A VAE approach to sample multivariate extremes

Nicolas Lafon, Philippe Naveau, Ronan Fablet

A VAE approach to sample multivariate extremes

Anonymous Authors[1]

Abstract

Rapidly generating accurate extremes from an observational dataset is crucial when seeking to estimate risks associated with the occurrence of future extremes which could be larger than those already observed. Many applications ranging from the occurrence of natural disasters to financial crashes are involved. This paper details a variational auto-encoder (VAE) approach for sampling multivariate extremes. The proposed architecture is based on the extreme value theory (EVT) and more particularly on the notion of multivariate functions with regular variations. Experiments conducted on synthetic datasets as well as on a dataset of discharge measurements along Danube river network illustrate the relevance of our approach.

1. Introduction

Simulating samples from an unknown distribution is a task that various studies have successfully tackled recently in the machine learning (ML) community. This has led to the emergence of generative algorithms, such as generative adversarial networks (GAN) (Goodfellow et al., 2020) or VAEs (Kingma & Welling, 2013; Rezende et al., 2014). ML tasks focus on average behaviors rather than rare events and these methods were not tailored to generate extremes and extrapolate upon the largest value of the training dataset. This is a major drawback when dealing with extremes since accurately sampling extremes provides reference examples for assessing risk in worst-case scenarios. This corresponds to the two-dimensional problem sketched in Figure 1 to which we propose to provide a solution. We seek to consistently generate samples in an extreme region (black square) from observations (blue dots) none of which belong to the extreme region. In this context, the EVT characterizes the probabilistic structure of extreme events and provides a theoretically-sound statistical framework to analyze them. Heavy-tail analysis (Resnick, 2007), in its broadest sense, is a branch of EVT that studies phenomena governed by power laws. Data modeled by heavy-tailed distributions cover a wide range of application fields, e.g., hydrology (Anderson & Meerschaert, 1998; Rietsch et al., 2013), particle motion (Fortin & Clusel, 2015), finance (Bradley & Taqqu, 2003), Internet traffic (Hernandez-Campos et al., 2004), and risk management (Chavez-Demoulin & Roehrl, 2004; Das et al., 2013). Recently, this area of research has gained some interest in the ML community. Some work has shown the potential of bridging the gap between ML and EVT on different aspects, for example dimensionality reduction (Drees & Sabourin, 2021), quantile function approximation (Pasche & Engelke, 2022), outlier detection (Rudd et al., 2017), or classification in tail regions (Jalalzai et al., 2018). Concerning the generation of extremes, ML methods could also integrate EVT tools. To our knowledge, only GANs have been applied to extreme sampling problems. Different GAN approaches have been considered, exploiting various strategies. Feder et al. (2020) relied on heavy-tailed latent variables, as do Huster et al. (2021) who proved that the output random variable of a neural network obtained from a heavy-tailed latent random variable have the same extremal behavior as the latent variable. In Boulaguiem et al. (2022), all marginals were first fitted to heavy-tailed distributions, namely Pareto distributions (see, e.g., Tencaliec et al., 2019). Then they were transformed into a uniform distribution. A multivariate copula (Embrechts, 2009) is learned using a GAN. The inverse quantile function of each previously fitted Pareto distributions was applied to the GAN outputs. This allows to match the behavior of the tail of each margin but

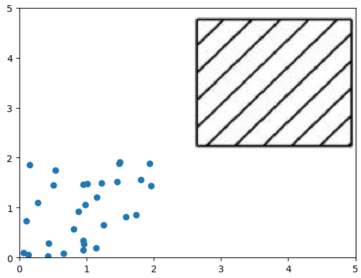

Figure 1. How to sample from observations (blue dots) in extreme regions (black square) to estimate probability of rare events?

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

imposes a specific shape to the whole distribution. Bhatia et al. (2021) proceeded empirically by recursively training GANs from tail samples up to the targeted return level. Based on the observation that a neural network with rectified linear units (ReLU) cannot directly map the interval $[0, 1]$ to the quantile function of a heavy-tailed law, Allouche et al. (2022) proposed a GAN to learn a transformation of this quantile function. They proved that the approximation made by GAN of the quantile functions of marginal laws does converge to the true quantile functions.

In this work, we propose a VAE framework to generate extremes. To our knowledge, this is the first attempt to address whether bridging VAE and EVT can bring fruitful extreme sampling schemes. In addition, recent research suggest that state-of-the-art likelihood-based models, including VAEs, could capture the spread of the true distribution better than GANs (see, e.g., Razavi et al., 2019; Nash et al., 2021).

This paper is organized as follows. We recall the basic principles of VAE and EVT in Section 2. We introduce formally the proposed VAE framework for multivariate extremes in Section 3 and detail the associated training setting in Section 4. Section 5 is dedicated to experiments.

2. Background

2.1. Sampling with VAE

To generate a sample from a random variable $\mathbf{X}$, VAE proposes a sampling strategy based on two steps:

- A sample $\mathbf{z}$ is drawn from a latent vector -or prior- $\mathbf{Z}$ with prior distribution $p_\alpha(\mathbf{z})$ parametrized by α;
- The desired sample is obtained by sampling from the conditional pdf $p(\mathbf{x}|\mathbf{z})$.

Since $p(\mathbf{x}|\mathbf{z})$ is in general not known, one uses an approximation $p_\theta(\mathbf{x}|\mathbf{z})$ parametrized by θ, referred to as the likelihood. The purpose is then to find the parametrization which enables to generate the most realistic samples of $\mathbf{X}$. To do so, VAE framework introduces a target distribution $q_\phi(\mathbf{z}|\mathbf{x})$ parametrized by ϕ to approximate the true posterior distribution. The training phase then comes to maximize the evidence lower bound (ELBO) with respect to the set of parameters (α, ϕ, θ). Formally, given $(\mathbf{x}^{(i)})_{i=1}^N$ N independent samples of $\mathbf{X}$, we have

$$-\log(p(\mathbf{x}^{(i)}) \geq L(\mathbf{x}^{(i)}, \alpha, \theta, \phi),$$

with L the ELBO cost given by

$$L(\mathbf{x}^{(i)}, \alpha, \theta, \phi) = -D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\alpha(\mathbf{z})\right) + E_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\right]. \quad (1)$$

The ELBO cost on the whole dataset is obtained by averaging eq. (1) over the N samples of $\mathbf{X}$. To infer the set of parameters (α, ϕ, θ) by neural network functions of the data, Kingma & Welling (2013) and Rezende et al. (2014) derived a specific training scheme for ELBO optimization. The authors allowed the cost function defined by eq. (1) to be approximated by an unbiased Monte Carlo estimator differentiable with respect to both θ and ϕ. This Monte Carlo estimator is given for a data point by

$$\hat{L}(\mathbf{x}^{(i)}, \alpha, \theta, \phi) = -D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\alpha(\mathbf{z})\right) + \frac{1}{L}\sum_{l=1}^{L} p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}), \quad (2)$$

where $\mathbf{z}^{(i,l)}$ are samples from the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$. To make this expression differentiable, a reparametrization trick is used. In an explicit reparametrization setting, a function g_ϕ has to be find, such that

$$q_\phi(\mathbf{z}|\mathbf{x}) = g_\phi(\mathbf{x}, \epsilon), \quad (3)$$

with ϵ a chosen random variable, and g_ϕ differentiable with respect to ϕ. When explicit reparametrization is not feasible, we may exploit implicit reparametrization gradients (see Figurnov et al., 2018). Details about implicit reparametrization are to be found in Appendix D.

2.2. Univariate extremes

A crucial notion regarding extreme values is the so-called regular variation property. A random variable X is said to be regularly varying with tail index $\alpha > 0$, if

$$\lim_{t\to+\infty} P(X > tx \mid X > t) = x^{-\alpha}. \quad (4)$$

Regularly varying functions belongs to the larger class of generalized Pareto (GP) (Pickands III, 1975) distributions, where the survival function is defined by

$$\bar{H}_{\sigma,\xi}(y) = \left(1 + \xi\frac{y}{\sigma}\right)_+^{-1/\xi}, \quad (5)$$

where $a_+ = 0$ if $a < 0$. The scalar ξ is called the shape parameter and, for $\xi > 0$ is related to the tail index by $\xi = \frac{1}{\alpha}$. The case $\xi > 0$ corresponds to heavy-tailed distribution.

This parametric modeling is motivated by its stability with respect to thresholding, namely if a random variable Y has a GP survival distribution $\bar{H}_{\sigma,\xi}$, then $P(Y > y + v \mid Y > v) = \bar{H}_{\sigma_v,\xi}(y)$ where σ_v depends on the value of the threshold.

A simple yet efficient way to sample from a GP distribution with parameters ξ and σ is to multiply an inverse gamma distributed random variable with shape $\frac{1}{\xi}$ and rate σ by an unit and independent exponential one. This multiplicative feature is essential for understanding the pivotal role of inverse-Gamma random variable in our sampling scheme in Section 3.1.

2.3. Multivariate extremes

A key concept to represent multivariate heavy-tailed data is the following multivariate regular property (see, e.g. Resnick, 2007, for details) which extends the notion developed in eq. (4). Let $\mathbf{X}$ be a random vector in $(\mathbb{R}^+)^d$. To define multivariate regular variations, we need to decompose $\mathbf{X}$ into a radial component $R = X_1 + \cdots + X_d = ||\mathbf{X}||$ and an angular component of the d-dimensional simplex $\mathbf{\Theta} = \frac{\mathbf{X}}{||X||}$. $\mathbf{X}$ has multivariate regular variation if the two following properties are fulfilled:

- The radius R is regularly varying as defined in eq. (4);
- There exists a probability measure $\mathbf{S}$ defined on the d-dimension simplex such that $(R, \mathbf{\Theta})$ verify:

$$P\left(\mathbf{\Theta} \in \bullet \;\middle|\; R > r\right) \xrightarrow{w} \mathbf{S}, \tag{6}$$

  with $\xrightarrow{w}$ standing for weak convergence metric. $\mathbf{S}$ is called limit angular measure.

The last characterization (eq. (6)) indicates that given that the radius is above a sufficiently high threshold, the conditional distribution of the radius and the angle can be considered independent. This is of key interest to address tail events of the kind of $\{\mathbf{X} \in C\}$ when $u = \inf\{||\mathbf{x}||, \mathbf{x} \in C\}$ is large.

3. Proposed VAE architecture

We propose the following main three-step scheme to generate a sample $\mathbf{x}^{(i)}$ of a multivariate regularly-varying random vector:

- Using a VAE, a radius $r^{(i)}$ is drawn from an univariate heavy-tailed distribution R (see Section 3.2);
- Conditionally on the drawn radius $r^{(i)}$, we sample $\mathbf{\Theta}^{(i)}$ an element of the d-dimensional simplex from the conditional distribution $\mathbf{\Theta}|[R = r^{(i)}]$ while forcing the independence between radius R and angle $\mathbf{\Theta}$ for larger value of the radius. We use a conditional VAE for this purpose (see Section 3.3);
- We multiply compoment-wise the angle vector by the radius to obtain the desired sample, i.e. $\mathbf{x}^{(i)} = r^{(i)}\mathbf{\Theta}^{(i)}$.

The polar decomposition we used has important advantages. First, one can generate elements of the simplex with a given radius and study the dependence between variables at a given extreme level. Second, the polar decomposition also offers a great flexibility in modeling the dependence between variables, including for the limiting angular distribution, which is not the case for a GAN ReLU with heavy-tailed latent variables for instance (see Theorem E.2). The rest of this section details the architecture of the VAEs chosen to sample the heavy-tailed radius and the conditional angle.

3.1. Idealized multiplicative framework for sampling heavy-tailed radii

We model R through a latent variable Z_{rad}. In many applications, the joint pdf $f(z_{rad}, r)$ is assumed to be Gaussian and this leads to a L2 term of the type $\mathbf{E}_{Z_{rad}\sim q}||z_{rad} - r||^2_\Sigma$ in $L(q)$ (eq. (1)). Still, this limits the resulting ELBO maximization approach to light-tailed distributions. Therefore, it does not seem relevant for phenomena that are based on a multiplicative structure rather than an additive one, the later being ideal for Gaussian assumptions. We then move away from the L2-norm and the Gaussian hypothesis and, instead, we focus on heavy-tailed distributions introduced in Section 2.2. In this context, additional assumptions are needed.

Condition 3.1. Z_{rad} follows the inverse-gamma pdf defined by

$$f_{\mathbf{Inv\Gamma}}(z_{rad}; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z_{rad}^{-\alpha-1} \exp\left(-\beta/z_{rad}\right), \tag{7}$$

with α and β two strictly positive constants.

Condition 3.2. R is linked to Z_{rad} throughout a multiplicative model with a positive random coefficient A, i.e.

$$R \overset{\mathrm{d}}{=} A \times Z_{rad}, \tag{8}$$

where $\overset{\mathrm{d}}{=}$ corresponds to a equality in distribution and the random variable A is absolutely continuous and independent of Z. We also assume that $0 < \mathbf{E}A^{\alpha+\epsilon} < \infty$ for some positive ϵ.

Under these conditions, a direct application of Breiman's lemma (Breiman, 1965) implies the following statement.

Proposition 3.3. *If conditions 3.1 and 3.2 are fulfilled , R is heavy-tailed with tail index α.*

Notice that if A follows an exponential distribution then R follows a GP distribution (see eq. (5) with $\xi = \frac{1}{\alpha}$), which models exceedances above a large threshold as seen in 2.2. This idealized framework is the starting point for our thinking on heavy-tailed radius generation for multivariate data with VAEs. The theoretical properties that ensure that the tail index of R has the required value provides a structure towards which we have endeavored to strive as R tends to infinity.

3.2. Sampling from heavy-tailed radius distributions

To tailor the VAE framework introduced in 2.1 to heavy-tailed random variables, we set the prior Z_{rad} as an inverse gamma distribution with parameters α and 1, namely:

$$p_\alpha(z_{rad}) = f_{\mathbf{Inv\Gamma}}(z_{rad}\,;\,\alpha, 1). \tag{9}$$

Notice that as if X follows an inverse gamma distribution with parameters α and β, then for each $c > 0$, cX is an inverse gamma with parameters α and $c\beta$. Consequently, and without loss of generality, we set parameter β equal to 1.

From conditions 3.1 and 3.2, the conditional R and Z_{rad}, given respectively that Z_{rad} and $R = r$, can be expressed as

$$R|[Z_{rad} = z_{rad}] \overset{\mathrm{d}}{=} Az_{rad}, \tag{10}$$

$$Z_{rad}|[R = r] \overset{\mathrm{d}}{=} \frac{r}{A}. \tag{11}$$

To perform inference within a VAE framweork, we need to choose a parametric form for the likelihood. Recall from Condition 3.2 that A needs to have a tail lighter than Z_{rad} to satisfy the moment condition $\mathbf{E}A^{\alpha+\epsilon}$ for some positive ϵ. With this in mind, choosing the approximate likelihood in a light-tailed distribution family seems relevant. The target could be chosen either bounded if R is bounded away from zero or light to heavy-tail if R seems to have non-null probability on each open set containing 0. Moreover, as R is a positive random variable, we must ensure that negative values for either target and likelihood cannot occur. Overall, We choose the following parametrizations:

$$\begin{aligned} p_\theta(r|z_{rad}) &= f_{\mathbf{\Gamma}}\left(r\ ;\ \alpha_\theta(z_{rad}), \beta_\theta(z_{rad})\right) \\ q_\phi(z_{rad}|r) &= f_{\mathbf{InvGamma}}\left(z_{rad}\ ;\ \alpha_\phi(r), \beta_\phi(r)\right), \end{aligned} \tag{12}$$

with $f_{\mathbf{\Gamma}}$ (resp. $f_{\mathbf{InvGamma}}$) the pdf of a Gamma (resp. inverse Gamma) distribution. In this context, α_θ, β_θ, α_ϕ, β_ϕ are neural networks functions with parameters θ and ϕ. By introducing the inverse-gamma parameterizations for the prior p_α and the target q_ϕ in the ELBO cost of VAE (eq. (2)), we obtain the following proposition.

Proposition 3.4. *Given expression (7) and (12) for prior and target distributions, the KL divergence in eq. (2) is given by:*

$$\begin{aligned} &D_{KL}\left(q_\phi(z_{rad}|r)||p_\alpha(z_{rad})\right) = \\ &(\alpha_\phi(r) - \alpha)\psi(\alpha) - \log\frac{\Gamma(\alpha_\phi(r))}{\Gamma(\alpha)} \\ &+\alpha\log\beta_\phi(r) + \alpha_\phi\frac{1-\beta_\phi(r)}{\beta_\phi(r)}, \end{aligned} \tag{13}$$

where Γ and ψ stands respectively for the gamma and digamma functions.

This proposition means that we can optimize our model with respect to α, and learn the tail index of the distribution directly from data.

Notice that in the approximated posterior, as the parameters of the distribution are learned functions of the samples from the prior, the assumption of independence between A and Z_{rad} in Condition 3.2 is not necessarily verified. Therefore, neither does Breiman's lemma and this could affect the tail index preventing it from having the desired value. We have to constraint the function α_θ and β_θ of the target to ensure that the generated R has tail index α. To do so we consider the following proposition.

Proposition 3.5. *If the function $\alpha_\theta(.)$ is a strictly positive constant and the function $\beta_\theta(.)$ satisfies*

$$\lim_{z_{rad}\to+\infty}\beta_\theta(z_{rad}) \propto \frac{1}{z_{rad}}, \tag{14}$$

then R has a tail index equal to α.

In practice, in the implementation of the method, we force to satisfy $\beta_\theta(.)$ eq. (14) but leave $\alpha_\theta(.)$ more flexible, constraining it only to have a strictly positive finite limit at infinity. Experimentally, this seems to be enough to obtain the desired behavior for the extremes.

3.3. Sampling on the multivariate simplex

Once we choose the framework for the radial part of our dataset, sampling from the angular component part still remains. After generating R according to the framework developed in section 3.2, we design a conditional VAE (see, e.g., Zhao et al., 2017) to sample on the multivariate simplex conditionally on a previously sampled radius, namely conditional distribution $\mathbf{\Theta}|R$. This angular VAE has a latent variable $\mathbf{Z}_{ang}$ with a multivariate normal prior. The target is also parameterized by multivariate normal distributions, with mean and standard deviation function of the hidden variable and the observation data. The likelihood could be parametrized by a projection of a normal distribution on the multivariate simplex denoted $\mathbf{\Pi}$, or directly by a Dirichlet distribution. We summarize hereafter the chosen parametrisation where the likelihood is the projection of multivariate normal:

$$\begin{aligned} \mathbf{Z_{ang}} &\sim \mathcal{N}(0, I_{N_z}), \\ p_\nu(\mathbf{\Theta}|\mathbf{z}_{ang}, r) &\sim \mathbf{\Pi}\left(\mathcal{N}(\mu_\nu(\mathbf{z}_{ang}, r), \mathbf{\Sigma}_\nu(\mathbf{z}_{ang}, r))\right), \\ p_\omega(\mathbf{Z}_{ang}|\mathbf{s}, r) &\sim \mathcal{N}\left(\mu_\omega(\mathbf{s}, r), \mathbf{\Sigma}_\omega(\mathbf{s}, r)\right), \end{aligned}$$

where N_z is the dimension of the latent space, μ_ν, $\mathbf{\Sigma}_\nu$, μ_ω and $\mathbf{\Sigma}_\omega$ are neural network functions with parameters ν and ω. The dependency on R for the target and the likelihood has been made explicit to turn the framework conditional. Notice that to enforce the independence between the radius and the sphere when $r \to +\infty$, we make sure that the functions μ_ν and $\mathbf{\Sigma}_\nu$ satisfy the following necessary condition:

Condition 3.6. μ_ν and $\mathbf{\Sigma}_\nu$ are such that there exist two z-

varying functions μ_∞ and $\mathbf{\Sigma}_\infty$ which verify for each $\mathbf{z}_{ang}$

$$\lim_{r\to+\infty} \mu_\nu(\mathbf{z}_{ang}, r) = \mu_\infty(\mathbf{z}_{ang})$$
$$\lim_{r\to+\infty} \mathbf{\Sigma}_\nu(\mathbf{z}_{ang}, r) = \mathbf{\Sigma}_\infty(\mathbf{z}_{ang})$$

4. Learning framework

4.1. Training settings

In our numerical experiments, we consider the following parameterization of the neural architectures. We set the number of hidden layers of the neural network function of the probabilistic encoders and decoders to 2 for radii generation and to 3 for the sampling on the multivariate simplex. Concerning the latter, projected multivariate normal, and Dirichlet distribution have both been considered to parameterize the approximate likelihood. The optimized cost is the ELBO cost, which is given in its general formulation by the eq. (2). The metric chosen to infer the best parametrization of our approach is the ELBO cost on the validation set. The training is limited to 5000 epochs, and the learning rate set to 10^{-4} for radius generation training and 10^{-5} for angular generation training. Concerning radius generation, let us note that, depending on the experiments, the parameter α of the prior can either be supposed known or unknown. When unknown, α can be directly optimized by gradient descent as a model parameter using eq. (13). We used the Adam optimizer (Kingma & Ba, 2014). From a code perspective, we made extensive use of the Tensorflow and Tensorflow-Probability libraries. The whole code is freely available[1].

4.2. Performance assessment

For radii, log quantile quantile plots (see Resnick, 2007, for detailed examples), abbreviated as log-QQ plots, are graphical methods we use to informally assess the goodness-of-fit of our model to data. This method consists in plotting the empirical quantiles of a sample generated by our approach vs the empirical quantiles of the experimental data. If the fit is good, the plot should be roughly linear. We use the VAE cost (eq. (2)) on a given dataset as a numerical indicator to compare the radius distribution obtained with our VAE approach to a vanilla VAE not tailored for extremes. Another criterion that we apply is an estimator of the KL divergence, as well as one of its variants introduced by Naveau et al. (2014). This variant gives an estimator of the KL divergence upon a given threshold (see Appendix C.1).
Concerning the whole generated samples, we investigate several other criteria. We computed the Wasserstein distance between large samples generated by our model and true samples. If we select a threshold u, we can compute the Wasserstein distance above this threshold by restricting the samples to the points which have a radius greater than u. In this context, we consider a rescaled version of the Wasserstein distance upon a threshold divided by the square of this threshold (see Appendix C.2). To compute the Wasserstein distances, we use pre-implemented functions from the Python Optimal Transport package[2] (see Flamary et al., 2021).
We have seen that for a multivariate regularly varying random vector, the radius and the angle can be considered independent in the limit of an infinite radius (see eq. (6)). In practice, one can consider the radius and the angle independent by choosing a sufficiently large radius. Wan & Davis (2019) have established a criterion to detect whether the respective distributions of the radius and the angle can be considered as independent, and thus to choose the corresponding limiting radius. This allows us to compare the limiting radii between the true data and the generated data. To do so, the authors propose a testing framework to calculate a p-value that follows a uniform distribution if the distributions of the radius and the angle are independent, and that is close to 0 otherwise (see Appendix C.3).

5. Experiments

We conduct experiments on multivariate datasets, both synthetic and real. The synthetic dataset have a heavy-tailed radius distribution and the angular distribution on the multivariate simplex is a Dirichlet distribution which parameters varies according to the radius. The real dataset corresponds to a monitoring of Danube river network discharges.

5.1. Notations and benchmarked approaches

We refer to our generative approach as ExtVAE if we assume that the tail index α is known, and as UExtVAE if the tail index is learned by minimizing the analytical expression in eq. (13). If we restrict ourselves to the radii generated by ExtVAE and UExtVAE via the procedure described in Section 3.2, we denote respectively ExtVAE$_r$ and UExtVAE$_r$. We compare our approach with standard VAE (see Cemgil et al., 2020), i.e. with normal distribution for prior, target and likelihood, indicated by the acronym StdVAE. We also compare our approach with, ParetoGAN which is the GAN for generating extremes proposed by Huster et al. (2021). The ParetoGAN is a Wasserstein GAN (see Arjovsky et al., 2017) with Pareto latent variables. Given the difficulty of training a GAN, as well as the number of factors that can influence the results it produces, we empirically adjusted the ParetoGAN architecture to provide a sensible GAN baseline in our experiments. Though our parameterization may not

[1] Implementation available at https://anonymous.4open.science/r/SubmissionICML-0263

[2] see https://pythonot.github.io/quickstart.html

Table 1. Mean VAE cost on radius R_1 (see eq. (2)) training, validation and test dataset. These are abbreviated in Train, Val and Test loss.

Approach	Train loss	Val loss	Test loss
StdVAE	1.21	4.81	$+\infty$
ExtVAE$_r$	0.88	1.10	1.12
UExtVAE$_r$	0.95	1.12	1.15

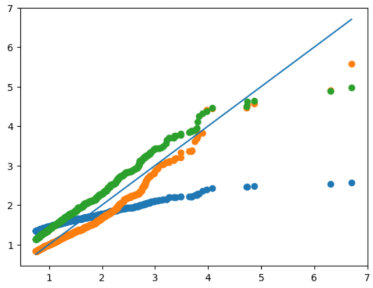

Figure 2. Log-QQ plot between the upper decile of 10000 radii samples of StdVAE (blue dots), ExtVAE$_r$ (orange dots), UExtVAE$_r$ (green dots) and the upper decile of the test dataset of R_1. The dots should lie close to the blue line

be optimal, our interest goes beyond a simple quantitative intercomparison in exploring and understanding the differences between the proposed VAE approach and GANs in their ability to represent and sample extremes.

5.2. Synthetic dataset

We first consider a synthetic dataset with a 5-dimensional heavy-tailed random variable with a tail index $\alpha = 1.5$. We detail the simulation setting in Appendix A.1. The training dataset consists of 250 samples, compared to 750 for the validation dataset and 10000 for the test dataset.

In Table 1 and Figure 2, we study the ability of the benchmarked VAE schemes to sample heavy-tailed radius distribution. The log-QQ plots given in Figure 2 illustrate further that ExtVAE$_r$ and UExtVAE$_r$ schemes relevantly reproduce the linear tail pattern of the radius distribution while this is not the case for StdVAE. Figure 3 evaluates, for the compared methods, the evolution of the KL divergence between the true distribution and the simulated ones above a varying quantile u (eq. (16)). Again, the StdVAE poorly matches the target distribution with a clear increasing trend for quantiles greater than $u = 0.3$. Conversely, the KL divergence is much smaller and much more stable for ExtVAE$_r$ and UExtVAE$_r$ schemes, especially for large quantile values. Interestingly, for the different criteria, the results obtained with UExtVAE$_r$ are very close or even indistinguishable from those obtained with ExtVAE$_r$. This suggests that the estimation of the tail index is accurate. In order to better assess the robustness of this estimation, we report the evolution of the tail index of UExtVAE$_r$ as a function of the training epochs for randomly chosen initial values (Figure 4). Given the expected uncertainty in estimating the tail index (see Appendix B), UExtVAE$_r$ estimates (Figure 4) are globally consistent. We report very good estimation patterns since the reported curves tend to get closer to the true value as the number of epochs increase, although it might show some bias when initial value is far from the true tail index value.

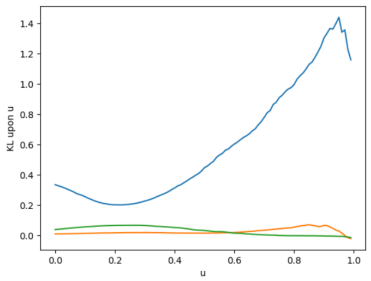

Figure 3. KL divergence between the radius distribution of the benchmarked VAE models and the target heavy-tailed distribution: we display the KL divergence (see eq. (16)) above quantile u for u varying from 0 to 1 for StdVAE (blue curve), ExtVAE$_r$ (orange curve) and UExtVAE$_r$. Numerically speaking, we sampled 10000 from each distribution and u is taken as the quantile of the sampled reference dataset.

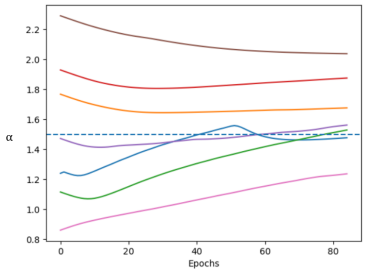

Figure 4. Evolution of the tail index α of UExtVAE$_r$ during the training procedure: we report the value of the tail index as a function of the training epochs for training runs from different initial values. The initial values of α are sampled uniformly between 0.8 to 2.25. The true value of α is 1.5 (dashed horizontal line).

We now focus on the five-dimensional heavy-tailed case-study. The best parametrization for the likelihood of the conditional VAE is a Dirichlet parametrization. An important advantage of our approach is the ability to generate samples on the simplex for a given radius as detailed in Section 3.3. If we send the radius to infinity, we can estimate the limit angular measure (eq. (6)). Figure 5 displays this limit angular measure projected onto the last two components of the simplex for the true limit angular distribution, our ExtVAE approach and the ParetoGAN. For the latter, we approximate the limit angular distribution by the empirical distribution above a very high threshold. The ExtVAE

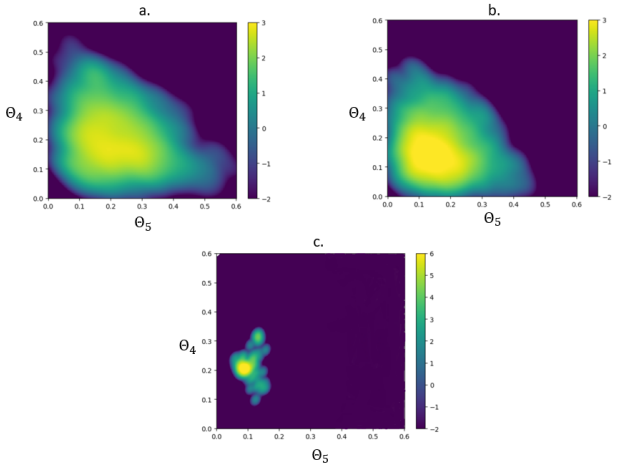

Figure 5. Log probability of the estimated limit angular measure obtained with a. ExtVAE, b. true distribution, c. ParetoGAN, projected on axes 4 and 5 (named θ_4 and θ_5). The estimation is based on 10000 samples of each at a high value of radius, typically above 10, which corresponds to the upper percentile of R_1 distribution.

shows a good agreement with the true distribution, though not as sharp. By contrast, the distribution sampled by the ParetoGAN tends to reduce to a single mode. The spatial direction of ParetoGAN extremes could therefore be erroneously interpreted as deterministic. This confirms the result of Theorem E.2.

Beyond the limit angular distribution, we asses the sampling performance of the benchmarked schemes through an approximation of the Wasserstein distance (eq. (17)) between 10000 generated items and the test set. The ExtVAE performs slightly better than the ParetoGAN (5.37 vs. 6.80). This is highlighted for high quantiles in Figure 6. We plot the Wasserstein distance upon a radius threshold, dividing by the square of the threshold, for ExtVAE and ParetoGAN. We focus on radius thresholds above 2, which corresponds to the highest decile. The ExtVAE performs again better than the ParetoGAN, especially for radius values between 2 and 4, corresponding roughly to quantiles between 0.90 and 0.95. We may recall that the ParetoGAN relies on the minimization of a Wasserstein metric, whereas the ExtVAE relies on a likelihood criterion. Therefore, we regard these results as an illustration of the better generalization performance of the ExtVAE, especially for the extremes.

At last, we estimate the threshold at which the respective distribution of radius and angle can be considered as independent following the criterion proposed by Wan & Davis (2019). Although, by construction, there is no radius value from which there is a true independence, the estimator gives a radius above which one can approximately consider that the limit distribution is reached. We compare in Figure 7 the p-values of sampled data as a function of the chosen

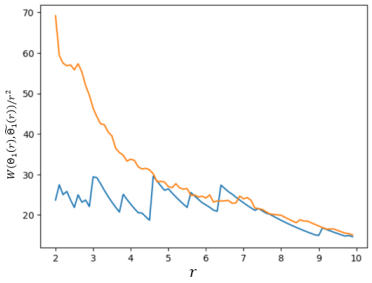

Figure 6. Wasserstein distance upon radius threshold r divided by the square of r calculated between 10000 samples drawn from generative approaches and test set. In orange, the generative method is the ParetoGAN and in blue it is our. The considered thresholds are above 2, which is roughly the upper decile of the radii distribution.

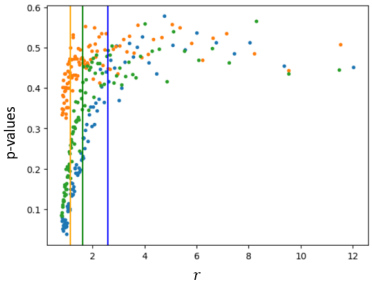

Figure 7. P-values at different radius threshold, (see Wan & Davis, 2019) for the test dataset (green points), 10000 samples of the ExtVAE (orange points), and 10000 samples of the ParetoGAN (blue points). The vertical bars correspond to the threshold below which the p-values are less than 0.45. Above this threshold, the radius and the angle can be roughly considered as independent.

threshold for the true distribution, the ExtVAE and the ParetoGAN. The ExtVAE slightly underestimates the radius of the threshold compared to the true data (1.3 vs. 1.6), while the ParetoGAN leads to a large overestimatation (2.7 vs. 1.6). This illustrates further that the ExtVAE better captures the statistical features of high quantiles than ParetoGAN does. We regard the polar decomposition considered in the ExtVAE as the key property of the ExtVAE to better render the asymptotic distributions between the radius and the angle.

5.3. Danube river discharge case-study

Our second experiment addresses a real heavy-tailed multivariate dataset. We consider the daily time series of river flow measurements over 50 years at five stations of the Danube river network (see Appendix A.2 for further details). River flow data are widely acknowledged to depict heavy-tailed distributions (Katz et al., 2002). In reference to the numbering of the stations (see Figure 9), we note the random variables associated with the considered stations X_{23}, X_{24}, X_{25}, X_{26}, X_{27}. We consider a training dataset

of 730 daily measurements, the remaining 17486 measurements form the test dataset. We focus on the question raised in introduction (see Figure 1): can we extrapolate and generate consistent samples in extreme areas not observed during the training phase? We focus on extreme areas of the form $\bigcap_{i=23}^{27} X_i > u_i$ with u_i predefined thresholds. This corresponds to flows exceeding predefined thresholds at several stations. The estimation of the probabilities of occurrence of such events is key to the assessment of major flooding risks along the river.

Our experiments proceed as follows. We train generative schemes on the training set as detailed in Section 3. For this case-study, the best parametrization for the likelihood of the angular part of the UExtVAE is a projection of a multivariate normal distribution. As evaluation framework, we generate for each trained model a number of samples of the size of the test dataset, and we compare the proportion of samples that satisfy a given extreme event to those of the test dataset. We consider extreme events corresponding to quantile values of 0.9 and 0.99. Table 2 synthesizes this analysis for the StdVAE and the UExtVAE[3]. As illustrated, the training dataset does not include extreme events for the 0.99 quantiles. Interestingly, the UExtVAE samples such extreme events with the same order of magnitude of occurrence as in the test dataset. For instance, the proportion of samples that satisfy $A_{26}^{(0.99)}$ and $A_{27}^{(0.99)}$ is consistent with that observed in the test dataset, respectively 0.2% and 0.18% against 0.4% and 0.25%. By contrast, the StdVAE cannot generalize beyond the training dataset. Though not as good as the UExtVAE for the 0.9 quantiles, the StdVAE truely generates events above these thresholds. However, the StdVAE does not generate any element in $A_{26}^{(0.99)}$ and $A_{27}^{(0.99)}$.

Note that the tail index of the radius of the discharge data set is not known a priori. Asadi et al. (2015) reports an estimate of 3.5 ± 0.5 considering only the summer months. In our case, the tail index of the trained UExtVAE is of 4.5. It is slightly higher than the value found by Asadi et al. (2015), which means a less heavy-tail distribution. Indeed, half of the annual maxima occurs in June, July or August, typically due to heavy summer rain events. Thus, we expect the summer months to lead to a heavier tail than the all-season dataset.

Conclusion

This study bridges VAE and EVT to address the generative modeling of multivariate extremes. Following the concept of multivariate regular variations, we propose a polar decomposition and combine a generative model of heavy-tailed radii with a generative model on the sphere conditionally to the radius. Doing so, we explicitly address the dependence between the variables at each radius, and in particular the limit angular distribution. Experiments performed on synthetic and real data support the relevance of our approach compared with vanilla VAE schemes and GANs tailed for extremes. In particular, we illustrate the ability to consistently sample extreme regions that have been never visited during the training stage.

Our contribution naturally advocates for future work, especially for extensions to multivariate extremes in time and space-time processes (Basrak & Segers, 2009; Liu et al., 2012) as well as to VAE for conditional generation problems (Zheng et al., 2019; Grooms, 2021).

[3]We do not include in these experiments the ParetoGAN. We did not succeed in training a satisfactory version of the ParetoGAN for the river flow dataset.

Table 2. Evaluation of the generation of multivariate extremes for the Danube river dataset: we report the proportion (in %) of elements satisfying $A_j^{(q)}$ in the training and test datasets as well as datasets sampled from the trained StdVAE and UExtVAE with the same size as the test dataset, where $A_j^{(q)} = \bigcap_{i=23}^{j} X_i > u_i^{(q)}$ with $u_i^{(q)}$ the value of the flow at q quantile for station i in test set. We report this analysis for quantile values of 0.9 and 9.99.

	q = 0.9			
	Train	Test	UExtVAE	StdVAE
$A_{25}^{(q)}$	5.9	6.6	5.0	3.8
$A_{26}^{(q)}$	4.9	6.0	4.6	3.3
$A_{27}^{(q)}$	3.8	5.1	4.1	2.5

	q = 0.99			
	Train	Test	UExtVAE	StdVAE
$A_{25}^{(q)}$	0.0	0.48	0.22	0.01
$A_{26}^{(q)}$	0.0	0.4	0.2	0.0
$A_{27}^{(q)}$	0.0	0.25	0.18	0.0

References

Allouche, M., Girard, S., and Gobet, E. Ev-gan: Simulation of extreme events with relu neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022.

Anderson, P. L. and Meerschaert, M. M. Modeling river flows with heavy tails. *Water Resources Research*, 34(9): 2271–2280, 1998.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Asadi, P., Davison, A. C., and Engelke, S. Extremes on river networks. *The Annals of Applied Statistics*, 9(4): 2023–2050, 2015.

Basrak, B. and Segers, J. Regularly varying multivariate time series. *Stochastic processes and their applications*, 119(4):1055–1080, 2009.

Bhatia, S., Jain, A., and Hooi, B. Exgan: Adversarial generation of extreme samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6750–6758, 2021.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science*, 1, 2022.

Bradley, B. O. and Taqqu, M. S. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier, 2003.

Breiman, L. On some limit theorems similar to the arc-sin law. *Theory of Probability & Its Applications*, 10(2): 323–331, 1965.

Cemgil, T., Ghaisas, S., Dvijotham, K., Gowal, S., and Kohli, P. The autoencoding variational autoencoder. *Advances in Neural Information Processing Systems*, 33: 15077–15087, 2020.

Chavez-Demoulin, V. and Roehrl, A. Extreme value theory can save your neck. *ETHZ publication*, 2004.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Das, B., Embrechts, P., and Fasen, V. Four theorems and a financial crisis. *International Journal of Approximate Reasoning*, 54(6):701–716, 2013.

Drees, H. and Sabourin, A. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943, 2021.

Embrechts, P. Copulas: A personal view. *Journal of Risk and Insurance*, 76(3):639–650, 2009.

Feder, R. M., Berger, P., and Stein, G. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Physical Review D*, 102(10):103504, 2020.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

Fortin, J.-Y. and Clusel, M. Applications of extreme value statistics in physics. *Journal of Physics A: Mathematical and Theoretical*, 48(18):183001, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Grooms, I. Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Quarterly Journal of the Royal Meteorological Society*, 147(734):139–149, 2021.

Hernandez-Campos, F., Marron, J., Samorodnitsky, G., and Smith, F. D. Variable heavy tails in internet traffic. *Performance Evaluation*, 58(2-3):261–284, 2004.

Huster, T., Cohen, J., Lin, Z., Chan, K., Kamhoua, C., Leslie, N. O., Chiang, C.-Y. J., and Sekar, V. Pareto gan: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pp. 4523–4532. PMLR, 2021.

Jalalzai, H., Clémençon, S., and Sabourin, A. On binary classification in extreme regions. *Advances in Neural Information Processing Systems*, 31, 2018.

Katz, R. W., Parlange, M. B., and Naveau, P. Statistics of extremes in hydrology. *Advances in water resources*, 25 (8-12):1287–1304, 2002.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Liu, Y., Bahadori, T., and Li, H. Sparse-gev: Sparse latent space model for multivariate extreme value time serie modeling. *arXiv preprint arXiv:1206.4685*, 2012.

Mhalla, L., Chavez-Demoulin, V., and Dupuis, D. J. Causal mechanism of extreme river discharges in the upper danube basin network. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):741–764, 2020.

Nash, C., Menick, J., Dieleman, S., and Battaglia, P. W. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

Naveau, P., Guillou, A., and Rietsch, T. A non-parametric entropy-based approach to detect changes in climate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):861–884, 2014.

Pasche, O. C. and Engelke, S. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*, 2022.

Pickands III, J. Statistical inference using extreme order statistics. *the Annals of Statistics*, pp. 119–131, 1975.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Resnick, S. I. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Rietsch, T., Naveau, P., Gilardi, N., and Guillou, A. Network design for heavy rainfall analysis. *Journal of Geophysical Research: Atmospheres*, 118(23):13–075, 2013.

Rudd, E. M., Jain, L. P., Scheirer, W. J., and Boult, T. E. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):762–768, 2017.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G. Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31(2):e2582, 2019.

Wan, P. and Davis, R. A. Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166, 2019.

Xie, X. *Analysis of Heavy-Tailed Time Series*. PhD thesis, University of Copenhagen, Faculty of Science, Department of Mathematical . . . , 2017.

Zhao, T., Zhao, R., and Eskenazi, M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.

Zheng, C., Cham, T.-J., and Cai, J. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019.

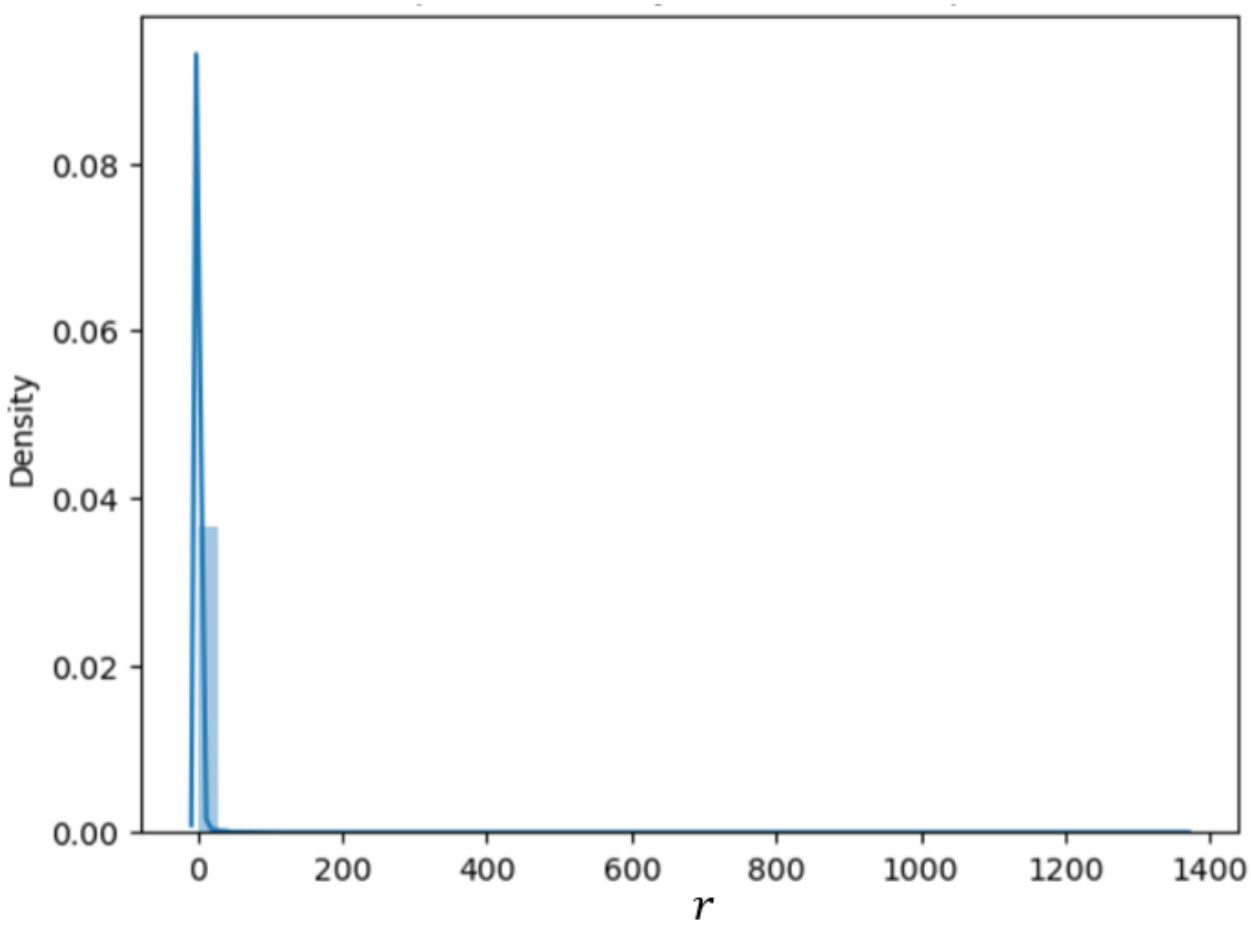

Figure 8. Empirical densities of synthesized radii R_1 for a batch of 1000 samples.

A. Datasets

A.1. Synthesized datasets

We sample in a space of dimension 5. We consider a sampling setting for the radius distribution denoted R_1:

$$R_1 \sim 2\mathbf{U} \times \mathbf{InvGamma}(\alpha_1 = 1.5\ ;\ \beta = 0.6)$$

with U uniform on $[0, 1]$ and π Bernoulli with parameter 0.3. The radius distribution is heavy-tailed with tail index α_1. The detailed expression of the conditional angular distribution $\mathbf{\Theta_1}|R_1 = r$ is given hereafter:

$$\mathbf{\Theta_1}|R_1 = r \sim Dir\left(\alpha_1(r), \alpha_1(r), \alpha_2(r), \alpha_2(r), \alpha_2(r)\right), \tag{15}$$

where $\alpha_1(r) = 3\left(2 - min(1, 1/2r)\right)$, $\alpha_2(r) = 3\left(1 + min(1, 1/2r)\right)$, and Dir stands for Dirichlet distribution. Figure 8 gives the empirical pdf of R_1 for a sample of 1000 values.

A.2. Danube river network discharge measurements

The upper Danube basin is an European river network which covers a large part of Austria, Switzerland and of the south of Germany. Figure 9 shows the topography of the Danube basin as well as the locations of the 31 stations at which daily measurements of river discharge are available for a 50 years time window. Danube river network dataset is make available by

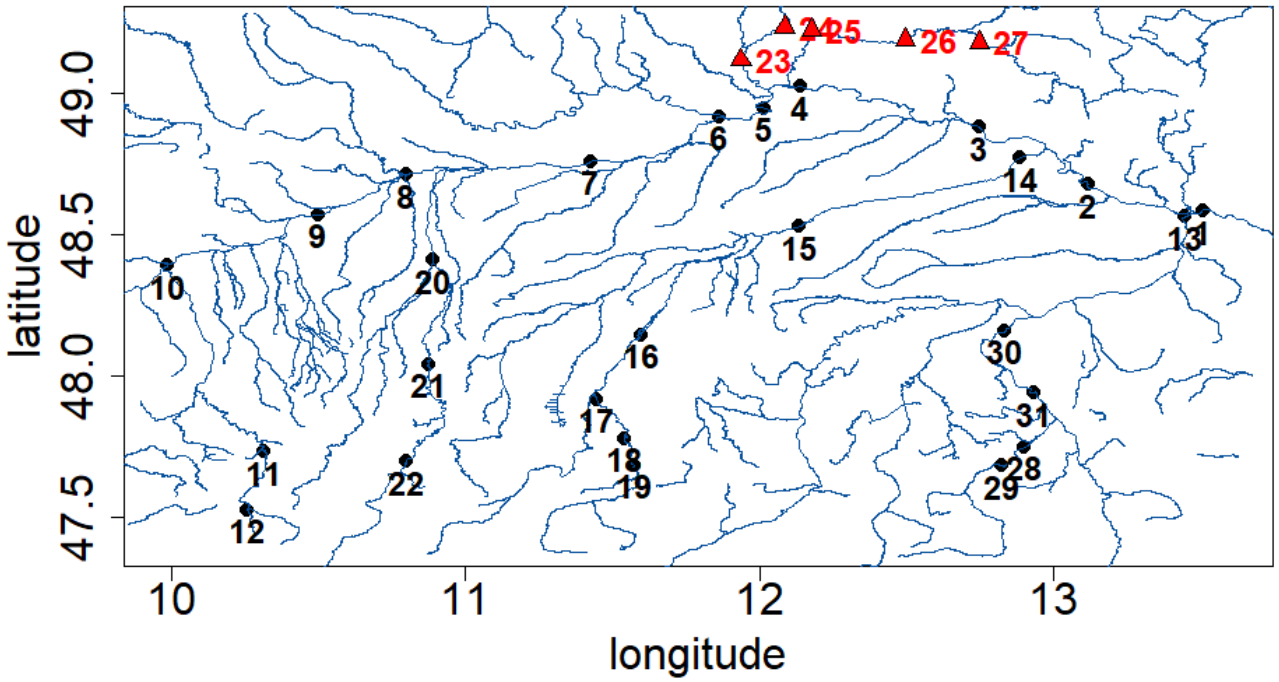

Figure 9. Topographic map of the upper Danube basin with 31 available gauging stations. A dataset of 50 years of daily measurements is considered (from 1960 to 2010). our training set consists of all measurements for the 5 stations indicated by red triangles

the Bavarian Environmental Agency at `http://www.gkd.bayern.de`. As river discharges usually exhibit heavy-tailed distribution, this dataset have been extensively studied in the community of multivariate extremes (see, e.g. Mhalla et al., 2020; Asadi et al., 2015). We consider measurements from a subset of 5 stations (red triangles in Figure 9) from which we want to sample.

B. Tail index estimation

Estimating the tail index of an univariate distribution from samples is not an easy task. To see this, we drew the Hill plot (see e.g Resnick, 2007, Section 4.4), (Xie, 2017, Section 2.2) for R_1 in Figure 10. The Hill plot is a common tool in the extreme value community for estimating the tail index of a distribution. If the graph is approximately constant from a certain order statistics, this constant is an estimator of the inverse of the tail index. We note that the Hill plot is of little use in this case because the graph does not exhibit clearly a plateau. Other methods are also broadly used to estimate the tail index within the extreme value community such as maximum likelihood estimation. It involves fitting a GP distribution (eq. (5)) to the subset of data above a certain threshold (see Coles et al., 2001, for details). For example, on train dataset of R_1, the maximum likelihood estimation gives an estimation of 1.28 for the tail index when the threshold corresponds to a 0.8-quantile while it becomes 1.67 for a 0.9-quantile.

C. Criteria

C.1. KL divergence upon threshold

Let us assume that we have n samples $\mathbf{R_{true}} = (R^1_{true}, R^2_{true}, ..., R^n_{true})$ from the true radius distribution and m samples $\mathbf{R_{gen}} = (R^1_{gen}, R^2_{gen}, ..., R^m_{gen})$ from a generative approach. Let denote $\tilde{\bar{F}}_{true}$, $\tilde{\bar{F}}_{gen}$ empirical estimators of the tail functions chosen to be non-zero above the upper observed value. Then the empirical estimate $\hat{K}_u(\mathbf{R_{true}}, \mathbf{R_{gen}})$ of the KL

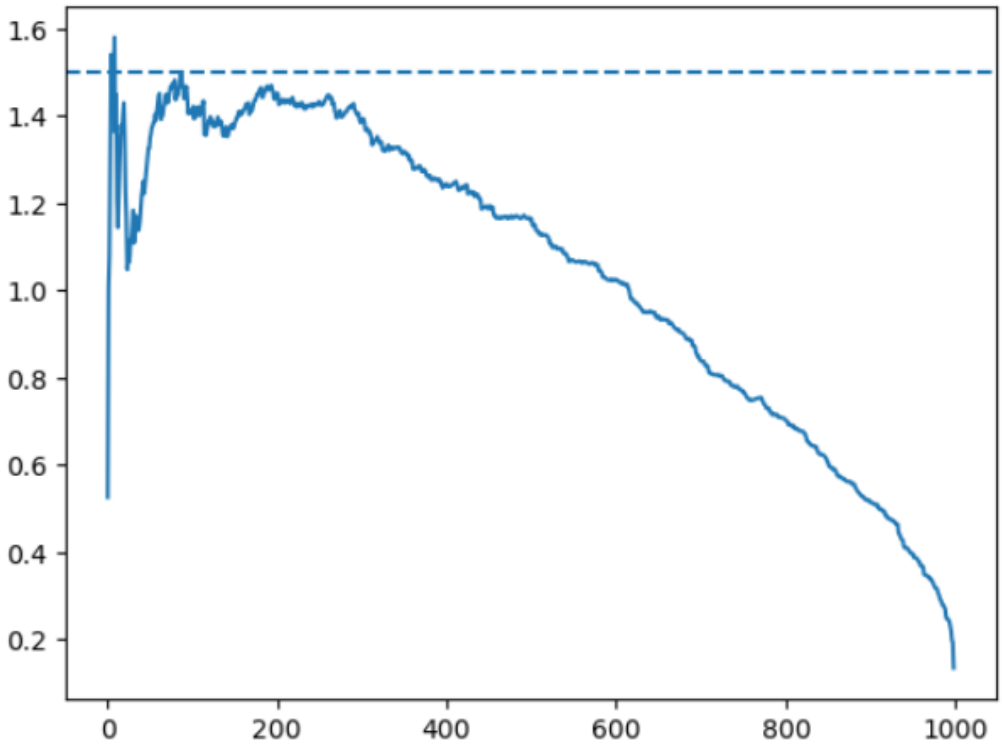

Figure 10. Hill plot for the 1000 R_1 samples of train and validation set (blue curve), the dashed line indicates the true value of the tail index, i.e. 1.5.

divergence beyond a threshold u is given by:

$$\hat{K}_u(\mathbf{R_{true}}, \mathbf{R_{gen}}) = \begin{aligned} &-1 - \tfrac{1}{N_n} \textstyle\sum_{i=1}^{m} \log\left(\frac{\tilde{\bar{F}}_{gen}(max(R^i_{gen},u))}{\tilde{\bar{F}}_{gen}(u)}\right) \\ &-1 - \tfrac{1}{M_m} \textstyle\sum_{i=1}^{n} \log\left(\frac{\tilde{\bar{F}}_{true}(max(R^i_{true},u))}{\tilde{\bar{F}}_{true}(u)}\right), \end{aligned} \tag{16}$$

where N_n and M_m are the number of samples above threshold u respectively among $\mathbf{R_{true}}$ and $\mathbf{R_{gen}}$.

C.2. Wasserstein distance

Assume we have n samples $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ from a random vector X and m samples $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)$ from another random vector with same dimension. Then, the Wasserstein distance we used is defined by:

$$W(\mathbf{X}, \mathbf{Y}) = \left(\min_{\gamma \in \mathbb{R}_+^{n \times m}} \sum_{i,j} \gamma_{i,j} ||\mathbf{x}_i - \mathbf{y}_j||_2 \right)^{\frac{1}{2}}, \tag{17}$$
$$\text{with } n\gamma\mathbf{1} = \mathbf{1} \; ; \; m\gamma^T\mathbf{1} = \mathbf{1},$$

with $\mathbf{1}$ a vector filled with ones, and $||.||_2$ the euclidean distance. The rescaled version of the Wasserstein distance beyond a threshold r is then given by :

$$W_r(\mathbf{X}, \mathbf{Y}) = \frac{W(\mathbf{X_r}, \mathbf{Y_r})}{r^2}, \tag{18}$$

where $\mathbf{X_r}$ (respectively $\mathbf{Y_r}$) consists of the elements of $\mathbf{X}$ (respectively $\mathbf{Y}$) of norm greater than r.

C.3. Threshold selection

To assess independence between radius and angular distributions, Wan & Davis (2019) relies on the following hypothesis testing framework:

- H_0: R_1/r_n and $\mathbf{\Theta_1}$ are independent given $R_1 > r_n$;
- H_1: R_1/r_n and $\mathbf{\Theta_1}$ are not independent given $R_1 > r_n$.

Considering this, the authors propose a p-value for computing H_0 with respect to H_1, such that the p-value follows a uniform distribution if H_0 is true and is close to 0 when H_1 is true. Thus, for a given threshold, when we average the p-values, we should find about 0.5 if H_0 is true and closer to 0 when H_1 is true. Let define the empirical distance covariance (Székely et al., 2007) between n observations $\{\mathbf{X_i}\}$ of a random vector $\mathbf{X}$ and n observations $\mathbf{Y_i}$ of a random vector $\mathbf{Y}$ by:

$$T_n(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} ||\mathbf{X_i}-\mathbf{X_j}||_2||\mathbf{Y_i}-\mathbf{Y_j}||_2 + \frac{1}{n^4}\sum_{i,j,k,l=1}^{n} ||\mathbf{X_i}-\mathbf{X_j}||_2||\mathbf{Y_k}-\mathbf{Y_l}|| - \frac{2}{n^3}\sum_{i,j,k=1}^{n} ||\mathbf{X_i}-\mathbf{X_j}||_2||\mathbf{Y_i}-\mathbf{Y_k}||_2,$$

with $||.||_2$ the euclidean distance. Note that $\mathbf{X}$ and $\mathbf{Y}$ have not necessarily equal sizes.
Let us consider $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N})$ a sequence of observed vector of $\mathbb{R}^d$, with a polar decomposition $(R, \mathbf{\Theta})$. Given a decreasing sequence of candidate radius threshold r_k, we want to find the smallest such as radius and angle of the polar decomposition of $\mathbf{X}$ can be considered independent. For a fixed threshold r_k, we first restrict the dataset to observations which have radii above r_k. Then, we randomly choose a subsample $\{\frac{R_i}{r_k}, \mathbf{\Theta_i}\}$ of size n_k. We can then compute $T_{n,k}$, the empirical covariance between the radii and angles within the subsample. To compute a p-value of $T_{n,k}$ under the assumption that the conditional empirical distribution is a product of the conditional marginals, we take a large number L of subsamples of size n_k of $\{R_i\}$ on the one hand, and of $\{\mathbf{\Theta_i}\}$ on the other hand. We can compute the empirical covariances $\{\tilde{T}_{n,k}\}_{l=1}^{L}$ between radii and angles. The p-value pv_k of $T_{n,k}$ is the empirical value of $T_{n,k}$ relative the $\{\tilde{T}_{n,k}\}_{l=1}^{L}$. This process is repeated m times which produces m estimates of pv_k. The p-value is then the mean of the estimates. If the radius and angular distribution are independent, the p-value should be around 0.5, otherwise it is closer to 0.

D. Implicit reparametrization

When it comes to optimization of the cost of eq. (2), explicit reparametrization (see eq. (3)) is not feasible for the proposed framework of eq. (12). Leveraging the work of Figurnov et al. (2018), we use an implicit reparametrization. It consists in differentiating the Monte Carlo estimator of $E_{q_\phi(Z|r^{(i)})}[f(Z)]$ using the following:

$$\nabla_\phi E_{q_\phi(Z|r^{(i)})}[f(Z)] = -E_{q_\phi(Z|r^{(i)})}[\nabla_z f(z)\nabla_\phi F_{q_\phi}(z)(\nabla_z F_{q_\phi}(z))^{-1}],$$

with F_{q_ϕ} the cumulative distribution function of q_ϕ. An implicit reparametrization of Gamma distribution, as well as inverse Gamma and many others, is availalble as a Tensorflow package named TensorflowProbability[4].

E. GAN and limit angular measure

In a GAN setting, one can think of the following framework for generating multivariate heavy-tailed data:

Assumption E.1. We generate heavy-tail random vector in the non-negative orthant through a generator:

$$\mathbf{X} = G(\mathbf{Z}),$$

with G a ReLU neural network and $\mathbf{Z}$ a d-dimensional random vector with i.i.d heavy-tailed margins.

This generating strategy is used by Feder et al. (2020) and Huster et al. (2021). Huster et al. (2021) proved that in the one dimensional case, $\mathbf{X}$ is heavy-tailed with same shape parameter as $\mathbf{Z}$. In the limit of extreme values, one can ask oneself what are the dependency structures between the variables that such a model can represent. This corresponds to the limit angular measure defined in eq. (6). If we designate $\mathbf{S}_G$ as the probability measure on the sphere when $||\mathbf{X}|| \to \infty$, we can state the following theorem:

[4]Details could be found at https://www.tensorflow.org/probability

Theorem E.2. *Under Assumption E.1,* $\mathbf{S}_G$ *is concentrated on a number of axis less than d.*

This result means that in the limit of infinite radius, $\mathbf{X}$ is a.s located on a specific axis. While extracting certain principal directions in extreme regions is a useful tool for comprehensive analysis of a dataset (Drees & Sabourin, 2021), it is severely lacking in flexibility when it comes to generating extreme samples. To circumvent this difficulty, we chose to describe the data by its polar decomposition and to generate the angle and the radius separately. Namely, we write $\mathbf{X} = (R, \mathbf{\Theta})$ as explained in section 2.3. This allows to make explicit the dependency structure of the data whatever the radius is, especially for large radii. We can then obtain more varied limit angular measures than measures concentrated on a finite number of axes. Figure 5 illustrates this difference by showing the limit angular density for both ParetoGAN and our approach.

F. Proofs

F.1. Proof of Proposition 3.5

The pdf of Y is given by :

$$\begin{aligned} p(y) &= \int_x p(y|x)p(x)dx \\ &= y^{a_\theta - 1} \int_x \frac{\beta_\theta(x)^{a_\theta}}{\Gamma(a_\theta)} e^{-y\beta_\theta(x)} p(x)dx \\ &= \frac{y^{a_\theta-1}}{\Gamma(a_\theta)} \mathbf{E}_{x\sim p_\alpha}[\beta_\theta(x)^{a_\theta} e^{-y\beta_\theta(x)}] \end{aligned}$$

Let assume $\beta_\theta(x) = \frac{c}{x} + \epsilon(x)$ with $\lim_{x\to+\infty} \epsilon(x) = o(\frac{1}{x})$. Using the change of variables $z = \frac{1}{x}$, we can rewrite :

$$\begin{aligned} \mathbf{E}_{x\sim p_\alpha}[\beta_\theta(x)^{a_\theta} e^{-y\beta_\theta(x)}] &= \mathbf{E}_{z\sim\Gamma(\alpha,1)}[\beta_\theta \left(\frac{1}{z}\right)^{a_\theta} e^{-y\beta_\theta\left(\frac{1}{z}\right)}] \\ &= \mathbf{E}_{z\sim\Gamma(\alpha,1)} \left[\left(cz + \epsilon(\frac{1}{z})\right)^{a_\theta} e^{-y\left(cz+\epsilon(\frac{1}{z})\right)}\right] \end{aligned}$$

We state $\tilde{\epsilon}(x) = \epsilon\left(\frac{1}{x}\right)$, then:

$$\mathbf{E}_{x\sim p_\alpha}[\beta_\theta(x)^{a_\theta} e^{-y\beta_\theta(x)}] = c^{a_\theta} y^{-a_\theta} \mathbf{E}_{\tilde{z}\sim\Gamma(\alpha,\frac{1}{y})} \left[\left(\tilde{z} + \frac{y}{c}\tilde{\epsilon}(\frac{\tilde{z}}{y})\right)^{a_\theta} e^{-\left(c\tilde{z}+y\epsilon(\frac{\tilde{z}}{y})\right)}\right].$$

One can then show using dominated convergence that :

$$\mathbf{E}_{x\sim p_\alpha}[\beta_\theta(x)^{a_\theta} e^{-y\beta_\theta(x)}] \underset{y\to+\infty}{\sim} C y^{-\alpha-a_\theta}$$

From this equivalent, we can infer that for $y \to +\infty$, $p(y) \propto y^{-\alpha-1}$. From Karamata's theorem ((see Resnick, 2007)), the tail function of Y is regularly varying with index α. It suffices to conclude that Y has tail index α. □

F.2. Proof of Theorem E.2

The proof proceeds by a series of step. First, we note that the latent vector has a limit angular distribution located on the basis axis. Then, we proove that several transformations of a random vector which limit angular distribution concentrated on some axes has still limit angular distribution concentrated on axes. The studied transformations are: multiplication by a matrix, addition of a bias, mapping with a ReLU unit. By applying iteratively this steps, we prove that $\mathbf{X}$ has a limit angular measure concentrated on axes for any ReLU neural network G.
First the limit angular measure of $\mathbf{Z}$ is concentrated on the basis axis:
$e_i = (0, \cdots, 1, \cdots, 0)$, with $i = 1, \cdots, d$.
A proof is given in (Resnick, 2007), Section 6.5. Let denote $\mathbf{S_Z}$ this limit angular distribution.
For the following, we need to specify that a random vector $\mathbf{Y}$ has multivariate regular variation if there exists a function $b \to \infty$ and a Radon measure $\mu_\mathbf{Y}$ such that:

$$\lim_{t\to\infty} tP\left(\frac{\mathbf{Y}}{b(t)} \in \bullet\right) \xrightarrow{v} \mu_\mathbf{Y}(\bullet).$$

Lemma F.1. *If the d-dimensional random vector $\mathbf{Y}$ has multivariate regular variation with limit angular measure concentrated on axes, and $\mathbf{W}$ is a $n \times d$ matrix, then $(\mathbf{WY})_+$ has regular variation and its limit angular measure is concentrated on axes.*

Proof.

$$\lim_{t\to\infty} tP\left(\frac{(\mathbf{WY})_+}{b(t)} \in \bullet\right) = tP\left(\frac{\mathbf{Y}}{b(t)} \in \mathbf{W}^{-1}(\bullet)\right),$$
$$= \mu_{\mathbf{Y}} \circ \mathbf{W}^{-1}(\bullet).$$

$(\mathbf{WY})_+$ has regular variation. Moreover if the limit angular measure of $\mathbf{Y}$ is concentrated on $d' \leq d$ lines $\bigcup_{i=1}^{d'}\{t\mathbf{e}_i,\ t>0\}$, then $\mathbf{WY}$ is concentrated on $\bigcup_{i=1}^{d'}\{t(\mathbf{We}_i)_+,\ t>0\}$. Note that $\mathbf{WY}$ is then concentrated on a number of axes less or equal to d'.

□

Lemma F.2. *If the d-dimensional random vector $\mathbf{Y}$ has multivariate regular variation with limit angular measure concentrated on axes, and $\mathbf{b}$ is a d-dimensional vector, then $(\mathbf{Y}+\mathbf{b})_+$ has regular variation and its limit angular measure is concentrated on axes.*

Proof.

$$\lim_{t\to\infty} tP\left(\frac{(\mathbf{Y}+\mathbf{b})_+}{b(t)} \in \bullet\right) = \lim_{t\to\infty} tP\left(\frac{\mathbf{Y}}{b(t)} \in \bullet\right) \xrightarrow{v} \mu_{\mathbf{Y}}(\bullet).$$

□

From Lemma F.1 and Lemma F.2 we get that for any random vector with multivariate regular variation and limit angular measure concentrated on lines, any matrix $\mathbf{W}$ and bias b, $(\mathbf{WY}+\mathbf{b})_+$ has multivariate regular variation with limit angular measure concentrated on lined. This transformation corresponds to a layer of a ReLU neural network. Applying iteratively this transformation to the initial latent random vector $\mathbf{Z}$, we obtain that $\mathbf{X} = G(\mathbf{Z})$ has multivariate regular variation with limit angular measure concentrated on lines.