Modeling Extremes with d-max-decreasing Neural Networks (Appendix)

A BACKGROUND ON EXTREME VALUE THEORY

The main idea behind extreme value theory (EVT) is to establish a form of the central limit theorem for the maxima of appropriately scaled random variables. EVT characterizes the behavior of the maxima of n independent and identically distributed (i.i.d.) random variables $X_1, \cdots, X_n$ with continuous distribution function F. More precisely, let $M_n = \max_{1\leq i\leq n} X_i$, then there exists sequences of real numbers $a_n > 0$ and b_n such that the limit $\mathbb{P}\left[(M_n - b_n)/a_n \leq x\right] \to H(x)$ as $n \to \infty$ is non-degenerate. We then say that F is in the maximum domain of attraction of H or equivalently $F \in \text{MDA}(H)$. This limit is fully identified by the generalized extreme value (GEV) distribution given by:

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right), & \text{if } \xi \neq 0 \\ \exp\left(-e^{-x}\right), & \text{if } \xi = 0 \end{cases} \tag{1}$$

where $1 + \xi x > 0$ and ξ is the shape parameter indicating the thickness of the tail. The following theorem due to Fisher, Tippet and Gnedenko, stated in de Haan and Ferreira [2010], states the fundamental result of EVT.

Theorem 1 (Fisher-Tippet-Gnedenko Theorem, stated in de Haan and Ferreira [2010])**.** *If $F \in MDA(H)$, and if the limit H exists then it belongs to the class of GEV distributions, i.e. $H = H_\xi$ for some real number ξ.*

A.1 MAIN DEFINITIONS AND THEOREMS

Theorem 2 (Extreme value copula)**.** *If C is a $d-$variate extreme value copula then there exists a tail dependence function $\ell : [0,\infty)^d \to [0,\infty)$ such that:*

$$C\left(u_1, \cdots, u_d\right) = e^{-\ell(-\log u_1, \cdots, -\log u_d)}, \tag{2}$$

where $(u_1, \cdots, u_d) \in (0,1]^d$. Using the homogeneity property of ℓ, the extreme value copula C can be rewritten as:

$$C\left(u_1, \cdots, u_d\right) = e^{\left(\sum_{k=1}^d \log u_k\right) A\left(\frac{\log u_1}{\sum_{k=1}^d \log u_k}, \cdots, \frac{\log u_d}{\sum_{k=1}^d \log u_k}\right)}, \tag{3}$$

where A is known as the Pickands dependence function, which can be thought of as the restriction of ℓ to the unit simplex $\Delta_{d-1} = \{\mathbf{w} = (w_1, \cdots, w_d) \in [0,\infty)^d : \sum_{k=1}^d w_k = 1\}$. The Pickands function A is known to be d-max-decreasing and satisfies:

$$\max_{1\leq k\leq d} w_k \leq A(w_1, \cdots, w_d) \leq 1, \tag{4}$$

for all $\mathbf{w} = (w_1, \cdots, w_d) \in \Delta_{d-1}$.

Definition 1 (Tail dependence function)**.** *A function $\ell : [0,\infty)^d \to [0,\infty)$ is a tail dependence function if for all $(x_1, \cdots, x_d) \in [0,\infty)^d$, the following conditions are satisfied:*

- *(i) ℓ is d-max-decreasing and homogeneous of order 1, i.e. $\ell(cx_1, \cdots, cx_d) = c\,\ell(x_1, \cdots, x_d)$, for all $c > 0$.*
- *(ii)* $\max_{1\leq k\leq d} x_k \leq \ell(x_1, \cdots, x_d) \leq \sum_{k=1}^d x_k$.

A.2 SPECTRAL DECOMPOSITION OF STATIONARY MAX-STABLE PROCESSES

Stationary max-stable processes can be intuitively interpreted as i.i.d. samples from infinite dimensional extreme value distributions (i.e. distributions over functions). A stationary max-stable process can be decomposed by the spectral representation defined in De Haan et al. [1984] which we recall in Proposition 1.

Proposition 1 (Spectral Representation of Max-Stable Processes [De Haan et al., 1984])**.** *Suppose that $M(t)$ has unit Fréchet margins and is stationary. Then, $M(t)$ can be written as:*

$$M(t) = \max_{i \geq 1} \xi_i Y_i^+(t), \quad t \in \mathcal{T}. \tag{5}$$

$\{Y_i(t)\}_{i \geq 1}$ are i.i.d. copies of a continuous stochastic process Y defined on $\mathcal{T}$ such that $\mathbb{E}[Y^+(t)] = 1$ with $Y^+(t) = \max\{0, Y(t)\}$ and ξ_i is the i^{th} realization of an independent Poisson point process on $[0, \infty)$ with intensity $\xi^{-2} d\xi$.

B D-MAX-DECREASING FUNCTIONS

We use the definition given in Hofmann [2009]. A function $A(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}$ is d-max-decreasing if and only if for any $\mathbf{x} \leq \mathbf{y} \leq 0$ and any subset $E \subsetneq \{1, \ldots, d\}$:

$$\sum_{\substack{\mathbf{m}=\{0,1\}^d, \\ m_j = 1 \text{ if } j \in E}} \left[(-1)^{d+1-\sum_{j \leq d} m_j} \left(\sum_{j \leq d} -y_j^{m_j} x_j^{1-m_j} \right) A \left(\frac{y_1^{m_1} x_1^{1-m_1}}{\sum_{j \leq d} y_j^{m_j} x_j^{1-m_j}}, \cdots, \frac{y_d^{m_d} x_d^{1-m_d}}{\sum_{j \leq d} y_j^{m_j} x_j^{1-m_j}} \right) \right] \geq 0,$$

and $A(\mathbf{e}_i) = 1$ where $\mathbf{e}_i$ is the canonical basis function. Moreover, from Hofmann [2009, Theorem 3.1.1], the following three characterizations are equivalent:

1. The function
$$\exp\left(- \sum_{j=1,\ldots,d} x_j A \left(\frac{x_1}{\sum_{j=1,\ldots,d} x_j}, \cdots, \frac{x_d}{\sum_{j=1,\ldots,d} x_j} \right) \right)$$
defines a multivariate extreme value distribution;
2. There exists a spectral measure Λ such that
$$A(\mathbf{w}) = \int_{\Delta_{d-1}} \max_{k=1,\ldots,d} w_k s_i \, \mathrm{d}\Lambda(\mathbf{s}), \quad \mathbf{w} \in \Delta_{d-1};$$
3. $A(\mathbf{w})$ is d-max-decreasing.

Next, we note the nesting property of d-max-decreasing functions, given in Hofert et al. [2018, Section 2.2], that hierarchies of spectral measures define valid EVDs, i.e.

$$A(\mathbf{w}) = \mathbb{E}\left[s_1^{(2,1)} \mathbb{E}\left[\max_{k=1,\ldots,d} s_k^{(1,1)} w_k \right], \ldots, s_d^{(2,1)} \mathbb{E}\left[\max_{k=1,\ldots,d} s_k^{(1,d)} w_k \right] \right].$$

We use part of this property in the next section to define the d-max neural network.

C PROOFS

C.1 D-MAX-DECREASING NEURAL NETWORKS

We partition the proof into the 1-layer case and the n-layer case. We assume that all parameters $\theta \in [0, 1]$ for the purposes of the proof.

Background: $D-$norms. $D-$norms are norms defined as

$$\|\mathbf{x}\|_D := d\,\mathbb{E}_{\boldsymbol{\theta}}\left[\max_{k=1\dots d}(|x_k|\theta_k)\right], \quad \mathbf{x}\in\mathcal{X}\subset\mathbb{R}^d, \tag{6}$$

where $\boldsymbol{\theta}\in[0,1]^d$ and $\mathbb{E}[\theta_k]=1/d$, for $i=1\dots d$. We note that the condition $\mathbb{E}[\theta_k]=1/d$ is not necessary for the $d-\max$ decreasing property, though it leads to unit exponential margins for convenience during inference, see Fougères et al. [2013, Section 2.1] or Hofmann [2009, Definition 5.4.1] for more on this property. The key condition is that the expectation in (6) is taken with respect to the distribution of $\boldsymbol{\theta}$ which has support only on nonnegative real numbers. Taking $\mathcal{X}$ to be the unit simplex, we see that a $D-$norm defines a Pickands dependence function, and by the spectral representation of the Pickands function, all Pickands functions are $D-$norms. The main property we will use throughout the proof is that compositions of $D-$norms are also $D-$norms. This property is well established in, for example, Hofert et al. [2018] and Hofmann [2009]. We finally note that all $D-$norms satisfy the $d-$max decreasing property defined as shown in Hofmann [2009, Theorem 3.1.1] .

1-Layer Case.

Proof. Recall that the 1-layer dMNN is given by

$$A_\theta^{(1)}(\mathbf{w}) = \max\left(L^{(1)}(\mathbf{w}) + (1-L^{(1)}(\mathbf{e})^T\mathbf{w}), \max_{k=1\dots d} w_k\right), \quad \mathbf{w}\in\Delta_{d-1} \tag{7}$$

$$L^{(1)}(\mathbf{w}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\max_{k=1\dots d} w_k\theta_{kj}\right)_j \tag{8}$$

The expression (8), corresponding to the first term in (7), is a valid $D-$norm since it is the expectation with respect to a nonnegative spectral measure. The second term $(1-L^{(1)}(\mathbf{e})^T\mathbf{w}) = \sum_{i=1}^d(1-\mathbb{E}[\theta_i])w_i$ is also a $D-$norm since $(1-\mathbb{E}[\theta_i])$ is positive, for $i=1\dots d$, and thus it is also an expectation with respect to a nonnegative spectral measure. In fact, it is equivalent to $\|\text{diag}(1-\mathbb{E}[\boldsymbol{\theta}])\mathbf{w}\|_1$. The combination $L^{(1)}(\mathbf{w}) + (1-L^{(1)}(\mathbf{e})^T\mathbf{w})$ is then the sum of two $D-$norms, equivalent to a composition with the $\|\cdot\|_1$ norm, which is again a $D-$norm. Finally, the outer $\max$ with $\max_{i=1\dots d} w_i$, a $D-$norm corresponding to dependence, is yet another composition of $D-$norms. This results in a function that is $d-\max$ decreasing and concludes the proof for the single layer case. □

n-Layer Case.

Proof. We first show the base case (the 2-layer case), then show that the general n-layer case follows. We focus on the composition of intermediate layers, since the technique for proving the output layer is a $D-$norm follows from the 1-layer case. Recall that the 2-layer dMNN is given by

$$A_\theta^{(2)}(\mathbf{w}) = \max\left(L^{(2)}(\mathbf{w}) + (1-L^{(2)}(\mathbf{e})^T\mathbf{w}), \max_{k=1\dots d} w_k\right), \quad \mathbf{w}\in\Delta_{d-1}$$

$$L^{(2)}(\mathbf{w}) = \frac{1}{n_2}\sum_{j=1}^{n_2}\left(\ell^{(2)}\left(\ell^{(1)}(\mathbf{w})\right)\right)_j$$

Let $\ell^{(1)}(\mathbf{w})$ have width n_1. Then the output of $\ell^{(1)}$ is given by the following vector

$$\ell^{(1)}(\mathbf{w}) = \begin{pmatrix} \mathbb{E}_{\boldsymbol{\theta}^{(1,1)}\sim\lambda^{(1,1)}}\left[\max_{k=1\dots d}\theta_k^{(1,1)}w_k\right] \\ \vdots \\ \mathbb{E}_{\boldsymbol{\theta}^{(1,n_1)}\sim\lambda^{(1,n_1)}}\left[\max_{k=1\dots d}\theta_k^{(1,n_1)}w_k\right] \end{pmatrix}, \quad \boldsymbol{\theta}\in[0,1]^d. \tag{9}$$

Each row in (9) is a $D-$norm where the expectation is taken over a delta function centered at $\boldsymbol{\theta}$, i.e. $\lambda^{(1,j)} = \delta(\boldsymbol{\theta}^{(1,j)})$ for $j=1,\dots,n_1$. Therefore, the property of $D-$norms is preserved for each row of (9). By analogy to $\ell^{(1)}$ in (9), the property

of $D-$norms is preserved for $\ell^{(2)}$ in (10):

$$\ell^{(2)}(\mathbf{w}) = \begin{pmatrix} \mathbb{E}_{\boldsymbol{\theta}^{(2,1)}\sim\lambda^{(2,1)}}\left[\max_{k=1\ldots n_1}\theta_k^{(2,1)}w_k\right] \\ \vdots \\ \mathbb{E}_{\boldsymbol{\theta}^{(2,n_2)}\sim\lambda^{(2,n_2)}}\left[\max_{k=1\ldots n_1}\theta_k^{(2,n_2)}w_k\right] \end{pmatrix}, \quad \boldsymbol{\theta}\in[0,1]^d. \tag{10}$$

We then use the nesting property of $D-$norms given in Hofert et al. [2018] such that $\ell^{(2)}(\ell^{(1)}(\boldsymbol{w}))$ is a $D-$norm, and by the same construction, $\ell^{(n)}(\ell^{(n-1)}(\cdots(\ell^{(1)}(\boldsymbol{w}))))$ is a $D-$norm and is thus $d-\max$-decreasing. Following the arguments in the 1-layer case for the output layer then completes the proof. □

C.2 UNIVERSAL APPROXIMATION

Our proof that our architecture is an universal approximator of Pickand's copula functions is constructive. Recall that every Pickands function has the form

$$A(\mathbf{w}) = \mathbb{E}_{\mathbf{s}\sim\lambda(\Delta_{d-1})}\left[\max_{k=1\ldots d} s_k w_k\right], \quad \mathbf{w}\in\Delta_{d-1}, \tag{11}$$

where λ is a spectral measure with $\text{supp}(\lambda)=\Delta_{d-1}$. We now construct a single layer dMNN, with width n, by sampling n independent and identically distributed (i.i.d.) samples $\mathbf{s}^{(1)},\ldots,\mathbf{s}^{(n)}\sim\lambda(\Delta_{d-1})$, and setting

$$\tilde{A}_n(\mathbf{w}) = \frac{1}{n}\sum_{j=1}^{n}\max_{k=1\ldots d}(s_k^{(j)}w_k) \tag{12}$$

Before showing that $\tilde{A}_n$ converges uniformly to A, we show that it converges point-wise. Although this intermediary result is not needed to show uniform converge, its proof provides intuition while being less technical.

The copula $\tilde{A}_n$ converges pointwise to A, almost surely.

Proof. Consider the discrete distribution $\mathbb{\Lambda}_n$ given by the n i.i.d. samples $\mathbf{s}^{(1)},\ldots,\mathbf{s}^{(n)}\sim\lambda(\Delta_{d-1})$:

$$\mathbb{\Lambda}_n := \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{s}^{(i)}),$$

where $\delta(\mathbf{s})$ represents a Dirac measure at $\mathbf{s}$. By the law of large numbers, for every $\mathbf{w}\in\Delta_{d-1}$,

$$\begin{aligned} \mathbb{E}_{\mathbf{s}\sim\mathbb{\Lambda}_n}\left[\max_{k=1\ldots d}s_k w_k\right] &\xrightarrow{a.s.} \mathbb{E}_{\mathbf{s}\sim\lambda}\left[\max_{k=1\ldots d}s_k w_k\right], \quad n\to\infty \\ \implies A_{\boldsymbol{\theta}}(\mathbf{w}) &\xrightarrow{a.s.} A(\mathbf{w}). \end{aligned}$$

□

We now state and prove the main result regarding uniform convergence.

The empirical process

$$\mathbb{G}_n = \sqrt{n}\left(\tilde{A}_n - A\right)$$

weakly converges to a zero-mean Gaussian process as $n\to\infty$ where $\tilde{A}_n$ is a single layer dMNN with width n.

Proof. Let λ be the law given by the spectral measure $\lambda(\Delta_{d-1})$ and the discrete empirical spectral measure be given by $\mathbb{\Lambda}_n := \frac{1}{n}\sum_{j=1}^{n}\delta(\mathbf{s}^{(j)})$ for n i.i.d. samples $\mathbf{s}^{(1)},\ldots,\mathbf{s}^{(n)}$ from λ. We additionally write $\lambda f := \mathbb{E}_{\mathbf{s}\sim\lambda}[f(\mathbf{s})]$ as the expectation with respect to the measure λ. The empirical process $\mathbb{G}_n$ is defined by

$$\begin{aligned} \mathbb{G}_n &= \sqrt{n}\left(\tilde{A}_n - A\right) \\ &= \sqrt{n}\left(\mathbb{\Lambda}_n - \lambda\right)f \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\max_{k=1\ldots d}(s_k^{(j)}w_k) - \mathbb{E}_{\mathbf{s}\sim\lambda}\left[\max_{k=1\ldots d}(s_k w_k)\right]\right), \end{aligned}$$

where $f \in \mathcal{F}$ and

$$\mathcal{F} := \{f_w(s) := \max_{k=1\ldots d}(s_k w_k) : \mathbf{w} \in \Delta_{d-1}\}.$$

By the classical central limit theorem, for a given $\mathbf{w}$, $\sqrt{n}\left(\tilde{A}_n(\mathbf{w}) - A(\mathbf{w})\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, with $\sigma^2 \leq (1 - A(\mathbf{w}))(A(\mathbf{w}) - \frac{1}{d^2})$, since the random variable $\max_{k=1\ldots d}(w_k s_k) \in [1/d^2, 1]$ is bounded and has finite variance.

Our claim is that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ where $\mathbb{G}$ is a zero-mean Gaussian process for establishing uniform convergence over $\mathbf{w}$. We will now show that the function class given by $\mathcal{F}$ is $\lambda-$Donsker. To show this, we will show that the bracketing integral given by

$$\mathcal{J}_{[]}(1, \mathcal{F}, L_2(\lambda)) = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F} \cup 0, L_2(\lambda))}\, \mathrm{d}\epsilon \tag{13}$$

converges where the $L_2(\lambda)$ norm is defined as $\|f\|_{\lambda,2} = (\int f^2 d\lambda)^{1/2}$. A sufficient condition for convergence of (13) is to show that the logarithm of the bracketing number $N_{[]}$ grows at a rate slower than $O(\frac{1}{\epsilon^2})$. The function class $\mathcal{F}$ is indexed by $\mathbf{w} \in \Delta_{d-1}$ and is Lipschitz on $\mathbf{w}$. From Sen [2018, Lemma 2.14], the bracketing number of $\mathcal{F}$ is thus bounded above by the covering number of Δ_{d-1}, i.e.

$$N_{[]}(2\epsilon, \mathcal{F}, L_2(\lambda)) \leq N(\epsilon, \Delta_{d-1}, \|\cdot\|_2),$$

where the covering number of the unit simplex is asymptotically $O\left(\frac{1}{\epsilon^{d-1}}\right)$. The logarithm of the bracketing number then grows at a rate $\log N_{[]} \leq O\left((d-1)\log\frac{1}{\epsilon}\right) < O(\frac{1}{\epsilon^2})$. This proves that $\mathcal{F}$ is $\lambda-$Donsker and thus $\mathbb{G}_n \rightsquigarrow \mathbb{G}$. □

We can also make a statement on the covariance of the process. For any two points $w^{(i)}, w^{(j)}$, the covariance of the Gaussian process converges to

$$\begin{aligned}
Cov\left(A\left(w^{(i)}\right), A\left(w^{(j)}\right)\right) &= \mathbb{E}\left[\max_k s_k w_k^{(i)} \max_k s_k w_k^{(j)}\right] - \mathbb{E}\left[\max_k s_k w_k^{(i)}\right]\mathbb{E}\left[\max_k s_k w_k^{(j)}\right] \\
&= \mathbb{E}\left[\max_k s_k w_k^{(i)} \max_k s_k w_k^{(j)}\right] - A(w^{(i)})A(w^{(j)}) \\
&\leq \sqrt{Var\left(A\left(w^{(i)}\right)\right) Var\left(A\left(w^{(j)}\right)\right)} \\
&\leq \frac{1}{4}\left(1 - \frac{1}{d}\right)^2.
\end{aligned}$$

The final inequality comes from Popoviciu's inequality and the fact that $A(w) \in [1/d, 1]$. For an introduction to empirical processes, see the review given in Wellner [2005].

D SURVIVAL PROBABILITY ESTIMATION

One particularly useful task is estimating multi-dimensional survival probabilities rather than cumulative probabilities. More precisely, let $(\gamma_1, \cdots, \gamma_d) \in \mathbb{R}^d$ be a $d-$dimensional vector of thresholds, we are interested in calculating the following survival probability:

$$\mathbb{P}\left[M_n^{(1)} > \gamma_1, \cdots, M_n^{(d)} > \gamma_d\right] = \mathbb{P}\left[\bar{M}_n^{(1)} > \bar{\gamma}_1, \cdots, \bar{M}_n^{(d)} > \bar{\gamma}_d\right], \tag{14}$$

where $\bar{\gamma}_k = \frac{\gamma_k - b_n^{(k)}}{a_n^{(k)}}$, $k \in \{1, \cdots, d\}$.

To calculate this, we simply use a change-of-variable technique which we present in the following proposition. This approach is well known, and we only provide the proposition for completeness.

Proposition 2 (Survival Probability Computation)**.** *Let $G_k(x) := F_k^{-1}(1 - F_k(x))$ for $k \in \{1, \ldots, d\}$, then the random variables $G_k(\bar{M}_n^{(k)})$ and $\bar{M}_n^{(k)}$ have the same marginal CDF F_k, for $k \in \{1, \ldots, d\}$, and*

$$\mathbb{P}\left[\bar{M}_n^{(1)} > \bar{\gamma}_1, \cdots, \bar{M}_n^{(d)} > \bar{\gamma}_d\right] = \mathbb{P}\left[G_1(\bar{M}_n^{(1)}) < G_1(\bar{\gamma}_1), \cdots, G_d(\bar{M}_n^{(d)}) < G_d(\bar{\gamma}_d)\right]. \tag{15}$$

Proof. With the change-of-variable $G_k(x) := F_k^{-1}(1 - F_k(x))$ for $k \in \{1, \cdots, d\}$, it first follows that the random variables $G_k(\bar{M}_n^{(k)}) \sim F_k$:

$$\mathbb{P}(G_k(\bar{M}_n^{(k)}) \leq x) = \mathbb{P}(F_k^{-1}(1 - F_k(\bar{M}_n^{(k)})) \leq x) = \mathbb{P}(1 - F_k(\bar{M}_n^{(k)}) \leq F_k(x)) = F_k(x), \tag{16}$$

since $F_k(\bar{M}_n^{(k)})$ and $1 - F_k(\bar{M}_n^{(k)})$ follow the unit uniform distribution.

Moreover, the survival probability can be written as:

$$\begin{aligned}
\mathbb{P}\left[\bar{M}_n^{(1)} > \bar{\gamma}_1, \cdots, \bar{M}_n^{(d)} > \bar{\gamma}_d\right] &= \mathbb{P}\left[1 - F_1(\bar{M}_n^{(1)}) < 1 - F_1(\bar{\gamma}_1), \cdots, 1 - F_d(\bar{M}_n^{(d)}) < 1 - F_d(\bar{\gamma}_d)\right] \\
&= \mathbb{P}\left[G_1(\bar{M}_n^{(1)}) < G_1(\bar{\gamma}_1), \cdots, G_d(\bar{M}_n^{(d)}) < G_d(\bar{\gamma}_d)\right] \\
&= C\left(1 - F_1(\bar{\gamma}_1), \cdots, 1 - F_d(\bar{\gamma}_d)\right),
\end{aligned}$$

where C is the copula of $\left(G_1(\bar{M}_n^{(1)}), \cdots, G_d(\bar{M}_n^{(d)})\right)$. □

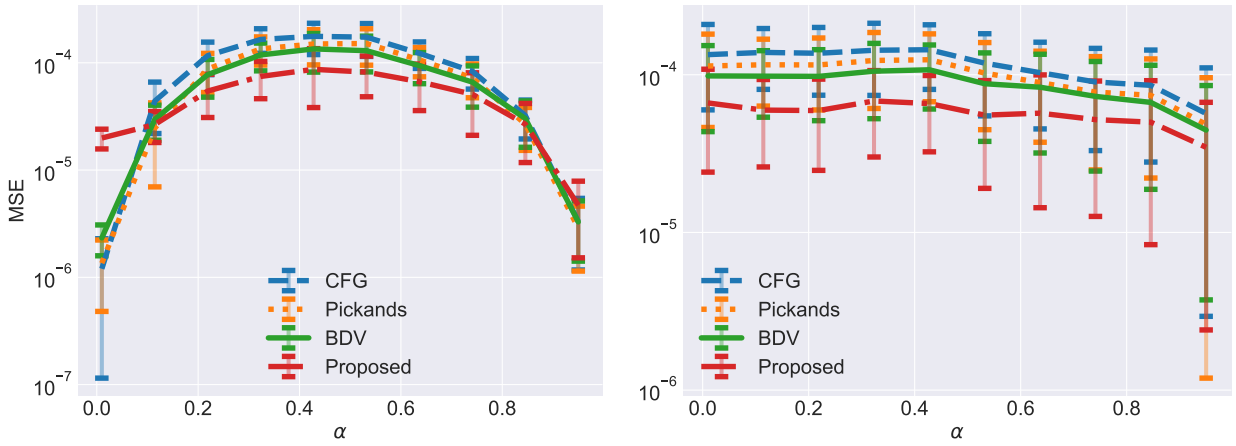
This proposition implies that the transformed variables $G_k(\bar{M}_n^{(k)})$ are samples from extreme value distributions. Then, we can fit Pickands dependence function to these transformed variables, and finally evaluate the corresponding extreme value copula on $(1 - F_1(\bar{\gamma}_1), \cdots, 1 - F_d(\bar{\gamma}_d))$. Details on how to estimate the survival probability in (14) are given in Algorithm 2.

E ADDITIONAL EXPERIMENTS AND FIGURES

Code for all experiments can be found at `https://github.com/alluly/dMNN`.

E.1 24 WIDTH 3 DEPTH ARCHITECTURE

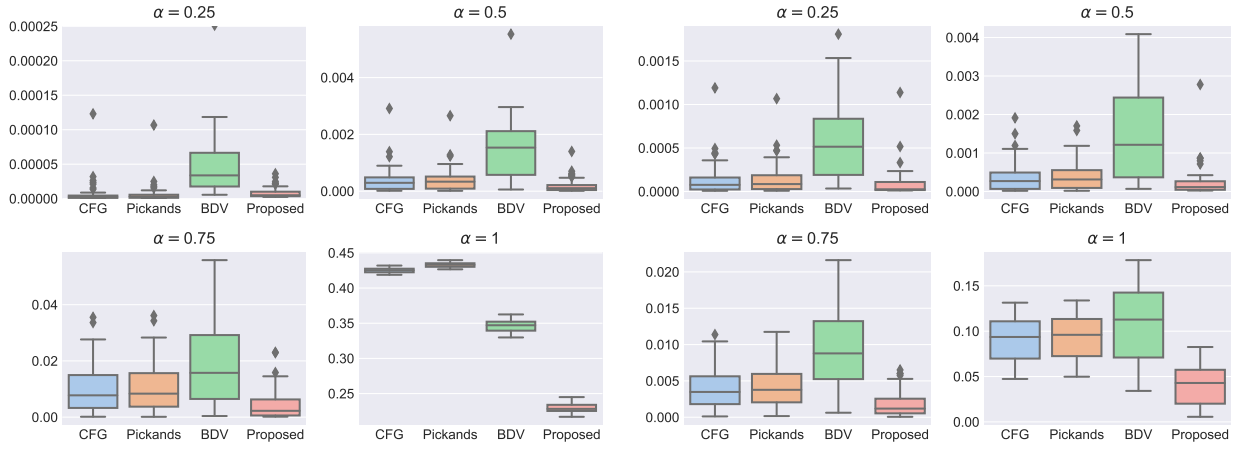
Here we repeat the experiments with a different architecture. All other hyperparameters are the same, the only difference is we increase the depth to 3 and use a width of 24 for each layer. Most of the results remain similar for the synthetic data but we see a change in the results for the real data, specifically, the Wind and Commodities data show a slight deterioration in performance. However, the variances are still high for the real experiments and not much can be said regarding the efficacy of any single method.



(a) A_{SL} MSE ($d = 2$)

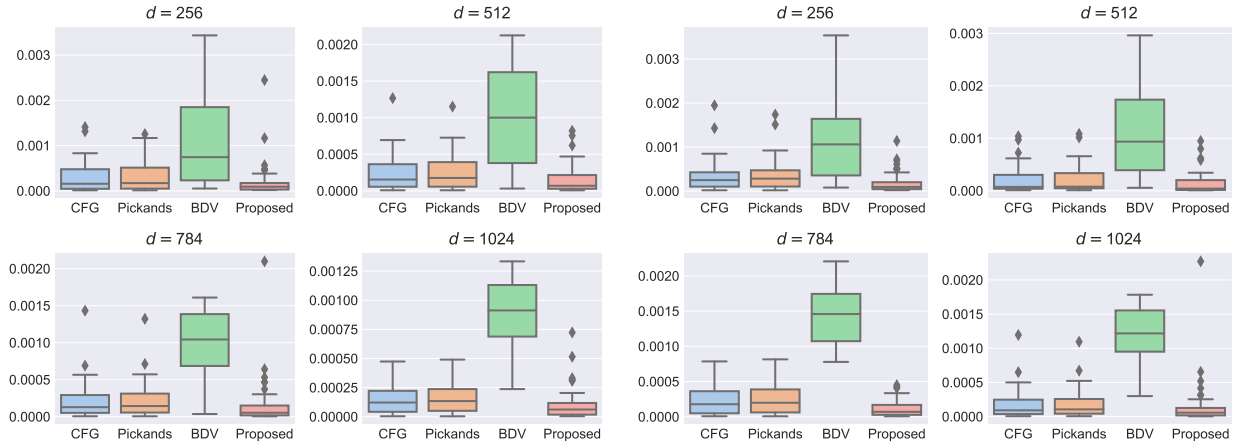
(b) A_{ASL} MSE ($d = 2$)

Figure 1: Using 24 width 3 depth architecture: MSE of survival probabilities for $d = 2$ with 100 samples for A_{SL} (1a) and A_{ASL} (1b). Thresholds are above the 75th percentile.



(a) A_{SL} MSE ($d = 256$)

(b) A_{ASL} MSE ($d = 256$)



(c) A_{SL} MSE ($\alpha = 0.5$)

(d) A_{ASL} MSE ($\alpha = 0.5$)

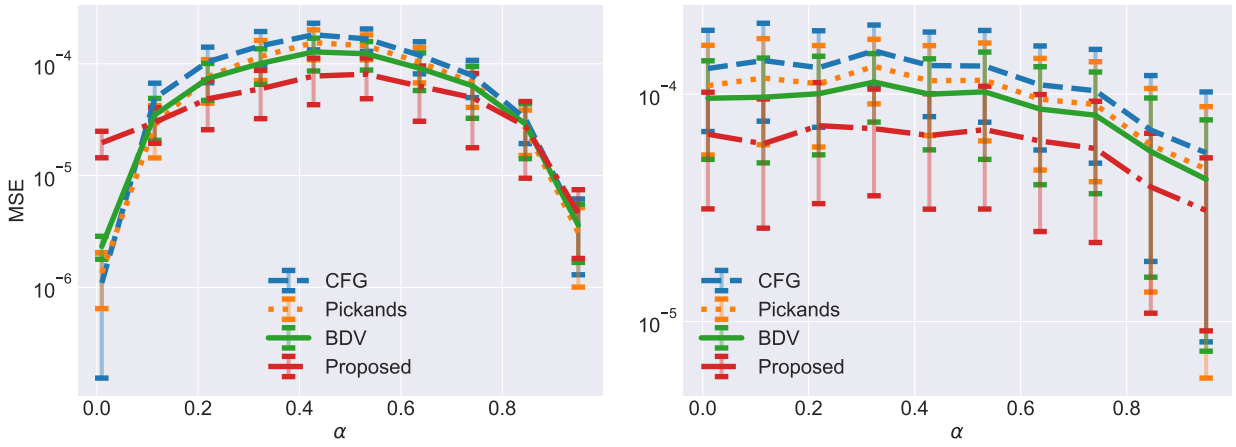
Figure 2: Using 24 width 3 depth architecture: Comparison of $||\hat{A}(w) - A(w)||_2^2$ for different estimators $\hat{A}$ for different dependence $\alpha = \{0.25, 0.50, 0.75, 1.0\}$ with fixed $d = 256$ (2a, 2b) and for fixed $\alpha = 0.5$ with different $d = \{256, 512, 728, 1024\}$ (2c, 2d). The truth models considered are A_{SL} (2a, 2c) and A_{ASL} (2b, 2d). Results are over 50 runs with 100 training samples for each run.

	d	Train/Test Length	PICKANDS	CFG	BDV	PROPOSED
Wind	10	day/week	$4.48(18.6)_{\times 10^{-4}}$	*4.15(15.1)*$_{\times 10^{-4}}$	$\mathbf{4.10(16.3)}_{\times 10^{-4}}$	$4.80(20.6)_{\times 10^{-4}}$
Ozone	4	day/week	$3.06(4.66)_{\times 10^{-2}}$	$3.86(6.10)_{\times 10^{-2}}$	*2.86(4.46)*$_{\times 10^{-2}}$	$\mathbf{2.82(4.38)}_{\times 10^{-2}}$
Commodities	10	week/month	$4.34(5.82)_{\times 10^{-3}}$	$4.33(5.71)_{\times 10^{-3}}$	$\mathbf{1.60(1.96)}_{\times 10^{-3}}$	*2.20(3.41)*$_{\times 10^{-3}}$
S&P 500	418	week/month	*3.02(21.2)*$_{\times 10^{-3}}$	*3.02(21.1)*$_{\times 10^{-3}}$	$6.28(35.2)_{\times 10^{-3}}$	$\mathbf{2.41(22.2)}_{\times 10^{-3}}$
Crypto	100	week/month	$1.06(2.85)_{\times 10^{-2}}$	*1.05(4.86)*$_{\times 10^{-2}}$	$1.34(3.44)_{\times 10^{-2}}$	$\mathbf{8.42(26.1)}_{\times 10^{-3}}$
COVID (NC)	100	week/week	$4.04(7.21)_{\times 10^{-2}}$	$4.04(7.19)_{\times 10^{-2}}$	*3.83(6.51)*$_{\times 10^{-2}}$	$\mathbf{5.22(12.8)}_{\times 10^{-3}}$
COVID (NY)	58	week/week	$2.74(10.4)_{\times 10^{-2}}$	$2.74(10.4)_{\times 10^{-2}}$	*2.25(7.75)*$_{\times 10^{-2}}$	$\mathbf{3.52(7.92)}_{\times 10^{-3}}$
COVID (CA)	58	week/week	$1.17(3.98)_{\times 10^{-2}}$	*1.19(3.87)*$_{\times 10^{-2}}$	$1.17(3.85)_{\times 10^{-2}}$	$\mathbf{3.27(9.28)}_{\times 10^{-3}}$

Table 1: MSE of different estimators in estimating maxima over two time scales for 24 width 3 depth architecture. Best and second best performances are marked in **bold** and *italic* respectively.

E.2 64 WIDTH 4 DEPTH ARCHITECTURE

Here we repeat the experiments with a different architecture. All other hyperparameters are the same, the only difference is we increase the depth to 4 and use a width of 64 for each layer. Most of the results remain similar for the synthetic data but we see a change in the results for the real data, specifically, the Wind and Commodities data show a slight deterioration in performance. However, the variances are still high for the real experiments and not much can be said regarding the efficacy of any single method.



(a) A_{SL} MSE ($d = 2$)

(b) A_{ASL} MSE ($d = 2$)

Figure 3: Using 64 width 4 depth architecture: (3a, 3b) MSE of survival probabilities for $d = 2$ with 100 samples for A_{SL} (3a) and A_{ASL} (3b). Thresholds are above the 75th percentile.

	d	Train/Test Length	PICKANDS	CFG	BDV	PROPOSED
Wind	10	day/week	$4.48(18.6)_{\times 10^{-4}}$	*4.15(15.1)*$_{\times 10^{-4}}$	$\mathbf{4.10(16.3)}_{\times 10^{-4}}$	$4.76(18.7)_{\times 10^{-4}}$
Ozone	4	day/week	$3.06(4.66)_{\times 10^{-2}}$	$3.86(6.10)_{\times 10^{-2}}$	*2.86(4.46)*$_{\times 10^{-2}}$	$\mathbf{2.73(4.25)}_{\times 10^{-2}}$
Commodities	10	week/month	$4.34(5.82)_{\times 10^{-3}}$	$4.33(5.71)_{\times 10^{-3}}$	$\mathbf{1.60(1.96)}_{\times 10^{-3}}$	*2.20(3.44)*$_{\times 10^{-3}}$
S&P 500	418	week/month	*3.02(21.2)*$_{\times 10^{-3}}$	*3.02(21.1)*$_{\times 10^{-3}}$	$6.28(35.2)_{\times 10^{-3}}$	$\mathbf{2.39(22.1)}_{\times 10^{-3}}$
Crypto	100	week/month	$1.06(2.85)_{\times 10^{-2}}$	*1.05(4.86)*$_{\times 10^{-2}}$	$1.34(3.44)_{\times 10^{-2}}$	$\mathbf{8.28(25.6)}_{\times 10^{-3}}$
COVID (NC)	100	week/week	$4.04(7.21)_{\times 10^{-2}}$	$4.04(7.19)_{\times 10^{-2}}$	*3.83(6.51)*$_{\times 10^{-2}}$	$\mathbf{4.93(12.1)}_{\times 10^{-3}}$
COVID (NY)	58	week/week	$2.74(10.4)_{\times 10^{-2}}$	$2.74(10.4)_{\times 10^{-2}}$	*2.25(7.75)*$_{\times 10^{-2}}$	$\mathbf{3.69(8.64)}_{\times 10^{-3}}$
COVID (CA)	58	week/week	$1.17(3.98)_{\times 10^{-2}}$	*1.19(3.87)*$_{\times 10^{-2}}$	$1.17(3.85)_{\times 10^{-2}}$	$\mathbf{1.10(4.78)}_{\times 10^{-3}}$

Table 2: MSE of different estimators in estimating maxima over two time scales for 64 width 4 depth architecture. Best and second best performances are marked in **bold** and *italic* respectively.

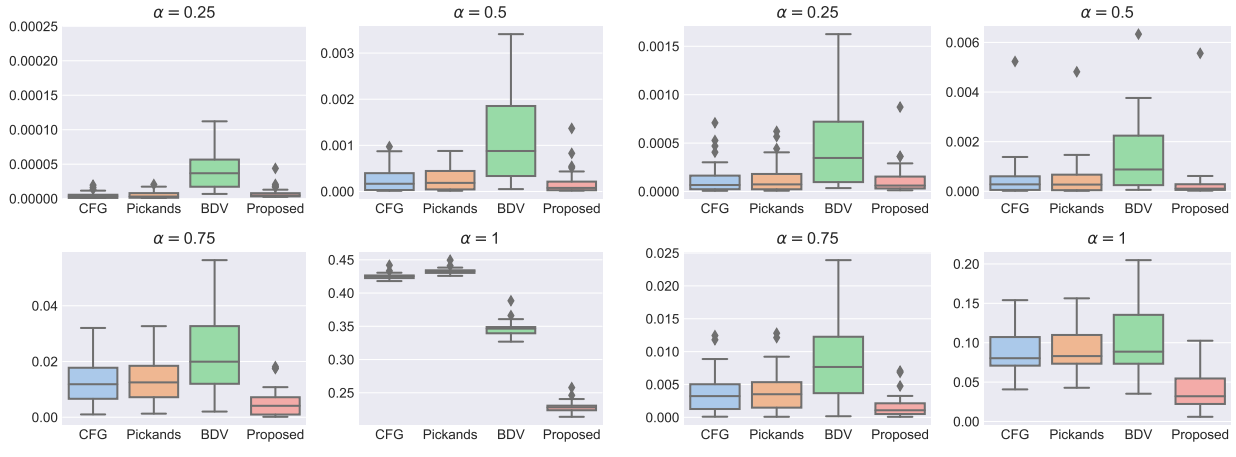
E.3 ESTIMATION COMPARISON

We finally add a few figures comparing the learned dependence functions between different architectures. We additionally provide a table comparing the results for different architectures on the real data experiments in Table 3.

F DATA DESCRIPTION

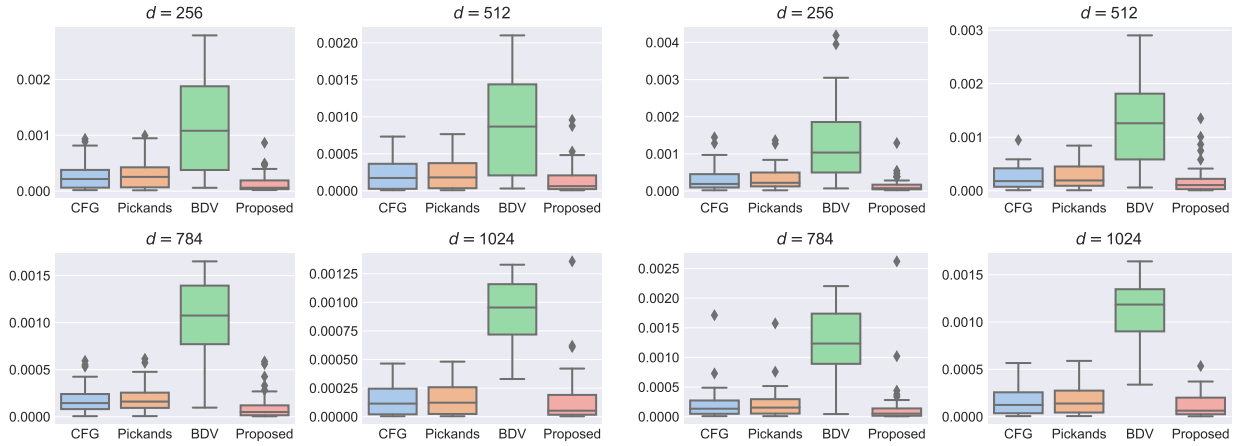
Synthetic Data

For the synthetic data experiments we consider samples of 100 points from each respective distribution. We use the full dataset for the batch size during training. We additionally sample 1000 points from the simplex for each data point during training.



(a) A_{SL} MSE ($d = 256$)

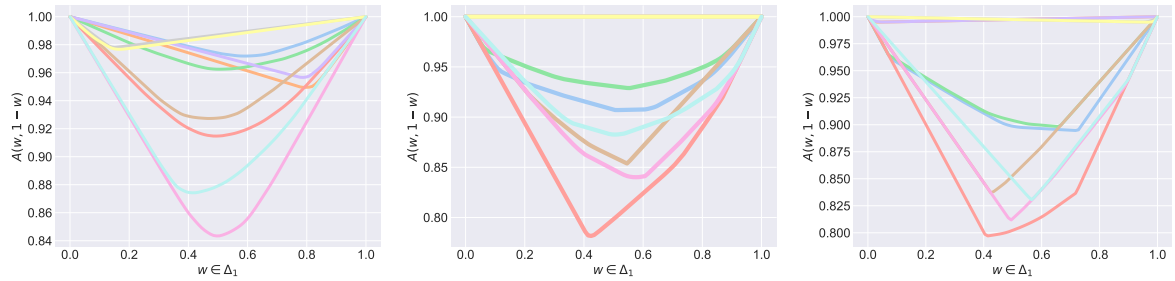
(b) A_{ASL} MSE ($d = 256$)



(c) A_{SL} MSE ($\alpha = 0.5$)

(d) A_{ASL} MSE ($\alpha = 0.5$)

Figure 4: Using 64 width 4 depth architecture: Comparison of $||\hat{A}(w) - A(w)||_2^2$ for different estimators $\hat{A}$ for different dependence $\alpha = \{0.25, 0.50, 0.75, 1.0\}$ with fixed $d = 256$ (4a, 4b) and for fixed $\alpha = 0.5$ with different $d = \{256, 512, 728, 1024\}$ (4c, 4d) for A_{SL} (4a, 4c) and A_{ASL} (4b, 4d). Results are over 50 runs with 100 training samples for each run.



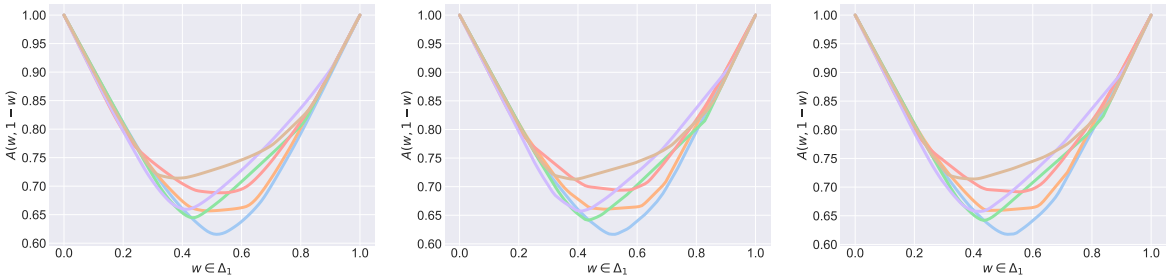
(a) 512 width 1 depth 2d Margins (b) 24 width 3 depth 2d Margins (c) 64 width 4 depth 2d Margins

Figure 5: Margin comparison for winds dataset.

Ozone Data

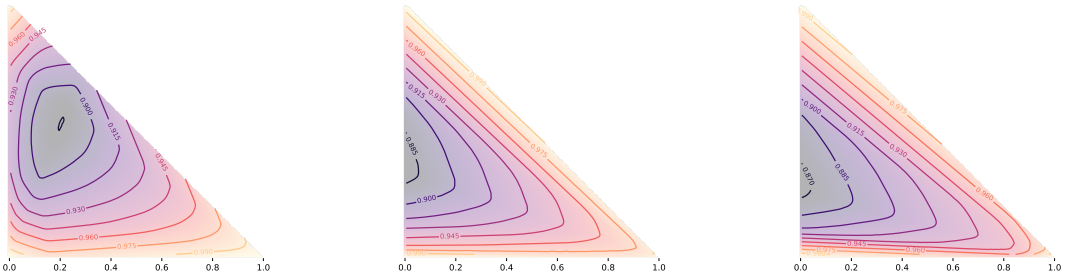
We consider ozone levels measured at 4 different stations in Sequoia National Park from data that can be downloaded from the National Park Service website [1]. The 4 stations are located at Ash Mountain, Lower Kaweah, Grant Grove and

[1] `https://ard-request.air-resource.com/data.aspx`



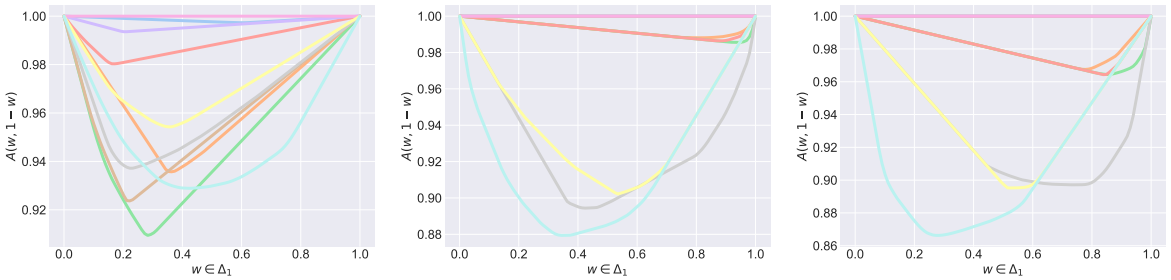
(a) 512 width 1 depth 2d Margins (b) 24 width 3 depth 2d Margins (c) 64 width 4 depth 2d Margins

Figure 6: Margin comparison for ozone dataset.



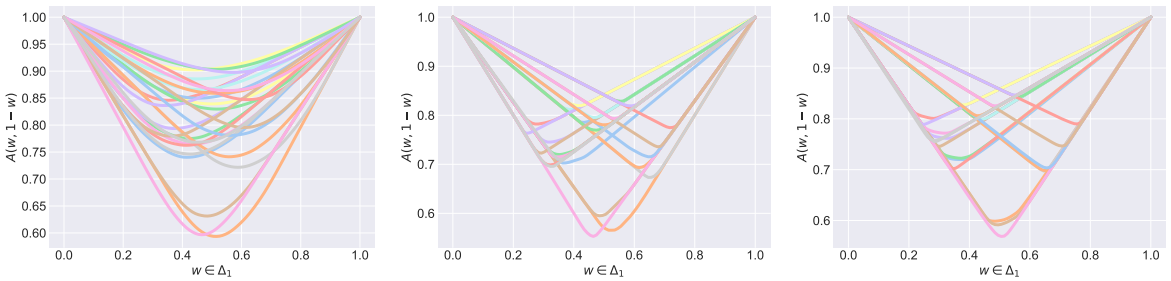
(a) 512 width 1 depth 3d Margins (b) 24 width 3 depth 3d Margins (c) 64 width 4 depth 3d Margins

Figure 7: Margin comparison for commodities dataset.



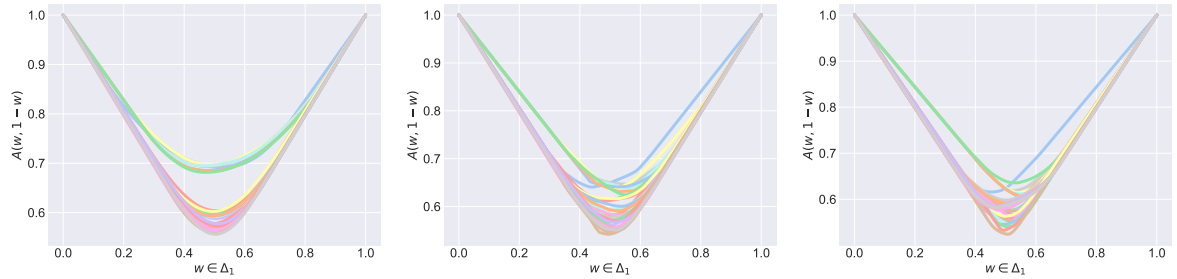
(a) 512 width 1 depth 2d Margins (b) 24 width 3 depth 2d Margins (c) 64 width 4 depth 2d Margins

Figure 8: Margin comparison for commodities dataset.



(a) 512 width 1 depth 2d Margins (b) 24 width 3 depth 2d Margins (c) 64 width 4 depth 2d Margins

Figure 9: Margin comparison for S&P 500 dataset.



(a) 512 width 1 depth 2d Margins (b) 24 width 3 depth 2d Margins (c) 64 width 4 depth 2d Margins

Figure 10: Margin comparison of different architectures for Cryptocurrencies dataset.

	512width $\times$ 1depth	24width $\times$ 3depth	64width $\times$ 4depth
Wind	$4.37(17.5)_{\times 10^{-4}}$	$4.80(20.6)_{\times 10^{-4}}$	$4.76(18.7)_{\times 10^{-4}}$
Ozone	$2.73(4.25)_{\times 10^{-2}}$	$2.82(4.38)_{\times 10^{-2}}$	$2.73(4.25)_{\times 10^{-2}}$
Commodities	$1.56(2.21)_{\times 10^{-3}}$	$2.20(3.41)_{\times 10^{-3}}$	$2.20(3.44)_{\times 10^{-3}}$
S & P	$2.41(22.2)_{\times 10^{-3}}$	$2.41(22.2)_{\times 10^{-3}}$	$2.39(22.1)_{\times 10^{-3}}$
Crypto	$8.57(26.4)_{\times 10^{-3}}$	$8.42(26.1)_{\times 10^{-3}}$	$8.28(25.6)_{\times 10^{-3}}$

Table 3: Comparison of 3 different architectures, in terms of MSE, on the real data experiments.

Lookout Point. We train the different models on daily maxima of ozone levels at the 4 different stations for the period from January 1984 to December 1996. To reduce the effect of seasonality, we do not train over the whole period, but we train different models on a single month (training month, e.g. June of each year) and compute accuracy on the consecutive month (validation month e.g. July of the same year). We additionally only look at summer months due to the increase of extreme events during that time. The accuracy is averaged with the specific validation month of each year over the whole period. We train on daily maxima and test on weekly maxima. For the experiments, we consider the following pair of (training/test) months: (June/July), (July/August), and (August/September).

California Wind Data

We are interested in modeling the extremal relationship of wind gusts between different locations in California during the summer months. We consider 10 locations in California illustrated in 11. We obtained the data from the Remote Automated Weather Station (RAWS) archive available at the online repository[2]. The RAWS data are collected from various time intervals from December 1989 to December 2020. We consider only the time points that occur in the intersection of all the data collected and where all values are valid (i.e. not NaNs or missing) for the summer months. Similarly to the ozone data, and in an effort to reduce seasonality, we consider the daily max wind gust for the different locations for a single month over all the years the data were collected. To evaluate the proposed method, we train and test on data from consecutive months and repeat for multiple sets of months in our dataset. Additionally, we train on daily max and test on monthly max using the following data splitting scheme (training/validation months): (June/July), (July/August), and (August/September).

Commodities Data

We consider the extreme dependency between different commodities such as Coffee, Copper, Corn, Crude Oil, Gold, Heating Oil, Natural Gas, Platinum, Silver and Wheat. We collect data of daily prices of the different commodities from January 2015 to December 2020 as published in [3]. For training, we consider weekly max drawdown over a year. We validate the performance by evaluating accuracy of monthly max drawdown over next three years. We consider the following pairs of ([training years],[validation years]): ([2015], [2016, 2017, 2018]), ([2016], [2017, 2018, 2019]), ([2017], [2018, 2019, 2020]).

[2] https://raws.dri.edu/index.html

[3] https://www.investing.com/commodities/

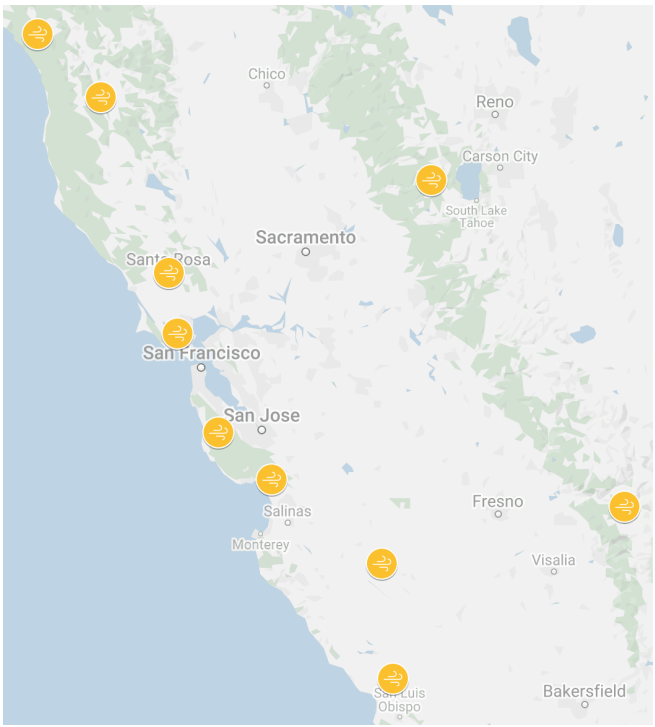


Figure 11: Locations (yellow circles) of weather stations sampled for the wind speed experiments. Figure generated via Google Maps.

S&P 500 Data

We obtain historical data from `https://www.alphavantage.co`[4]. We choose the components of the S&P 500 with sufficient history (resulting in 418 stocks). For training, we consider weekly max drawdown over a year. We validate the performance by evaluating accuracy of monthly max drawdown over next three years. We consider the following pairs of ([training years],[validation years]): ([2015], [2016, 2017, 2018]), ([2016], [2017, 2018, 2019]), ([2017], [2018, 2019, 2020]). For the full list of stocks, see the `sp_names.txt` file in the supplementary materials.

Cryptocurrencies Data

We obtain historical data from `https://coinmarketcap.com`[5] for 100 coins with the longest history. For training, we consider weekly max drawdown over a year. We validate the performance by evaluating accuracy of monthly max drawdown over next three years. We consider the following pairs of ([training years],[validation years]): ([2015], [2016, 2017, 2018]), ([2016], [2017, 2018, 2019]), ([2017], [2018, 2019, 2020]). For the full list of coins, see the `crypto_names.txt` file in the supplementary materials.

COVID-19 Data

We obtained COVID-19 case counts for the United States at the county level from `https://github.com/nytimes/covid-19-data`. We chose the states of North Carolina, New York, and California for the analysis. We consider training on weeks in 2020 and testing on weeks in 2021. We choose the same time scale due to the fact that cases were poorly counted in 2020 whereas in 2021 case counts were more accurately reported. Additionally, 2021 saw an increase in cases due to the arrival of new variants such as Delta and Omicron. For the conditional classification examples, we take the observation at the locations and compute the probability. If the probability is reported as > 0.5 then we consider it to be classified correctly. If it is < 0.5 then we consider it an incorrect classification. For the North Carolina data, we condition

[4]Alpha Vantage allows academic use as long as the website is cited.

[5]Coin Market Cap allows academic use as long as the website is cited (see FAQ page).

on the counties of: Mecklenburg, Wake, Guilford, Forsyth, and Cumberland and predict Durham. For the New York data, we condition on Westchester, Nassau, New York City, Suffolk, and Erie and predict Monroe (the data aggregated all the NYC counties into a single datapoint). For the California data, we condition on Los Angeles, San Diego, San Bernardino, Riverside, and Orange and predict Santa Clara.

G PICKANDS, CFG AND BDV ESTIMATORS

Pickands Estimator

The Pickands estimator Pickands [1981] is built following the transformations (6) and (7) in the paper. The estimator is obtained by exactly maximizing the likelihood (Equation (8) in the paper) resulting in the following non-parametric estimate:

$$\widehat{A}_{\text{Pickands}}(\mathbf{w}) = \left(\frac{1}{B} \sum_{i=1}^{B} Z_{w,i} \right)^{-1}. \tag{17}$$

CFG Estimator

The CFG estimator Capéraà et al. [1997] is constructed following the observation:

$$\mathbb{E} \log Z_w = -\log A(\mathbf{w}) - \gamma,$$

where $\gamma = -\int_0^\infty \log x e^{-x} dx$ denotes the Euler's contant. The CFG estimator is thus given by:

$$\widehat{A}_{\text{CFG}}(\mathbf{w}) = \exp\left[-\gamma - \frac{1}{B} \sum_{i=1}^{B} \log Z_{w,i} \right]. \tag{18}$$

In our main submission we use a similar estimator, with the correction term presented in Gudendorf and Segers [2011]:

$$\widehat{A}_{\text{CFG,C}}(\mathbf{w}) = \exp\left(\log \widehat{A}_{\text{CFG}}(\mathbf{w}) - \sum_{k=1}^{d} w_k \log\left(\widehat{A}_{\text{CFG}}(\mathbf{e}_k) \right) \right), \tag{19}$$

where $\mathbf{e}_k$ is the k-th canonical basis vector.

BDV Estimator

We propose an d-dimensional extension to the bivariate estimator described in Bücher et al. [2011]. We begin by defining the minimum distance estimator between the true CDF, $C(\mathbf{u})$ and the one estimated by the Pickands function $A(\mathbf{w})$.

$$\int_{[0,1]^d} \left[\log C(\mathbf{u}) - \sum_{k=1}^{d} \log u_k A\left(\frac{\log(\mathbf{u})}{\sum_k \log u_k} \right) \right]^2 d\mathbf{u} \tag{20}$$

$$= \int_{\Delta_{d-1}} \int_0^1 (\log C(y^{w_1}, \ldots, y^{w_d}) - \log(y) A(\mathbf{w}))^2 (-\log(y))^{d-1} \, dy \, d\mathbf{w}. \tag{21}$$

We have

$$\hat{C}(y^{w_1}, \ldots, y^{w_d}) = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}(F_1(\bar{M}_{n,i}^{(1)}) \leq y^{w_1}, \cdots, F_d(\bar{M}_{n,i}^{(d)}) \leq y^{w_d}) \tag{22}$$

$$= \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}(F_1(\bar{M}_{n,i}^{(1)})^{\frac{1}{w_1}} \leq y, \cdots, F_d(\bar{M}_{n,i}^{(d)})^{\frac{1}{w_d}} \leq y) \tag{23}$$

$$= \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}(\max_{1 \leq k \leq d} F_k(\bar{M}_{n,i}^{(k)})^{\frac{1}{w_k}} \leq y) \tag{24}$$

$$= \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}\left(\Gamma_{w,i} \leq y \right), \tag{25}$$

where $\Gamma_{w,i} = \exp(-Z_{w,i})$. Now, if we reorder these so that $\Gamma_{w,1} \leq \cdots \leq \Gamma_{w,B}$, we have that

$$\hat{C}(y^{w_1}, \ldots, y^{w_d}) = \begin{cases} 0 \text{ if } y < \Gamma_{w,1}, \\ \dfrac{i}{B} \text{ if } \Gamma_{w,i} \leq y < \Gamma_{w,i+1},\ i \in \{1, \ldots, B-1\}, \\ 1 \text{ if } \Gamma_{w,B} \leq y. \end{cases} \tag{26}$$

Because $\log \hat{C}(\cdots)$ is not defined if $y < \Gamma_{w,1}$, the following modified estimator is considered in Bücher et al. [2011].

$$\tilde{C}(y^{w_1}, \ldots, y^{w_d}) := \max\left\{C(y^{w_1}, \ldots, y^{w_d}), B^{-\gamma}\right\}, \tag{27}$$

where γ is any positive real greater or equal than $\frac{1}{2}$. For convenience, we choose $\gamma = 1$ so that:

$$\tilde{C}(y^{w_1}, \ldots, y^{w_d}) = \begin{cases} \dfrac{1}{B} \text{ if } y < \Gamma_{w,2}, \\ \dfrac{i}{B} \text{ if } \Gamma_{w,i} \leq y < \Gamma_{w,i+1},\ i \in \{2, \ldots, B-1\}, \\ 1 \text{ if } \Gamma_{w,B} \leq y. \end{cases} \tag{28}$$

Finally, as in Bücher et al. [2011], for any positive weight function $h : (0,1) \to \mathbb{R}_0^+$, let $h^*(y) := h(y)(\log y)^2$,

$$B_h := \int_0^1 h^*(y)\, dy \quad \text{and} \quad g(x) := -B_h^{-1} \int_0^x \frac{h^*(y)}{\log y}\, dy. \tag{29}$$

Then, letting $\Gamma_{w,0} = 0$, $\Gamma_{w,B+1} = 1$, we define the BDV estimator $\widehat{A}_{\text{BDV},h}$ as follows

$$\widehat{A}_{\text{BDV},h}(\mathbf{w}) = B_h^{-1} \int_0^1 \frac{\log \tilde{C}(y^{w_1}, \ldots, y^{w_d})}{\log y} h^*(y)\, dy \tag{30}$$

$$= B_h^{-1} \sum_{i=0}^{B} \int_{\Gamma_{w,i}}^{\Gamma_{w,i+1}} \frac{\log \tilde{C}(y^{w_1}, \ldots, y^{w_d})}{\log y} h^*(y)\, dy \tag{31}$$

$$= -\log \frac{1}{n} g(\Gamma_{w,2}) - \sum_{i=2}^{n} \log \frac{i}{n} \left(g(\Gamma_{w,i+1}) - g(\Gamma_{w,i})\right) \tag{32}$$

$$= -\sum_{i=2}^{n} \log \frac{i-1}{n} g(\Gamma_{w,i}) + \sum_{i=2}^{n} \log \frac{i}{n} g(\Gamma_{w,i}) \tag{33}$$

$$= \sum_{i=2}^{n} \log\left(1 + \frac{1}{i-1}\right) g(\Gamma_{w,i}) \tag{34}$$

In our main submission, we use a slightly modified estimator, which proved to have superior performance in our experiments. Recall that, if A is a Pickand's dependence function, we have $\max(\mathbf{w}) \leq A(\mathbf{w}) \leq 1$, which implies the true copula verifies:

$$\max(\mathbf{w}) \leq \frac{\log C(y^{w_1}, \ldots, y^{w_d})}{\log y} = A(\mathbf{w}) \leq 1. \tag{35}$$

Accordingly, we let

$$\text{clamp}_{a,b}(x) := \begin{cases} a & \text{if } x \leq a, \\ x & \text{if } a < x < b, \\ b & \text{if } x \geq b, \end{cases} \tag{36}$$

and define

$$\breve{C}(y^{w_1}, \ldots, y^{w_d}) = \exp\left(\text{clamp}_{\log y, \max(\mathbf{w}) \log y} \log \hat{C}(y^{w_1}, \ldots, y^{w_d})\right), \tag{37}$$

and

$$\widehat{A}_{\text{BDV,MM},h}(\mathbf{w}) = B_h^{-1} \int_0^1 \frac{\log \breve{C}(y^{w_1}, \ldots, y^{w_d})}{\log y} h^*(y)\, dy \tag{38}$$

$$= B_h^{-1} \sum_{i=0}^{B} \int_{\Gamma_{w,i}}^{\Gamma_{w,i+1}} \frac{\log \breve{C}(y^{w_1}, \ldots, y^{w_d})}{\log y} h^*(y)\, dy \tag{39}$$

Letting $\Gamma_{w,i}^{(\ell)} = \mathrm{clamp}_{\Gamma_{w,i},\Gamma_{w,i+1}}\left(\left(\frac{i}{n}\right)^{\frac{1}{\max(\mathbf{w})}}\right)$, $\Gamma_{w,i}^{(u)} = \mathrm{clamp}_{\Gamma_{w,i},\Gamma_{w,i+1}}\left(\frac{i}{n}\right)$ for $i \in \{0,\dots,B\}$ and $\eta(x) = B_h^{-1}\int_0^x h^*(y)\,dy$, we have

$$\begin{aligned}
\widehat{A}_{\text{BDV,MM},h}(\mathbf{w}) &= B_h^{-1}\sum_{i=0}^{B}\int_{\Gamma_{w,i}}^{\Gamma_{w,i+1}}\frac{\log \breve{C}(y^{w_1},\dots,y^{w_d})}{\log y}h^*(y)\,dy \\
&= B_h^{-1}\sum_{i=0}^{B}\int_{\Gamma_{w,i}}^{\Gamma_{w,i}^{(\ell)}}\max(\mathbf{w})h^*(y)\,dy + \int_{\Gamma_{w,i}^{(\ell)}}^{\Gamma_{w,i}^{(u)}}\log\frac{i}{n}\frac{h^*(y)}{\log y}\,dy + \int_{\Gamma_{w,i}^{(u)}}^{\Gamma_{w,i+1}}h^*(y)\,dy \\
&= \sum_{i=0}^{B}\max(\mathbf{w})\left(\eta(\Gamma_{w,i}^{(\ell)}) - \eta(\Gamma_{w,i})\right) - \log\frac{i}{n}\left(g(\Gamma_{w,i}^{(u)}) - g(\Gamma_{w,i}^{(\ell)})\right) + \eta(\Gamma_{w,i+1}) \\
&\quad - \eta(\Gamma_{w,i}^{(u)})
\end{aligned}$$

In our main submission, we use $\widehat{A}_{\text{BDV,MM},h}$ with $h(y) = \frac{1}{\log(y)}$.

H FURTHER DETAILS ON EXPERIMENTS

Architecture Details

For learning the Pickands dependence function, in all experiments in the manuscript we used 512 width and 1 depth dMNNs. Only the input layer was changed according to the input dimension. In order to force the weights to be positive, we use a weight clipping during training.

For the generative model experiments, we model p_z as a 128 d Gaussian random variable. The generator is a basic multi layer perceptron (MLP) with ReLU activations and batch norm. For all experiments, we use a width 256 and depth 2 MLP for the generator. The output is ensured to be positive through a final ReLU operation.

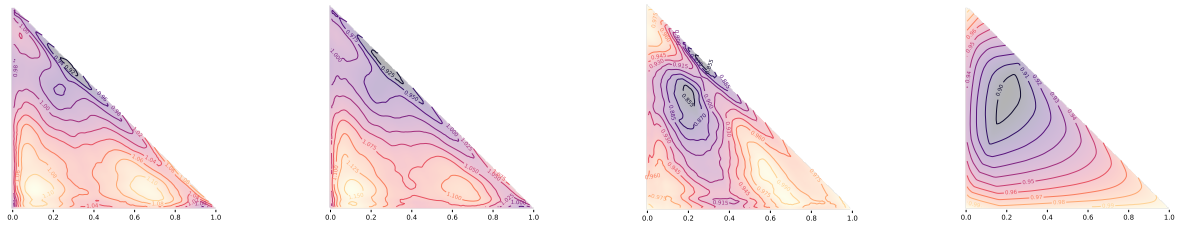
Hyperparameter Tuning

For learning the Pickands dependence experiments, we used the Adam Kingma and Ba [2014] optimizer for optimizing all parameters with learning rate 1×10^{-2} with a decay according to the `ReduceLROnPlateau` decay algorithm with a patience of 100 epochs. Each model was trained for 2000 epochs for the survival experiments and 4000 for the sampling experiments. For the sampling experiments, the generator was trained using Adam with learning rate 1×10^{-3}, $\beta_1 = 0.5$ and $\beta_2 = 0.99$ with exponential decay on the learning rate of 0.99998. Models for the generator were trained for 4000 epochs

Computational Resources

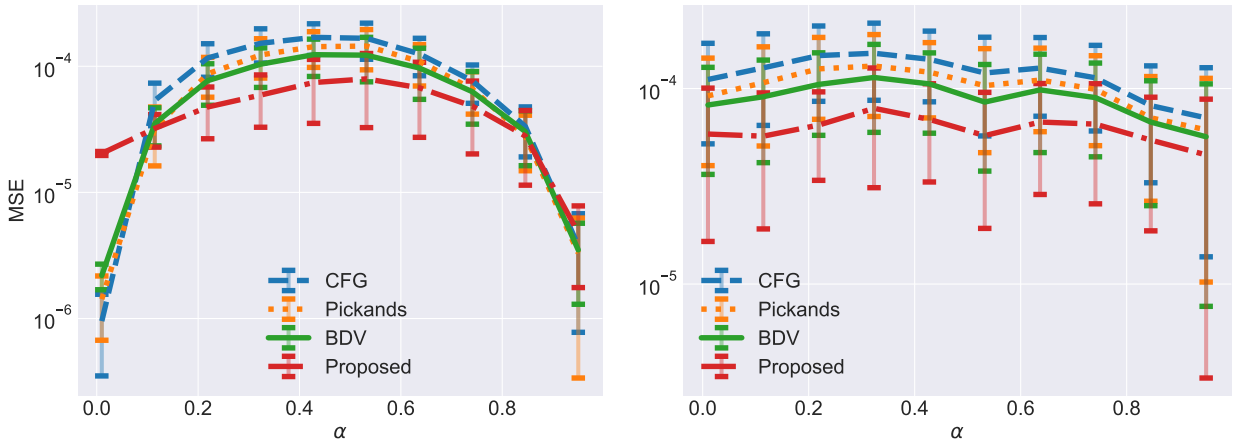
All experiments were run on an Nvidia RTX Titan GPU with an Intel Core i9-7900X CPU @ 3.30GHz and 64 GB of RAM.

I LARGER FIGURES



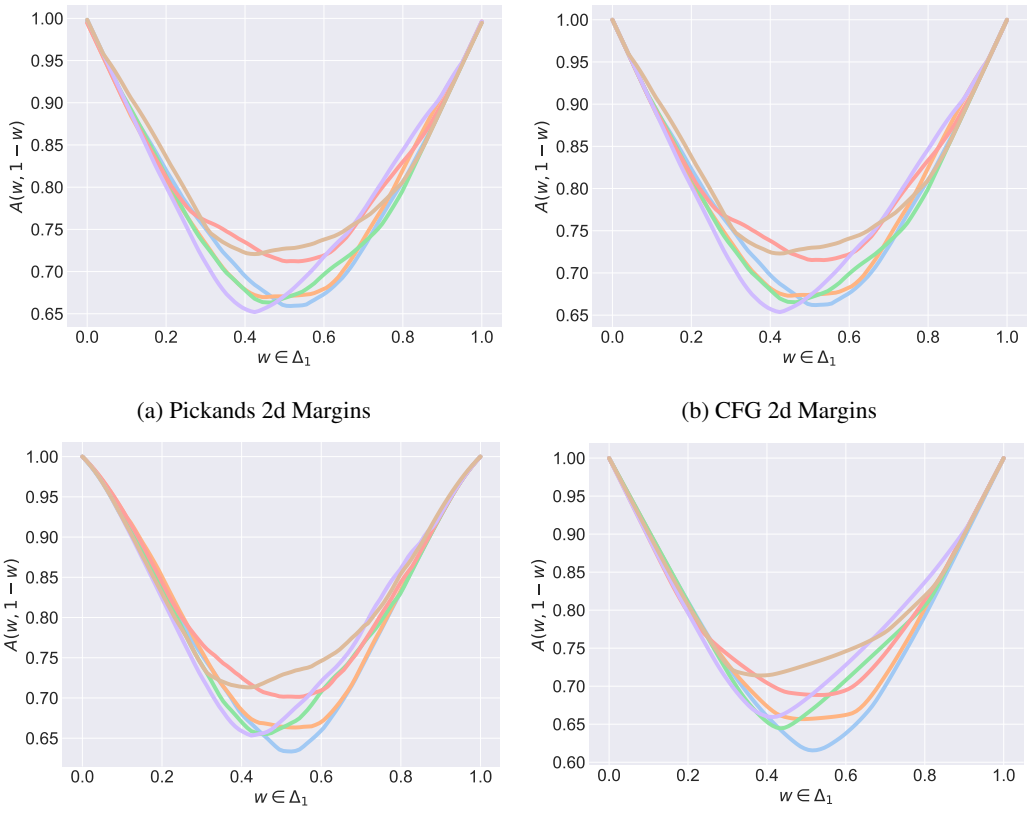
(a) Pickands 3d Margins (b) CFG 3d Margins (c) BDV 3d Margins (d) dMNN 3d Margins

Figure 12: (Larger figures from main text) Qualitative comparison of 3d margins from learned 10d MEV for the commodities dataset. The dMNN is the method that retains margins that are valid Pickands dependence functions as the others are non-convex and outside the required bounds. Contours plotted with solid line.



(a) A_{SL} MSE ($d = 2$) (b) A_{ASL} MSE ($d = 2$)

Figure 13: (Larger figures from main text) (13a, 13b) MSE of survival probabilities for $d = 2$ with 100 samples for A_{SL} (13a) and A_{ASL} (13b). Thresholds are above the 75th percentile.



(a) Pickands 2d Margins

(b) CFG 2d Margins

(c) BDV 2d Margins

(d) dMNN 2d Margins

Figure 14: (Additional figure.) Qualitative comparison of 6 2d margins from learned 4d MEV for the Ozone dataset. The dMNN is the method that retains margins that are valid Pickands dependence functions as the others are non-convex and outside the required bounds.

J ALGORITHMS

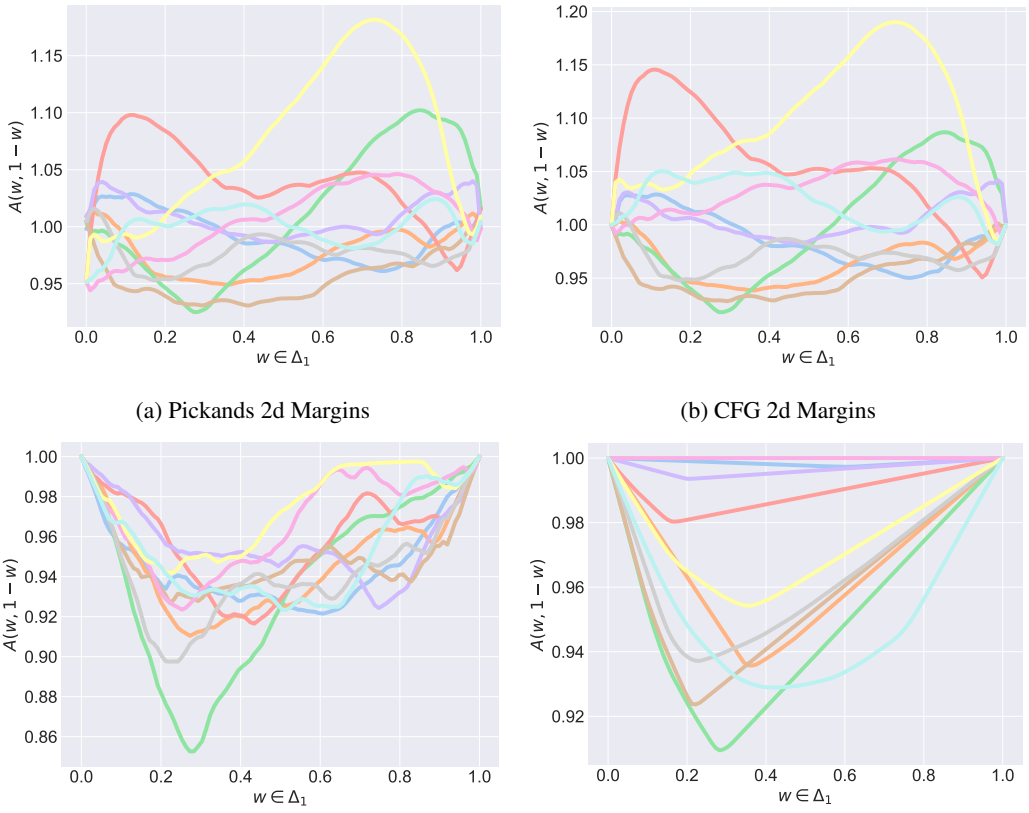
Here we provide algorithms for the estimation and sampling presented in the main content. We recall the transformations on $\bar{M}_k^{(n)}$:

$$\widetilde{M}_k^{(n)} = -\log(F_k(\bar{M}_k^{(n)})),\ \forall k \in \{1, \dots, d\}, \tag{40}$$

$$Z_w = \min_{k=1,\dots,d} \widetilde{M}_k^{(n)}/w_k. \tag{41}$$

Then, we have: $\mathbb{P}\left[Z_w > z\right] = e^{-zA(\mathbf{w})}$. Additionally, recall the definition of the copula in terms of A:

K SAMPLING EXAMPLES



(a) Pickands 2d Margins

(b) CFG 2d Margins

(c) BDV 2d Margins

(d) dMNN 2d Margins

Figure 15: (Additional figure.) Qualitative comparison of 10 2d margins from learned 10d MEV for the commodities. The dMNN is the method that retains margins that are valid Pickands dependence functions as the others are non-convex and outside the required bounds.

Algorithm 1 Fitting the Pickands-dMNN to Data

1: **Input:** $\left\{\left(X_1^{(i)}, \ldots, X_d^{(i)}\right)\right\}_{i=1}^{N}$, $N = B \times n$ samples of i.i.d. random vectors where B is the number of blocks of data and n is the size of each block.

2: Take component-wise maxima over each block: $\left\{\left(M_1^{(n,b)}, \ldots, M_d^{(n,b)}\right)\right\}_{b=1}^{B}$ where $M_k^{(n,b)} = \max_{i=(b-1)n+1,\ldots,bn} X_k^{(i)}$, $(k, b) \in \{1, \ldots, d\} \times \{1, \ldots, B\}$.

3: Fit a GEV to each component-wise maxima $\{M_k^{(n,b)}\}_{b=1}^{B}$, obtain $\{\bar{M}_k^{(n,b)}\}_{b=1}^{B}$, then estimate marginals F_k for each $k \in \{1, \ldots, d\}$.

4: **Initialize** the parameters $\boldsymbol{\theta} \geq 0$ of the dMNN
Repeat:

5: Randomly sample a minibatch of training data $\{\bar{M}_k^{(n,b)}\}_{b \in \text{batch}}$ and uniformly sample $\mathbf{w} \in \Delta_{d-1}$.

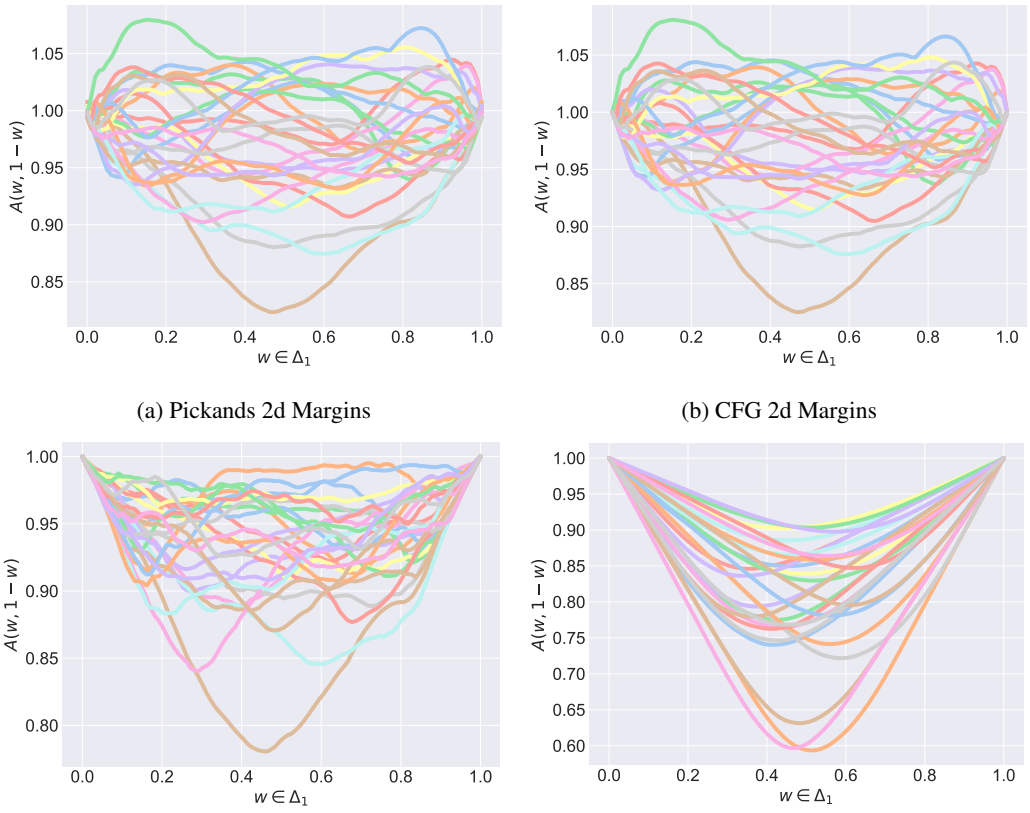
6: Transform samples according to Equations (40) and (41) to obtain transformed samples $\{Z_{w,b}\}_{b \in \text{batch}}$.

7: Compute gradient $\nabla_{\boldsymbol{\theta}} \sum_{b \in \text{batch}} \mathcal{L}\left(Z_{w,b}; \boldsymbol{\theta}\right)$.

8: Update $\boldsymbol{\theta}$ with Adam [Kingma and Ba, 2014]
Until convergence
Output: $A_{\boldsymbol{\theta}}^{\star}(\mathbf{w})$.

REFERENCES

Axel Bücher, Holger Dette, Stanislav Volgushev, et al. New estimators of the pickands dependence function and a test for extreme-value dependence. *The Annals of Statistics*, 39(4):1963–2006, 2011.



(a) Pickands 2d Margins

(b) CFG 2d Margins

(c) BDV 2d Margins

(d) *d*MNN 2d Margins

Figure 16: (Additional figure.) Qualitative comparison of 28 2d margins from learned 418d MEV for the S&P dataset. The *d*MNN is the method that retains margins that are valid Pickands dependence functions as the others are non-convex and outside the required bounds.

Algorithm 2 Estimating survival probabilities with the Pickands dependence function

1: **Input:** $\{\bar{M}_{n,b}^{(k)}\}_{b=1}^{B}$, thresholds: $(\gamma_1, \cdots, \gamma_d)$.

2: Train a model $A(\mathbf{w};\theta)$ with the transformed variables $\{(G_1(\bar{M}_{n,b}^{(1)}, \ldots, G_d(\bar{M}_{n,b}^{(d)}))\}_{b=1}^{B}$ using Algorithm 1 and obtain $A(\mathbf{w};\theta_*)$.

3: Evaluate the Pickands copula:

$$C\left(1 - F_1(\bar{\gamma}_1), \cdots, 1 - F_d(\bar{\gamma}_d)\right),$$

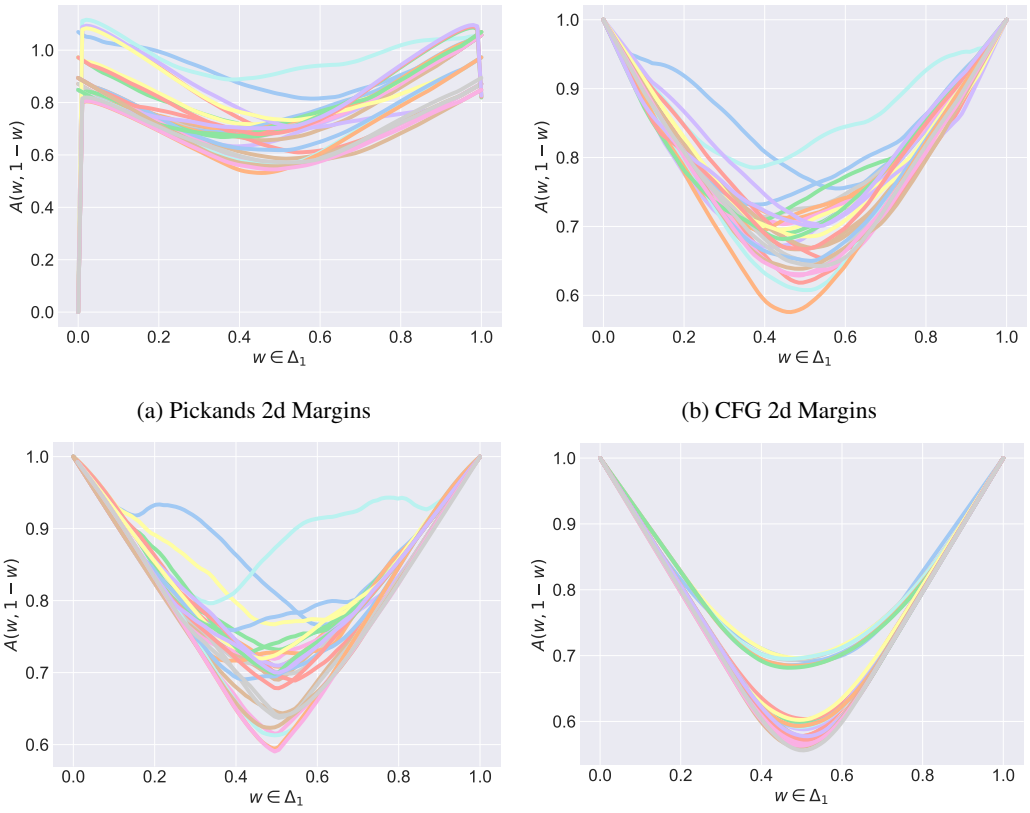
where C is calculated as in Equation (3) with $A = A(\mathbf{w};\theta_*)$.

Philippe Capéraà, A-L Fougères, and Christian Genest. A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3):567–577, 1997.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction (Springer Series in Operations Research and Financial Engineering)*. Springer, 2010.

Laurens De Haan et al. A spectral representation for max-stable processes. *The Annals of Probability*, 12(4):1194–1204, 1984.

Anne-Laure Fougères, Cécile Mercadier, and John P Nolan. Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, 116:109–129, 2013.



(a) Pickands 2d Margins

(b) CFG 2d Margins

(c) BDV 2d Margins

(d) dMNN 2d Margins

Figure 17: (Additional figure.) Qualitative comparison of 28 2d margins from learned 100d MEV for the Crypto dataset. The dMNN is the method that retains margins that are valid Pickands dependence functions as the others are non-convex and outside the required bounds.

Algorithm 3 Training a Generator for a Pickands Copula

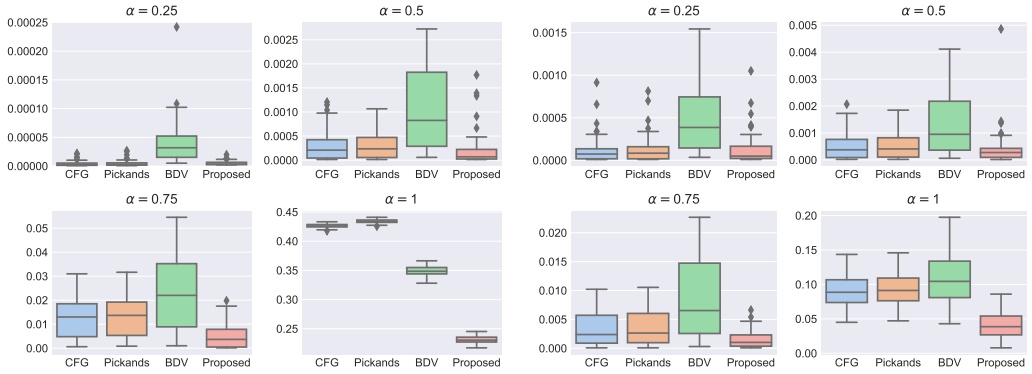
Input: $A(\mathbf{w})$, p_z, tolerance parameter ϵ
Initialize parameters ϕ of generator $G(\cdot;\phi)$
Sample $\{\mathbf{w}^{(j)}\}_{j=1}^{N_{\text{simplex}}}$ samples uniformly over Δ_{d-1}.
while $\sum_{j=1}^{N_{\text{simplex}}} \mathcal{L}(\mathbf{w}^{(j)};\phi) > \epsilon$ **do**
    Sample $\{\mathbf{w}^{(j)}\}_{j=1}^{N_{\text{simplex}}}$ samples uniformly over Δ_{d-1}.
    Sample $\{\mathbf{y}^{(i)}\}_{i=1}^{N_{\text{gen}}}$ where $\mathbf{y}^{(i)} = G(\mathbf{z}^{(i)};\phi), \mathbf{z}^{(i)} \sim p_z$ for $1 \leq i \leq N_{\text{gen}}$.
    Define $\eta(\mathbf{w},\mathbf{y}) = \max\{\mathbf{w} \odot \mathbf{y}\}$ with $\odot$ denoting the point-wise multiplication.
    Compute gradient w.r.t ϕ of $\sum_{j=1}^{N_{\text{simplex}}} \mathcal{L}(\mathbf{w}^{(j)};\phi)$ where:

$$\mathcal{L}(\mathbf{w}^{(j)};\phi) = (A(\mathbf{w}^{(j)}) - \frac{1}{N_{\text{gen}}} \sum_{i=1}^{N_{\text{gen}}} \eta(\mathbf{w}^{(j)}, \mathbf{y}^{(i)}))^2 + \left\| \frac{1}{N_{\text{gen}}} \sum_{i=1}^{N_{\text{gen}}} \mathbf{y}^{(i)} - 1 \right\|_2^2$$

    Update ϕ using Adam Kingma and Ba [2014].
end while
Output: $G(.;\phi_*)$.

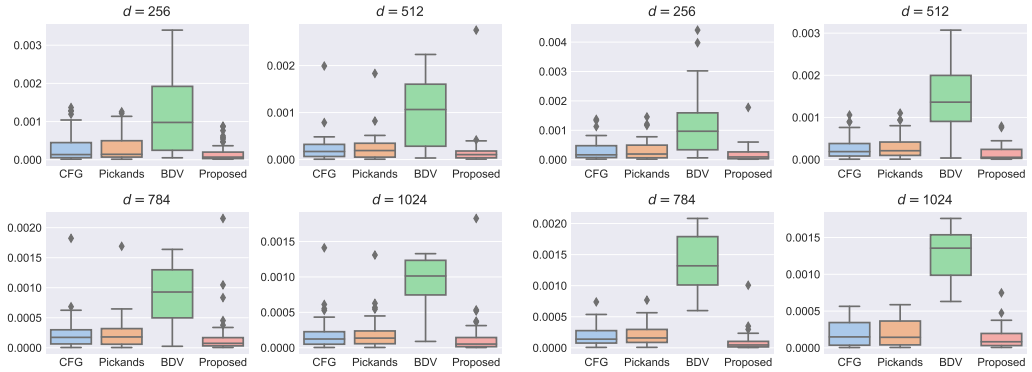
Gordon Gudendorf and Johan Segers. Nonparametric estimation of an extreme-value copula in arbitrary dimensions. *Journal of multivariate analysis*, 102(1):37–47, 2011.

Marius Hofert, Raphaël Huser, and Avinash Prasad. Hierarchical archimax copulas. *Journal of Multivariate Analysis*, 167:



(a) A_{SL} MSE ($d = 256$)

(b) A_{ASL} MSE ($d = 256$)



(c) A_{SL} MSE ($\alpha = 0.5$)

(d) A_{ASL} MSE ($\alpha = 0.5$)

Figure 18: (Larger figures from main text) Comparison of $||\hat{A}(\mathbf{w}) - A(\mathbf{w})||_2^2$ for different estimators $\hat{A}$ for different dependence $\alpha = \{0.25, 0.50, 0.75, 1.0\}$ and $d = 256$ (18a, 18b) and for fixed $\alpha = 0.5$ for $d = \{256, 512, 728, 1024\}$ (18c, 18d) for A_{SL} (18a, 18c) and A_{ASL} (18b, 18d). Results are over 50 runs with 100 training samples for each run.

Algorithm 4 Heuristic for Sampling From a Given Pickands Copula [Hofert et al., 2018, Algorithm 1]

Input: $A(\mathbf{w})$, $N_{\max} > 1 \in \mathbb{N}$
Optimize a generator $G(\cdot; \phi)$ using Algorithm 3.
for $i \in \{1, \ldots, N_{\max}\}$ **do**
  Generate $\mathbf{y}^{(i)}$ where $\mathbf{y}^{(i)} = G(\mathbf{z}^{(i)}; \phi_*), \mathbf{z}^{(i)} \sim p_z$.
  Sample $\{\xi^{(i)}\}_1^{N_{\max}}$ from the Poisson process by sampling $\epsilon_k \sim \text{Exp}(1)$ and $\xi^{(i)} = 1/\sum_{k=1}^{i} \epsilon_k$.
end for
Compute the component-wise maxima as: $M = \max_{1 \leq i \leq N_{\max}} \{\xi^{(i)} \odot \mathbf{y}^{(i)}\}$.
Output: M.

195–211, 2018.

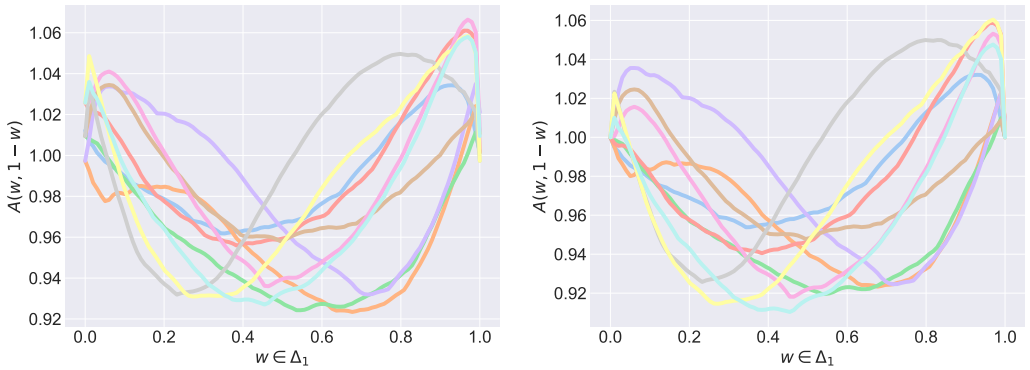
Daniel Hofmann. *Characterization of the D-norm corresponding to a multivariate extreme value distribution*. PhD thesis, Universität Würzburg, 2009.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *http://arxiv.org/abs/1412.6980*, 2014.

J Pickands. Multivariate extreme value distributions, bull. int. statist. 1981.

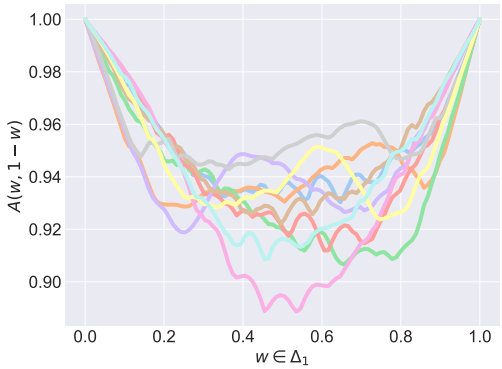
Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 2018.

Jon A Wellner. Empirical processes: Theory and applications. *Notes for a course given at Delft University of Technology*, 2005.

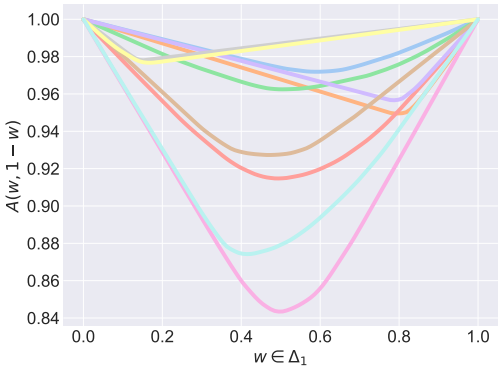


(a) Pickands 2d Margins

(b) CFG 2d Margins

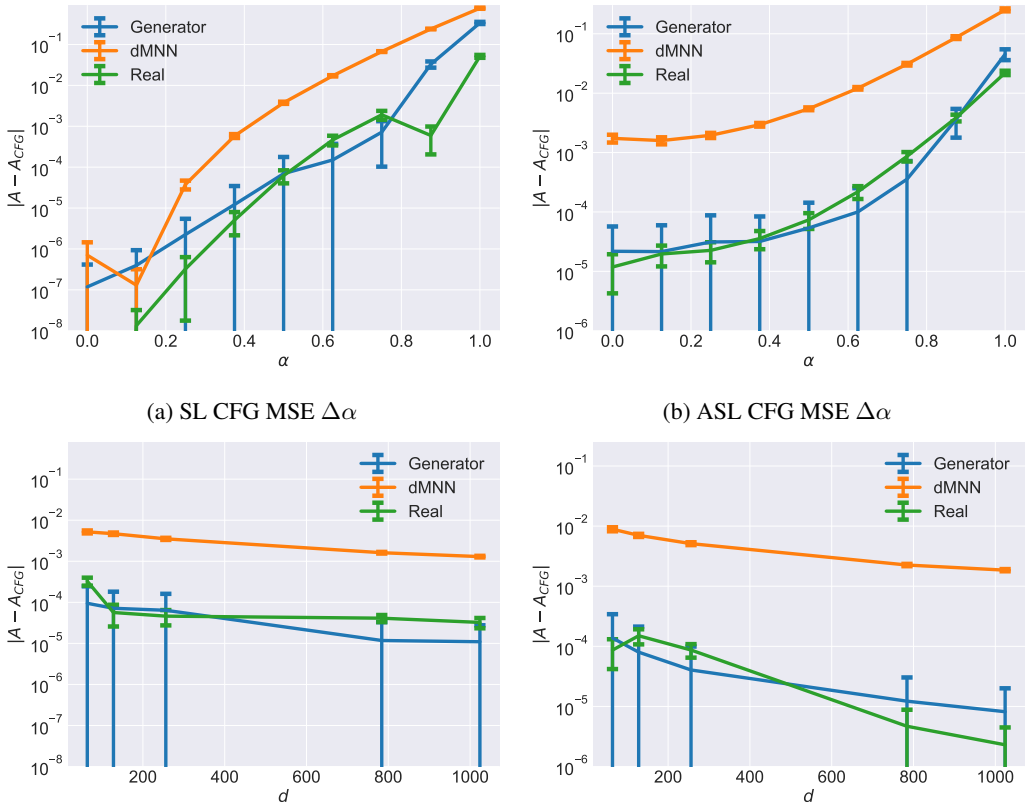


(c) BDV 2d Margins



(d) dMNN 2d Margins

Figure 19: (Larger figures from main text) Qualitative comparison of 10 out of 45 total 2d margins from learned 10d MEV for the California Winds dataset. The dMNN is the only method that retains margins that are valid Pickands dependence functions.



(a) SL CFG MSE $\Delta\alpha$

(b) ASL CFG MSE $\Delta\alpha$

(c) SL CFG MSE Δd

(d) ASL CFG MSE Δd

Figure 20: (Larger figures from main text) MSE of CFG estimate for 1000 samples and 1000 simplex points for $d = 225$ (20a, 20b) at various $\alpha \in (0, 1)$ and $\alpha = 0.5$ (20c, 20d) at $d = \{64, 128, 256, 784, 1024\}$ for A_{SL} (20a) and A_{ASL} (20b) for data sampled from generative model (blue), dMNN (orange), and exact sampled (green). Both models were trained with 1000 data points.