-}(z)$, where $\widetilde{Q}(\pi_i^{-1}(H) \cap \pi_j^{-1}(H))$ corresponds to an inner product in $\mathcal{L}^2$, i.e. joint probabilities between pairs correspond to an inner product in an infinite-dimensional vector space, and marginal probabilities are given by the inner product with the constant function $1 \in \mathcal{L}^2(\mu)$. It is, therefore, of supreme interest to develop a better understanding of the properties of these densities.

In order to map unbounded random variables to $(0, 1)$ we had chosen the transformation $h : \mathbb{R} \to (0, 1)$ such that

$$h'(u) = g(u) \propto [1 - F_{U^+}(u)]F_{U^-}(u), \tag{39}$$

where $[U^-, U^+]$ represents a "universe box". For Gumbel random variables, all variables use the same scale β (to ensure min-/max-stability), and therefore the parameter transformation itself has 2 parameters, which are the location parameters u^-, u^+ of U^-, U^+. We plot h for various values of these parameters in Figure 1, and note that the dependence on the parameters is quite intuitive.

Including the global β and the universe box parameters, we have introduced a family of distributions on $[0, 1]$ indexed by five parameters: $\mu^-, \mu^+, u^-, u^+, \beta$. To gain intuition about the unnormalized density representation, we plot these densities for various values of the parameters in Figure 2. We note that, while there are 5 parameters, there are only 4 degrees of freedom, as the densities are invariant when translating *all* location parameters. As $\beta \to 0$ we recover the "hard box" model of [Vilnis et al., 2018]. We also note that the model is extremely flexible, capable of capturing large regions in the center of the space as well as small areas in a corner or along a side.

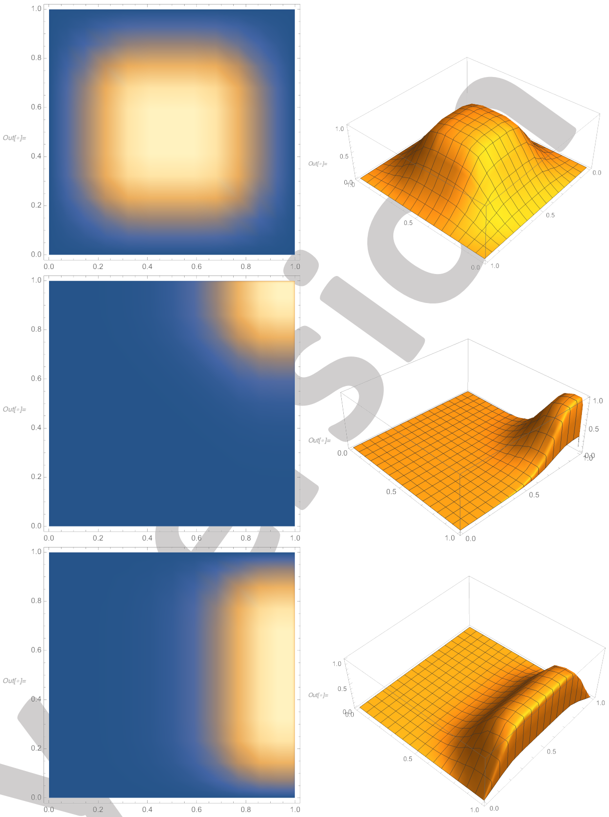

Figure 3: **Two-Dimensional Gumbel Boxes** We plot the unnormalized densities representing the transformed Gumbel boxes in $[0, 1]^2$. We use universe parameters $u^- = -5, u^+ = 5$, and $\beta = 1$. From top to bottom, we have plotted (a) $f(x; -3, -3, h)f(y; -3, -3, h)$, (b) $f(x; 3, 9, h)f(y; 3, 9, h)$, and (c) $f(x; 3, 9, h)f(y; -4, 4, h)$.

7 EXPERIMENTS

7.1 DATASET

In order to evaluate the ease of training and capability to effectively model real-world data, we create a dataset from MovieLens-25M which consists of movie ratings given by different users. Following [Li et al., 2019], we calculate the values for binary random variables indicating whether or not a user likes a given movie. Specifically, we consider a user to have liked a movie if they gave it a rating greater than 4. In order to ensure relevance, we select only popular movies, i.e., movies must have more than 200 ratings. This provided us with $160,369$ users and $3,093$ movies. Given this data, we then calculted $P(\text{MovieA})$, $P(\text{MovieB}, \text{MovieA})$, where the joint probability indicates the probability a given user liked both Movie A and Movie B, and used this to calculate $P(\text{MovieB} \mid \text{MovieA})$. Explicitly,

$$P(\text{MovieA}, \text{MovieB}) = \frac{\#\text{Rating}(\text{MovieA}, \text{MovieB})}{\#\text{Users}}.$$

The conditional probabilities are then calculated by diving the pair-wise joint probabilities with the marginals.

7.2 BASELINES

We compare our method with the following probabilistic embedding methods:

1. **Probabilistic Order Embeddings (POE)** [Lai and Hockenmaier, 2017]: A model which represents elements using infinite cones, integrated under the negative exponential measure.
2. **Probabilistic Box Embeddings (PBE)** [Vilnis et al., 2018]: This model can be viewed as representationally

equivalent to our model where the distributions are simply delta distributions, however to solve issues with disjoint boxes during training the authors introduce a surrogate loss function.

7.3 RESULT

A properly trained probabilistic embedding model should be able to predict joint probabilities $P(\mathrm{MovieA}, \mathrm{MovieB})$ when trained with marginal $P(\mathrm{MovieA})$ and conditional $P(\mathrm{MovieA} \,|\, \mathrm{MovieB})$ probabilities. We evaluate this for the two baselines and compare with our proposed method as well. The results are reported in Table 1.

Table 1: KL divergence (lower is better) for training and validation.

	Train	Train	Validation
	$P(A \,|\, B)$	$P(B)$	$P(A, B)$
POE	0.0052	0.0002	0.00003
PBE	0.0049	0.0006	0.00004
Gumbel Bounded	0.0022	0.00004	0.000009

We observe that all the probabilistic models including the baselines are able to learn both marginals and joint distribution reasonably well. However, our proposed method leaned a distribution which achieves a KL divergence to the target distribution that is a third of the other methods' divergence. We credit this to the smoothness of our model's training objective, as well as the flexibility of the box distributions.

8 RELATED WORK

Critical to our objective was defining a family of transformations $h\colon \mathbb{R} \to [0, 1]$. Furthermore, we wish this h to be sufficiently smooth such that the parameters of the Gumbel distribution can be learned after composition. While quite similar, this is different from the objective for learning a Normalizing Flow [Tabak and Turner, 2013, Rezende and Mohamed, 2015, Kobyzev et al., 2020], wherein the objective is to learn a chain of compositions such that the Jacobian of the composition is easy to compute. Furthermore, Normalizing Flows are used to sample and estimate densities from the transformed distribution, whereas in our case we are interested in analytically calculating an expectation.

Historically, the study of max-stable distributions focused on a more restricted notion of max-stability [Gnedenko, 1943, Gumbel, 1958, Pantcheva, 1985, Kunin, 1997]. That is, a distribution F is max-stable if there exists a sequence of strictly monotone continuous functions $(G_n\colon \mathbb{R} \to \mathbb{R})_{n \in \mathbb{N}}$ such that

$$F(x) = F^n(G_n(x)) \qquad \text{for all } x \in \mathbb{R}.$$

The motivation for such a definition is related to the asymptotic behavior of F^n with $n \to \infty$, i.e., the maximum between a numerable set of i.i.d. random variables. In contrast, the definition presented in this paper aims to take the maximum of a set of variables not necessarily identically distributed, as long as the set is finite and all variables belong to the same family. Moreover, this definitions gives enough freedom to choose a family for which the quantity of (34) can be analytically computed, while still getting gradient information with respect to the parameters. It is this last condition which prevented us from using the distributions over bounded support defined in [Pantcheva, 1985] and [Kunin, 1997].

There are many approaches to representation learning which use objects other than standard Euclidean vectors. In particular, a recent line of work explores the notion of representing elements using vectors in hyperbolic space. Nickel and Kiela [2017] consider a Poincaré disk model, and Nickel and Kiela [2018] consider an equivalent approach using the Lorentzian model, further extended by Law et al. [2019]. Ganea et al. [2018] extends this to a region-based representation where elements are represented by infinite cones in hyperbolic space, a concept originally introduced in Euclidean space by Vendrov et al. [2016]. None of these representations are probabilistic, however. Probabilistic order embeddings [Lai and Hockenmaier, 2017], which we compare with, extend Vendrov et al. [2016] probabilistically by integrating the space under the negative exponential measure. Vilnis et al. [2018] originally introduced the probabilistic box embedding model, and improvements to training were introduced in Li et al. [2019] and Dasgupta et al. [2020], however these approaches resulted in models which no longer provide methods for evaluating joint probabilities, and thus we do not compare directly to them. Non-probabilistic methods for embedding boxes have also been introduced in Subramanian and Chakrabarti [2018].

9 CONCLUSION

In this work, we actually introduce several novel families of distributions on $[0, 1]$. First, motivated by the requirements of box embeddings, we introduce a families of min- and max- stable distributions on a bounded domain for which an explicit expectation can be calculated. We also proved that any such family on $\mathbb{R}$ can be transformed to a family on any set homeomorphic to $\mathbb{R}$. Second, we observe that the densities corresponding to box embeddings themselves can be normalized to provide a novel distribution on $[0, 1]$ which is closed under multiplication and demonstrates impressive flexibility. Practically, we leverage these results by demonstrating the equivalence of learning box representations with latent noise and learning representations of these unnormalized densities. As a result, we derive a box model with latent noise which can benefit learning which provably maintains the probabilistic nature of the original box embedding model. The equivalent perspectives of region and density representation of this model provides new methods of analysis, extension, and insight which we plan to explore in the future.

Author Contributions

Michael Boratko conceived the idea and wrote the paper. Javier Burroni clarified a number of proofs and helped design the mathematical framework for analyzing probabilistic box embeddings. Shib Sankar Dasgupta wrote the code and ran the experiments, and Andrew McCallum provided helpful guidance and feedback.

Acknowledgements

Thanks to the members of the IESL at UMass Amherst for helpful discussions. This work is supported in part by the Center for Intelligent Information Retrieval and the Center for Data Science, in part by IBM Research AI through the AI Horizons Network, in part by University of Southern California subcontract no. 123875727 under Office of Naval Research prime contract no. N660011924032, and in part by the University of Southern California subcontract no. 89341790 under Defense Advanced Research Projects Agency prime contract No. FA8750-17-C-0106. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL, DARPA, or the U.S. Government.

References

Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. In *ACL*, 2017.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. *ICLR*, 2018.

Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. *arXiv preprint arXiv:1804.01882*, 2018.

Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pages 423–453, 1943.

EJ Gumbel. Statistics of extremes. In *Statistics of Extremes*. Columbia University Press, 1958.

Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 623–632, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6. doi: 10.1145/2806416.2806502. URL `http://doi.acm.org/10.1145/2806416.2806502`.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Boris Kunin. A new type of extreme value distributions. *Engineering fracture mechanics*, 58(5-6):557–570, 1997.

Alice Lai and Julia Hockenmaier. Learning to predict denotational probabilities for modeling entailment. In *EACL*, 2017.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3672–3681. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/law19a.html`.

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

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *arXiv preprint arXiv:1705.08039*, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2018.

Elisaveta Pantcheva. Limit theorems for extreme order statistics under nonlinear normalization. In Vladimir V. Kalashnikov and Vladimir M. Zolotarev, editors, *Stability Problems for Stochastic Models*, pages 284–309, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39686-4.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Sandeep Subramanian and Soumen Chakrabarti. New embedded representations and evaluation protocols for inferring transitive relations. *SIGIR 2018*, 2018.

Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

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

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *ICLR*, 2015.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272, 2018.