Journal of Machine Learning Research 23 (2022) 1-39 Submitted 6/21; Revised 3/22; Published 5/22

EV-GAN: Simulation of extreme events with ReLU neural networks

Michaël Allouche MICHAEL.ALLOUCHE@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique
Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau Cedex, France

Stéphane Girard STEPHANE.GIRARD@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
38000 Grenoble, France

Emmanuel Gobet EMMANUEL.GOBET@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique
Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau Cedex, France

Editor: Shakir Mohamed

Abstract

Feedforward neural networks based on Rectified linear units (ReLU) cannot efficiently approximate quantile functions which are not bounded, especially in the case of heavy-tailed distributions. We thus propose a new parametrization for the generator of a Generative adversarial network (GAN) adapted to this framework, basing on extreme-value theory. An analysis of the uniform error between the extreme quantile and its GAN approximation is provided: We establish that the rate of convergence of the error is mainly driven by the second-order parameter of the data distribution. The above results are illustrated on simulated data and real financial data. It appears that our approach outperforms the classical GAN in a wide range of situations including high-dimensional and dependent data.

Keywords: Extreme-value theory, neural networks, generative models

1. Introduction

Context of risks. Analyzing extreme events is an important issue in economics, engineering, and life sciences, among other fields, with significant applications such as actuarial risks (Asmussen and Albrecher, 2010), communication network reliability (Robert, 2003), aircraft safety (Prandini and Watkins, 2005), analysis of epidemics, and so forth... In the last two decades, it has taken even more importance in financial risk management, because of the increasing number of shocks and financial crises. Among the wide range of exercises in this field, stress test (European Banking Authority, 2014) has become a main guideline for the regulator in order to assess the banking system resilience against the realizations of various categories of risk (market, credit, operational, climate, etc). To this end, numerical simulation of unfavorable extreme (but plausible) scenarios is a major tool to study the consequences on these risks. Given a stochastic model of risks, various sampling schemes are available (for instance, using importance sampling (Bucklew, 2004, Chapter 4), MCMC with splitting – (Gobet and Liu, 2015), or interacting particles system – (Del Moral and Garnier, 2005)), with the potential advantage of reducing the statistical fluctuation over a

naive Monte Carlo method. Though presumably more informative for a given number M of samples, these methods suffer from a higher computational complexity (notably in high dimension): Thus, one might wonder how to get extra samples in an efficient way, by leveraging the previous M samples. Somehow, the situation is similar to a case where the previous samples are viewed as observed data and where we seek a data-driven method able to sample similarly to that empirical distribution, without necessarily the knowledge of the sampling method that has generated the observed data. This corresponds to the recent paradigm of Generative adversarial network (GAN) models initiated by (Goodfellow et al., 2014) or of Variational autoencoder (VAE) by (Kingma and Welling, 2014). The novelty in our work is relative to the context of risks, where we are interested in a generative data-based model able to reproduce – with high-fidelity – specific extreme statistical properties of a data set, while being fast in the simulation phase. This challenge arises both in the context of true historical data sets (see experiments in Section 4) or when the learning data set has been generated by sophisticated Monte Carlo methods.

Background results. Generally speaking, different types of generative models have been developed lately (Foster, 2019) and in this work, we focus on GANs, which have gained a tremendous popularity from the original work of (Goodfellow et al., 2014) and its extension using the Wasserstein distance (Arjovsky et al., 2017). Kuratowski Theorem (Bertsekas and Shreve, 1978, Chapter 7)-(Villani, 2009, p.8) ensures that any random variable X on $\mathbb{R}^d$ (and more generally on a Polish space) can be obtained by

$$X \overset{\mathrm{d}}{=} G(Z) \tag{1}$$

for some measurable function G and some latent random variable Z in dimension d' (see Lemma 7 in the Appendix for a constructive proof with $Z \sim \mathcal{U}([0,1])$ and $d' = 1$) such that for each mth marginal, $m \in \{1, \dots, d\}$, one has $\tilde{X}^{(m)} := G^{(m)}(Z) \overset{\mathrm{d}}{=} X^{(m)}$. This result is one key to understand the ability of GANs to simulate realistic samples in a space of high dimension d, starting from a latent space of moderate dimension d'. In practice, the selection of this latent dimension is an open problem in the generative neural networks literature. A GAN scheme is aimed at approximating the unknown G through a parametric family of neural networks (NN) $\mathcal{G} = \{G_\theta : \mathbb{R}^{d'} \to \mathbb{R}^d, \theta \in \Theta\}$ and to learn the optimal parameter $\theta^\star$ from a data set $\{X_i \in \mathbb{R}^d, i = 1, \dots, n\}$ of i.i.d samples from an unknown distribution p_X. It is performed by optimizing an objective function which can be interpreted as an adversarial game between a generator and a discriminator chosen in a parametric family of functions $\mathcal{D} = \{D_\phi : \mathbb{R}^d \to [0,1], \phi \in \Phi\}$. In other words, $D_\phi(x)$ represents the probability that an observation x is drawn from p_X. Both the generator and the discriminator are NNs with opposite objectives: The former tries to mimic real data which seem likely by the discriminator, while the latter tries to distinguish between the two sources. In (Goodfellow et al., 2014), this optimization problem is defined as:

$$\min_{\theta \in \Theta} \max_{\phi \in \Phi} \left[\mathbb{E}_{p_X} \left(\log D_\phi(X) \right) + \mathbb{E}_{p_Z} \left(\log \left(1 - D_\phi \left(G_\theta(Z) \right) \right) \right) \right].$$

Theoretical results on GANs have been established in (Biau et al., 2020a,b; Haas and Richter, 2020), see also (Remlinger et al., 2021; Wiese et al., 2020) for the generation of financial time-series and (Allouche et al., 2021) for the generation of fractional Brownian motion.

Extreme events generation using GANs has been investigated in Finance (Wiese et al., 2020), in meteorology (Bhatia et al., 2020), in cosmological analysis (Feder et al., 2020) and in anomaly detection (Dionelis et al., 2020). One strategy is to learn a light-tail model on some transformed data and then transform back the generator outputs for recovering the heavy-tailed data property. Possible transformations include the Lambert W function (Wiese et al., 2020). Another approach is to use directly a heavy-tailed latent variable in the GAN setting (Feder et al., 2020; Huster et al., 2021). It is shown in (Huster et al., 2021) that generator outputs follow the desired heavy-tailed distribution. Alternative metric spaces are also introduced to ensure the loss function to be finite. To be effective, the method however requires the accurate estimation of the tail-index associated with each heavy-tailed marginal distribution, which is a challenging task in extreme-value theory, see next paragraph for details: As a main difference, our approach does not require the estimation of tail-indices. Alternatively, in (Bhatia et al., 2020), a distribution shifting is first introduced in order to reduce the lack of training data in the extreme tails. Second, a GAN parametrization conditioned by samples drawn from a generalized Pareto distribution is fitted to the shifted data. Finally, an additional term representing some distance to a desired extremeness is added to the loss function. Although numerical results on images are promising, we do not think that the proposed parametrization gives theoretical support for generating extreme observations in the sense that no error or complexity bounds are provided in the NN architecture of the generator.

Our contributions. In a GAN setting, our purpose is to cope with two prominent issues, that are mostly related to extreme-value theory. First, the number of data available in extreme regions must be relatively small, by definition (even in the case of data that are output of sophisticated sampling methods). Second, we restrict to the challenging situation of heavy-tailed distributions (in the Fréchet maximum domain of attraction), where by definition, extreme data take very large values. Therefore, the usual GAN approach cannot work, as we now explain (and as the reader will check from our numerical experiments in Section 4). Consider for a while the case $d = d' = 1$ and say that G in (1) is approximated by a ReLU NN under the form

$$G_\theta(z) = \sum_{j=1}^{J} a_j \sigma(w_j z + b_j), \tag{2}$$

where $\sigma(x) := \max(x, 0)$ is the ReLU function, $\theta = \{(a_j, w_j, b_j),\ j = 1, \dots, J\} \in \Theta = \mathbb{R}^{3J}$ and J is the number of units in the hidden layer. On the one-hand, if the latent random variable Z were bounded, the output would be bounded (Huster et al., 2021, Proposition 1) and by no means, it would be a good candidate for fitting the distribution of the unbounded random variable X. On the other hand, taking for Z a Gaussian vector as it is often chosen, for example in (Bhatia et al., 2020), would lead to a light-tailed distribution for $G_\theta(Z)$ (Huster et al., 2021, Theorem 1) since G_θ is sublinear w.r.t. the input (Vladimirova et al., 2018), whereas we focus on the heavy-tail case. Similar arguments are given in (Wiese et al., 2019, Theorem 1) to emphasize that the generator cannot generate samples with heavier distribution than its inputs. Clearly, such a parameterization (2) of the generator cannot be efficient as far as extreme values are concerned. Note that deeper NN would not overcome this issue either.

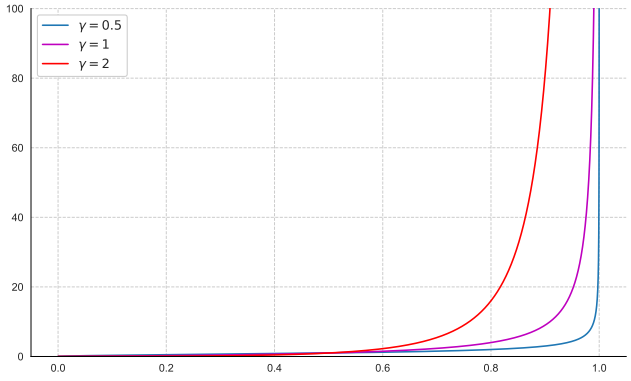


Figure 1: Quantile function associated with the Burr distribution $u \in (0,1) \mapsto q_X(u)$ with tail-index $\gamma \in \{1/2, 1, 2\}$ and second-order parameter $\rho = -1$, see Table 4 for the parameterization.

To introduce our new parametrization called Extreme-Value GAN (EV-GAN), consider first a real random variable X with cumulative distribution function F_X defined on $\mathbb{R}$. The inversion method by Von Neumann Eckhardt (1987) gives that one can set $G(u) := q_X(u) := \inf\{x : F_X(x) \geq u\}$ with $U \sim \mathcal{U}([0,1])$. Since we shall focus on distributions in the Fréchet maximum domain of attraction (de Haan and Ferreira, 2006, Theorem 1.2.1) with positive tail-index γ, the associated survival function $\bar{F}_X(x) := 1 - F_X(x)$ decays at rate $x^{-1/\gamma}$ when $x \to \infty$, which implies that $q_X(u)$ diverges as $u \to 1$ at rate $(1-u)^{-\gamma}$. The tail-index γ is thus the main driver of the behavior of extreme quantiles, see Figure 1 for an illustration. To be in a position to apply results such as the Universal approximation theorem (Cybenko, 1989) (any continuous function on $[0,1]$ can be approximated with arbitrary precision by a one hidden layer NN), we shall transform the quantile function to avoid divergence in the neighborhood of $u = 1$. To this end, for all $(u, y) \in [0,1) \times (0, \infty)$, let

$$H_u(y) = -\log(y) \,/ (\log(1-u^2) - \log(2)) \qquad \text{and} \qquad f^{\mathrm{TIF}}(u) = H_u(q_X(u)). \tag{3}$$

It will appear in the sequel that f^{TIF} is continuous on $[0,1]$ for all F_X in the Fréchet maximum domain of attraction, with $f^{\mathrm{TIF}}(u) \to \gamma$ as $u \to 1$; f^{TIF} is thus referred to as the Tail-index function (TIF). Therefore, a ReLU NN could well approximate f^{TIF} thanks to the Universal approximation theorem, but to get even better approximation, we shall consider a correction of the Tail-index function:

$$f^{\mathrm{CTIF}}(u) = f^{\mathrm{TIF}}(u) - \sum_{k=1}^{6} \kappa_k e_k(u), \quad u \in [0,1],$$

which enjoys higher regularity in the neighborhood of $u = 1$. See Paragraph 2.2 for a definition of functions $e_1, \ldots, e_6$ and coefficients $\kappa_1, \ldots, \kappa_6$: The functions $e_1, \ldots, e_6$ are universal (see (19)) and as such, they do not depend on the distribution parameters, only coefficients $\kappa_1, \ldots, \kappa_6$ may depend on them. Now use a NN to approximate the smooth function f^{CTIF},

deduce an approximation of f^{TIF}, and of the quantile function by composing with H_u^{-1} for each u (in view of (3)): All in all, we obtain the so-called EV-GAN parametrization defined for all $(z, x) \in [0,1] \times (0, \infty)$ as

$$G_\psi^{\mathrm{TIF}}(z) = H_z^{-1}\left(\sum_{j=1}^{J} a_j \sigma(w_j z + b_j) + \sum_{k=1}^{6} \kappa_k e_k(z)\right), \tag{4}$$

$$\text{with} \quad H_z^{-1}(x) := \left(\frac{1-z^2}{2}\right)^{-x}. \tag{5}$$

In the multidimensional setting $d > 1$ and $d' > 1$, our strategy of approximation consists in preserving the same parametric form for each marginal component, and in mixing the latent components to generate dependence between the d coordinates (see Corollary 6): The m-th coordinate will take the form, with $z = (z^{(1)}, \ldots, z^{(d)})$,

$$G_\psi^{\mathrm{TIF},(m)}(z^{(1)}, \ldots, z^{(d')}) = H_{z^{(m)}}^{-1}\left(\sum_{j=1}^{J} a_j^{(m)} \sigma\left(\sum_{i=1}^{d'} w_j^{(i)} z^{(i)} + b_j\right) + \sum_{k=1}^{6} \kappa_k^{(m)} e_k\left(z^{(m)}\right)\right). \tag{6}$$

Let us highlight that, in (6), the mth coordinate of the generator $G_\psi^{\mathrm{TIF}}(z)$ involves the mth coordinate of z which is a $d'-$ dimensional vector. The above construction of the EV-GAN generator thus constraints the latent dimension to be larger than the dimension of the data: $d' \geq d$. The architecture of the associated NN is illustrated on Figure 2 in the case $d = 2$ and $d' = 3$. We prove in Theorem 4 that the above EV-GAN parametrization converges uniformly coordinate-wise, in the log-scale of the $H_\cdot$-transform. Joint convergence for all coordinates is an open question, which is related to the delicate notion of upper tail dependence. However, numerical experiments in multidimensional settings fully support the relevance of this parametrization. We observe that tail dependencies are extremely well reproduced.

The rest of the paper is organized as follows. The transformation of the quantile function q_X associated with an heavy-tailed distribution F_X into a regular function f^{CTIF} is presented in Section 2: Under a second-order assumption, we show that f^{CTIF} can be uniformly approximated by a one hidden layer NN with some rate depending on the second-order parameter ρ, which plays a crucial role in extreme-value theory. Auxiliary results and technical proofs are postponed to the Appendix. The performance of the method is illustrated on simulated data (Section 4) and real financial data (Section 5). It is shown that, in both experiments, our approach largely outperforms the classic GAN method. Some conclusions and directions of future research are discussed in Section 6.

2. Main results

First, the construction of the proposed transformation of the quantile function is developed and its approximation by a ReLU NN is then investigated.

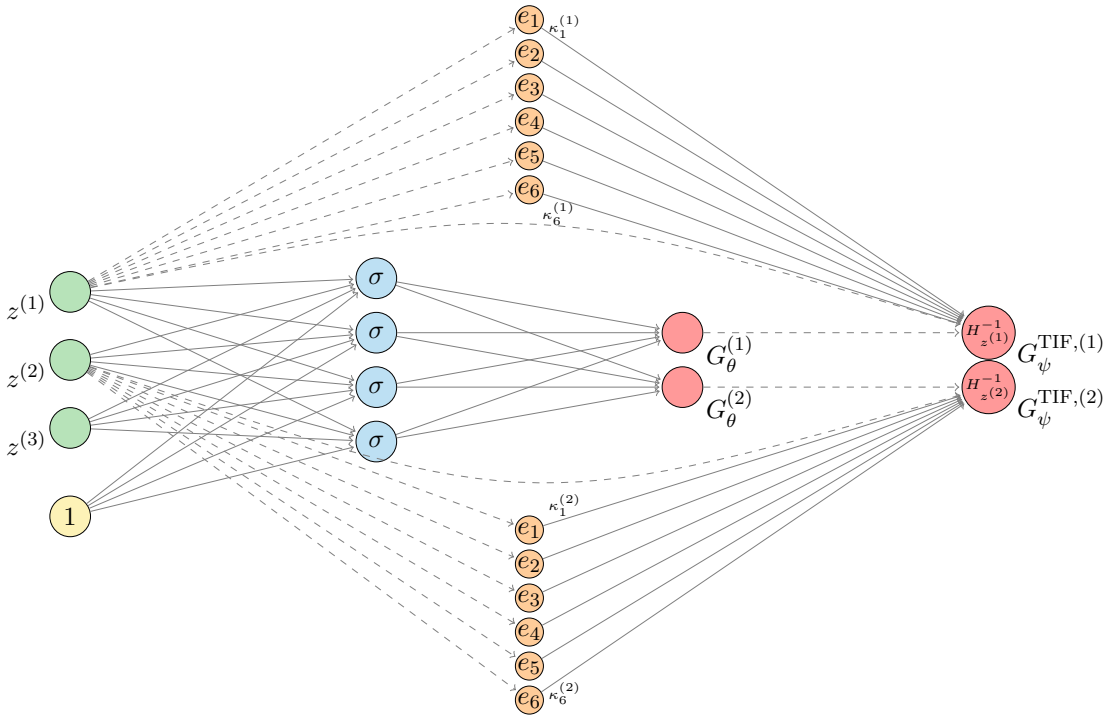


Figure 2: Generator of the EV-GAN with one hidden layer, $d' = 3$ and $d = 2$.

2.1 TIF regularity

In this section, we discuss the construction and the extension of (3). The objective is to build a tail-index function which may be well approximated by a NN. Let X be a real random variable and denote by F_X its cumulative distribution function supposed to be continuous and strictly increasing. We focus on the case of heavy-tailed distributions, *i.e.* when F_X is attracted to the maximum domain of Pareto-type distributions with tail-index $\gamma > 0$. From Bingham et al. (1987), the survival function $\bar{F}_X := 1 - F_X$ of such a heavy-tailed distribution can be expressed as

($\mathbf{H_1}$): $\bar{F}_X(x) = x^{-1/\gamma}\ell_X(x)$, where ℓ_X is a slowly-varying function at infinity *i.e.* such that $\ell_X(\lambda x)/\ell_X(x) \to 1$ as $x \to \infty$ for all $\lambda > 0$.

In such a case, $\bar{F}_X$ is said to be regularly-varying with index $-1/\gamma$ at infinity, which is denoted for short by $\bar{F}_X \in RV_{-1/\gamma}$. Similarly, we shall note $\ell_X \in RV_0$. The tail-index γ tunes the tail heaviness of the distribution function F_X. Assumption ($\mathbf{H_1}$) is recurrent in risk assessment, since actuarial and financial data are most of the time heavy-tailed, see for instance the recent studies Alm (2016); Chavez-Demoulin et al. (2014) or the monographs Embrechts et al. (1997); Resnick (2007). The Pareto distribution is the simplest example

of heavy-tailed distribution, since, in this case, ℓ_X in $(\mathbf{H_1})$ is constant. See Table 4 for more sophisticated examples. As a consequence of the above assumptions, the tail quantile function $x \mapsto q_X(1-1/x)$ is regularly-varying with index γ at infinity, see (de Haan and Ferreira, 2006, Proposition B.1.9.9), or, equivalently,

$$q_X(u) = (1-u)^{-\gamma} L\left(\frac{1}{1-u}\right), \tag{7}$$

for all $u \in (0,1)$ with $L \in RV_0$. Without of loss of generality, one can assume that $\eta := \mathbb{P}(X \geq 1) \neq 0$ and, since, we focus on the upper tail behavior of X, introduce the random variable $Y = X$ given $X \geq 1$. It follows that the quantile function of Y is given by

$$q_Y(u) = q_X(1-(1-u)\eta), \tag{8}$$

for all $u \in (0,1)$. Note that one could also assume $X \geq 1$ and set $\eta = 1$ in order to simplify the following derivations. Finally, we consider the Tail-index function (TIF) obtained by plugging (8) into (3):

$$f^{\mathrm{TIF}}(u) = -\frac{\log q_X\big(1-(1-u)\eta\big)}{\log(1-u^2) - \log 2}, \tag{9}$$

for all $u \in (0,1)$. Extra assumptions on F_X, or equivalently on L, are necessary such that f^{TIF} is differentiable. Consider the Karamata representation of the slowly-varying function L (de Haan and Ferreira, 2006, Definition B1.6):

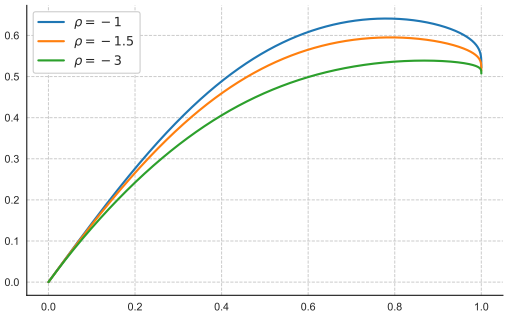
$$L(x) = c(x) \exp\left(\int_1^x \frac{\varepsilon(t)}{t}\,\mathrm{d}t\right), \tag{10}$$

where $c(x) \to c_\infty$ as $x \to \infty$ and ε is a measurable function such that $\varepsilon(x) \to 0$ as $x \to \infty$. Our second main assumption then writes:

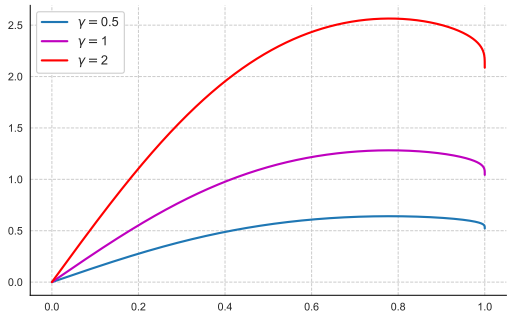
$(\mathbf{H_2})$: $c(x) = c_\infty > 0$ for all $x \geq 1$ and $\varepsilon(x) = x^\rho \ell(x)$ with $\ell \in RV_0$ and $\rho < 0$.

The assumption that c is a constant function is equivalent to assuming that L is normalized (Kohlbecker, 1958) and ensures that L is differentiable. As noted in Bingham et al. (1987), the normalization assumption is not restrictive since slowly-varying functions are of interest only to within asymptotic equivalence. The condition $\varepsilon \in RV_\rho$ with $\rho < 0$ entails that $L(x) \to L_\infty \in (0,\infty)$ as $x \to \infty$. The index of regular variation ρ is referred to as the second-order parameter. It is the main driver of the bias in the estimation of extreme quantiles from heavy-tailed distributions, see Table 4 for values of ρ associated with usual distributions. Besides, $(\mathbf{H_2})$ entails that F_X satisfies the so-called second-order condition which is the cornerstone of all proofs of asymptotic normality in extreme-value statistics. Interpretations and examples may be found in Beirlant et al. (2004) and de Haan and Ferreira (2006). We also refer to Gardes and Girard (2010, 2012) where a similar assumption is introduced in the framework of conditional extremes. Similarly, we shall also consider the assumption:

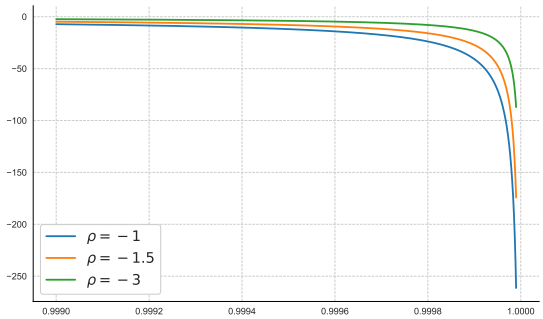
$(\mathbf{H_3})$: ℓ is normalized.



(a) $\gamma = 1/2$ and $\rho \in \{-3, -3/2, -1\}$



(b) $\rho = -1$ and $\gamma \in \{1/2, 1, 2\}$



(c) $\gamma = 1/2$ and $\rho \in \{-3, -3/2, -1\}$

Figure 3: (a,b): Tail-index function $u \in (0,1) \mapsto f^{\mathrm{TIF}}(u)$ associated with a Burr distribution for different values of tail-index γ and second-order parameter ρ, see Table 4 for parameterization details. (c): First derivative of Tail index function $u \in (0,1) \mapsto \partial_u f^{\mathrm{TIF}}(u)$.

The latter condition ensures that ℓ is differentiable on $(0,1)$ and thus that L and q_X are twice differentiable on $(0,1)$. Regularity properties of the above TIF can then be established, see Figure 3 for an illustration on the Burr distribution defined in Table 4.

Proposition 1

(i) If $(\mathbf{H_1})$ *holds, then* f^{TIF} *is a continuous and bounded function on* $[0,1]$, $f^{\mathrm{TIF}}(0) = 0$ *and* $f^{\mathrm{TIF}}(u) \to \gamma$ *as* $u \to 1$.

(ii) If, moreover, $(\mathbf{H_2})$ *holds, then* f^{TIF} *is continuously differentiable on* $(0,1)$ *and*

$$\partial_u f^{\mathrm{TIF}}(0) = \frac{\gamma + \varepsilon\,(1/\eta)}{\log(2)},$$

$$\partial_u f^{\mathrm{TIF}}(u) = \sum_{j=0}^{3} c_j \varphi_j(u) - \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)\log(1-u)}\,(1+o(1)) + \mathcal{O}\left(\frac{(1-u)}{\log(1-u)}\right), \qquad (11)$$

as $u \to 1$, *where* $c_0 = c_3 = \beta$, $c_1 = -\gamma/2$, $c_2 = (\gamma - \beta)/2$, $\beta = \gamma \log \eta - \log L_\infty$,

$$\varphi_0(u) = \frac{1}{(1-u)\left(\log(1-u)\right)^2} \text{ and } \varphi_j(u) = \frac{1}{\left(\log(1-u)\right)^j},\ u \in (0,1),\ j = 1,2,3.$$

It appears from Proposition $1(i)$ that, in contrast to the quantile function, the TIF is bounded on $[0,1]$. Remark that considering the simpler form $H_u(\cdot) = -\log(\cdot)/\log(1-u)$ in (3) would circumvent the problem as $u \to 1$ but would introduce an artifical singularity as $u \to 0$. Let us also highlight that $\varphi_0(u) \to \infty$ as $u \to 1$ making the first derivative of f^{TIF} unbounded as $u \to 1$, see Figure 3c for an illustration on the Burr distribution. Besides, $\varphi_j(u) \to 0$ as $u \to 1$ for all $j \in \{1,2,3\}$ while the second term in (11) tends to 0 if $\rho < -1$ or tends to ∞ if $\rho > -1$. Moreover, it is readily seen that $\partial_u \varphi_j(u) \to \infty$ as $u \to 1$ for all $j \in \{0,\dots,3\}$. As a conclusion, in the case where $\rho < -1$, Proposition $1(ii)$ suggests to build a twice differentiable version of f^{TIF} by removing the φ_j components, $j \in \{0,\dots,3\}$, in the neighborhood of $u = 1$. To this end, consider

$$f^{\mathrm{CTIF}}(u) := f^{\mathrm{TIF}}(u) - g(u) \sum_{j=0}^{3} c_j \Phi_j(u) - \gamma g(u) - \partial_u f^{\mathrm{TIF}}(0) h(u), \tag{12}$$

with, for all $u \in (0,1)$,

$$\begin{cases} g(u) = -4u^5 + 5u^4, \\ h(u) = u^3 - 2u^2 + u, \\ \Phi_0(u) = \varphi_1(u), \\ \Phi_1(u) = -\operatorname{li}(1-u), \\ \Phi_2(u) = \Phi_1(u) + (1-u)\varphi_1(u), \\ \Phi_3(u) = \Big(\Phi_1(u) + (1-u)(\varphi_1(u) + \varphi_2(u))\Big)/2. \end{cases} \tag{13}$$

Here, $\operatorname{li}(\cdot)$ denotes the logarithmic integral function defined as $\operatorname{li}(x) := \int_0^x \frac{1}{\log(t)}\,\mathrm{d}t$ for all $0 < x < 1$, with $\operatorname{li}(0) = 0$ and $\operatorname{li}(x) \to -\infty$ as $x \to 1$. Let us remark that $g(\cdot)$ and $h(\cdot)$ are two Hermite spline functions and that, by construction, $\partial_u \Phi_j(u) = \varphi_j(u)$, for all $j \in \{0,\dots,3\}$. The second term in (12) thus aims at removing the singular components in the first and second derivative of the TIF function in the neighborhood of $u = 1$. The additional terms $\gamma g(u)$ and $\partial f^{\mathrm{TIF}}(0) h(u)$ ensure that the TIF function as well as its first derivative vanish at $u = 0$. Regularity properties of f^{CTIF} are established in the next Proposition and illustrated on Figure 4 in the case of a Burr distribution.

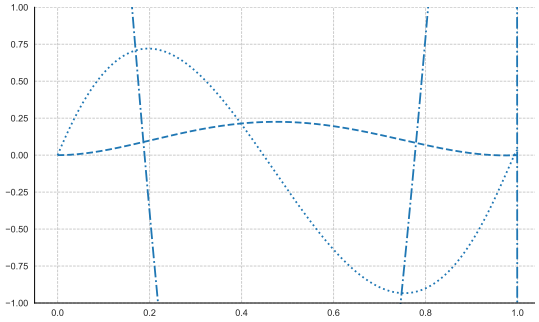
Proposition 2

(i) If $(\mathbf{H_1})$ *holds, then*

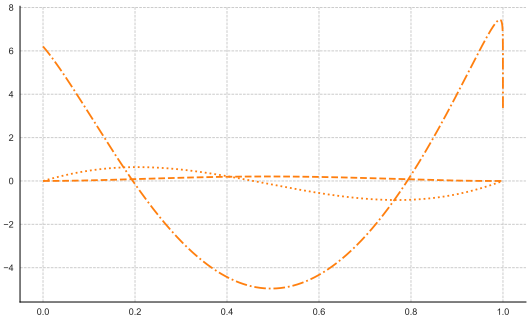
$$\lim_{u \to 0} f^{\mathrm{CTIF}}(u) = \lim_{u \to 1} f^{\mathrm{CTIF}}(u) = 0. \tag{14}$$

(ii) If, moreover, $(\mathbf{H_2})$ *holds with* $\rho < -1$, *then* f^{CTIF} *is continuously differentiable on* $[0,1]$ *and*

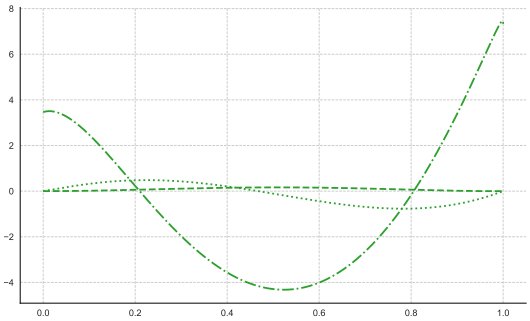
$$\lim_{u \to 0} \partial_u f^{\mathrm{CTIF}}(u) = \lim_{u \to 1} \partial_u f^{\mathrm{CTIF}}(u) = 0. \tag{15}$$



(a) CTIF with $\gamma = 1/2$ and $\rho = -1$



(b) CTIF with $\gamma = 1/2$ and $\rho = -3/2$



(c) CTIF with $\gamma = 1/2$ and $\rho = -3$

Figure 4: Illustration of the regularity properties of CTIF on a Burr distribution with $\gamma = 1/2$ and $\rho \in \{-3, -3/2, -1\}$. Corrected tail-index function $u \in (0,1) \mapsto f^{\mathrm{CTIF}}(u)$ (dashed line) and its first two derivatives $u \in (0,1) \mapsto \partial_u f^{\mathrm{CTIF}}(u)$ (dotted line) and $u \in (0,1) \mapsto \partial^2_{uu} f^{\mathrm{CTIF}}(u)$ (dash-dot line).

(iii) If, moreover, $(\mathbf{H_3})$ *holds, then* f^{CTIF} *is twice continuously differentiable on* $[0,1)$ *and*

$$\partial^2_{uu} f^{\mathrm{CTIF}}(u) = 20\gamma - 2\left(\frac{\gamma + \varepsilon(1/\eta)}{\log(2)}\right) - \frac{(1+\rho)\varepsilon\left(\frac{1}{\eta(1-u)}\right)}{(1-u)^2 \log(1-u)}\,(1+o(1)) + \mathcal{O}\left(\frac{1}{\log(1-u)}\right), \tag{16}$$

as $u \to 1$ *and*

$$\lim_{u\to 0} \partial^2_{uu} f^{\mathrm{CTIF}}(u) = \frac{5\gamma + \varepsilon(1/\eta)\left(5 + \rho + \frac{1}{\eta}\frac{\partial \ell(1/\eta)}{\ell(1/\eta)}\right)}{\log(2)} - 5\beta. \tag{17}$$

(iv) If, moreover, $\rho < -2$, *then* f^{CTIF} *is twice continuously differentiable on* $[0,1]$ *and*

$$\lim_{u\to 1} \partial^2_{uu} f^{\mathrm{CTIF}}(u) = 20\gamma - 2\left(\frac{\gamma + \varepsilon(1/\eta)}{\log(2)}\right). \tag{18}$$

It appears on Figure 4a that for $\rho = -1$, property (14) holds while first and second derivatives do not vanish at the boundaries of $[0, 1]$. When $\rho = -2$ (Figure 4b) both properties (14) and (15) are satisfied while the second derivative converges to a finite value in the neighborhood of 0, see (17), and diverges in the neighborhood of 1, see (16). Finally, $\rho = -3$ (Figure 4c) corresponds to the same situation, except that the second derivative also converges in the neighborhood of 1, see (18).

Let $I \subset \mathbb{R}$. Let us recall that a function $f : I \mapsto \mathbb{R}$ is Hölder continous with exponent $\alpha \in (0, 1]$ if the following quantity is finite

$$[f]_\alpha := \sup_{x \neq y \in I} \frac{|f(x) - f(y)|}{|x - y|^\alpha}.$$

This property is denoted for short by $f \in H^\alpha(I)$. The case $\alpha = 1$ corresponds to Lipschitz functions. We shall also note $C^m(I)$ the set of m-th continuously differentiable functions on I, $m \in \mathbb{N}$. Finally, for all $\alpha \in (0, 1]$ and $m \in \mathbb{N}$, we denote by $\mathcal{C}^{m,\alpha}(I)$ the Hölder space which consists of all functions $f \in C^m(I)$ such that $\partial^m f \in H^\alpha(I)$. In particular, $\mathcal{C}^{m+1}(I) \subseteq \mathcal{C}^{m,1}(I)$. Using these notations, and focusing on the case where $\rho < -1$, the regularity properties of f^{CTIF} provided by Proposition 2 can be simplified as:

Corollary 3 *Assume* $(\mathbf{H_1})$, $(\mathbf{H_2})$ *and* $(\mathbf{H_3})$ *hold.*

(i) If $-2 \leq \rho < -1$ *then* $f^{\mathrm{CTIF}} \in \mathcal{C}^{1,\alpha}([0, 1))$ *for all* $\alpha \in (0, -1 - \rho)$.

(ii) If $\rho < -2$ *then* $f^{\mathrm{CTIF}} \in C^2([0, 1])$.

It is thus clear that, the smaller ρ is, the more regular f^{CTIF} is, and therefore higher regularities could be obtained at the price of further restrictions on ρ. We are now in a position to investigate how a NN can approximate such a function.

2.2 Approximation error

Lemma 13 in Appendix B provides the minimum number of ReLU functions to approximate a $\mathcal{C}^{1,\alpha}$ function with a given precision ϵ. Combining this result with Corollary 3 yields the uniform approximation error of f^{CTIF} by a NN depending on the number of ReLU functions:

Theorem 4 *Assume* $(\mathbf{H_1})$, $(\mathbf{H_2})$ *and* $(\mathbf{H_3})$ *hold. Let* σ *be a ReLU function. For all* $J \geq 6$, *there exist* $(a_j, w_j, b_j) \in \mathbb{R}^3$, $j = 1, \ldots, J$ *such that:*

$$\sup_{u \in [0,1]} \left| f^{\mathrm{CTIF}}(u) - \sum_{j=1}^{J} a_j \sigma\left(w_j u + b_j\right) \right| \leq \frac{[\partial_t f^{\mathrm{CTIF}}]_\alpha}{4} \left\lceil \frac{J-3}{3} \right\rceil^{-\alpha-1} = \mathcal{O}\left(J^{-\alpha-1}\right),$$

where

1. $\alpha \in (0, -1 - \rho)$ *if* $-2 \leq \rho < -1$,
2. $\alpha = 1$ *if* $\rho < -2$.

Note that, for $\alpha = 1$, the above rate cannot be improved in general, owing to (Yarotsky, 2017, Theorem 6). Moreover, the previous approximation result can be interpreted in terms of Wasserstein-1 distance between the true data distribution and the simulated one. Indeed, in the univariate case, the Wasserstein-1 distance can be simplified as

$$W_1(q_Y, \widetilde{q}_Y) = \int_0^1 |q_Y(u) - \widetilde{q}_Y(u)| \, \mathrm{d}u,$$

where $u \mapsto \widetilde{q}_Y(u) := H_u^{-1}(G(u))$, with $H_u^{-1}(\cdot)$ defined in (5), is the EV-GAN approximation of the unknown quantile function $u \mapsto q_Y(u)$.

Corollary 5 *Assume conditions of Theorem 4 hold with $\gamma < 1$ and $\rho < -1$. Then, the Wasserstein-1 distance can be controlled as $W_1(q_Y, \widetilde{q}_Y) = \mathcal{O}\left(J^{-\alpha-1}\right)$.*

Note that $\gamma < 1$ is a necessary condition for the Wasserstein-1 distance to exist. In view of (12) and (13), letting

$$e_1(u) = g(u), \quad e_2(u) = h(u) \quad \text{and} \quad e_{k+3}(u) = g(u)\Phi_k(u) \ \text{ for } \ k = 0, \dots, 3 \tag{19}$$

in (4), the above approximation bounds on f^{CTIF} can be translated in terms of approximation bounds on f^{TIF} using the "enriched" NN. Note that the approximation space of TIF functions can be done for all components of a d-dimensional random variable, by following the principle (6), with a latent dimension $d' \geq d$. We obtain the final approximation result whose proof is now an easy combination of Corollary 3 and Theorem 4.

Corollary 6 *Let σ be a ReLU function. Let $X = (X^{(1)}, \dots, X^{(d)})^\top$ be a d-dimensional vector, with each component $X^{(m)}$ fulfilling $(\mathbf{H_1})$, $(\mathbf{H_2})$ and $(\mathbf{H_3})$ with parameters $(\gamma^{(m)}, \rho^{(m)})$. Let $\mathcal{G}_J^{d',d}$ be the approximation space of TIF functions made of $J \geq 6$ neurons:*

$$\mathcal{G}_J^{d',d} := \Bigg\{ G : z \in [0,1]^{d'} \mapsto G(z) = (G^{(1)}(z), \dots, G^{(d)}(z))^\top,$$
$$G^{(m)}(z) = \sum_{j=1}^{J} a_j^{(m)} \sigma\left(\sum_{i=1}^{d'} w_j^{(i)} z^{(i)} + b_j \right) + \sum_{k=1}^{6} \kappa_k^{(m)} e_k\left(z^{(m)}\right),$$
$$a_j^{(m)}, w_j^{(i)}, b_j, \kappa_k^{(m)} \in \mathbb{R} \Bigg\}.$$

Then,

$$\inf_{G \in \mathcal{G}_J^{d',d}} \sup_{m=1,\dots,d} \sup_{z \in [0,1]^{d'}} \left| f^{\mathrm{TIF},(m)}(z^{(m)}) - G^{(m)}(z) \right| = \mathcal{O}\left(J^{-\alpha-1}\right),$$

where

(i) $\alpha \in (0, -1 - \max_{m=1,\dots,d} \rho^{(m)})$ *if* $-2 \leq \rho^{(m)} < -1$ *for some* $m = 1, \dots, d$,

(ii) $\alpha = 1$ *if* $\rho^{(m)} < -2$ *for all* $m = 1, \dots, d$.

Here, we have defined

$$f^{\mathrm{TIF},(m)}(z^{(m)}) = -\frac{\log(q_{X^{(m)}}(1-(1-z^{(m)})\eta^{(m)}))}{\log\left(1-\left(z^{(m)}\right)^2\right)-\log 2}$$

as an natural extension of (9). For optimal parameters $a_j^{(m)}, w_j^{(i)}, b_j, \kappa_k^{(m)}$, the generative model for X is then

$$\tilde{X} = \left(H^{-1}_{Z^{(m)}}\left(G^{(m)}(Z)\right) : m = 1, \dots, d \right) \qquad \text{with} \quad Z \overset{\mathrm{d}}{=} \mathcal{U}([0,1]^{d'}). \tag{20}$$

In the above, one could restrict $G^{(m)}(z)$ to depend only on the m-th coordinate of z: it would not affect the potential quality of approximation of the m-th marginal of X but it would lead to a generative model with independent components which would be too restrictive. Mixing all latent components of z in $G^{(m)}(z)$ allows for generating dependence in the tails, while ensuring good fit of the marginals, as it will be checked in the subsequent experiments.

Observe that the worst second-order parameter $\rho^{(m)}$, *i.e.* the closest to -1, tunes the global accuracy of the EV-GAN through the convergence order $\alpha + 1$. Obtaining a similar result on the d-dimensional Wasserstein-1 distance is beyond the scope of this paper.

One may wonder if deeper ReLU NNs would help in better approximating the generative model for X. From the theoretical point of view, the benefit is unclear, in particular in view of (Yarotsky, 2017, Theorem 1) which states that a $\mathcal{C}^{1,1}$-function[1] can be approximated with error ϵ using a ReLU NN with depth $\mathcal{O}(\log(1/\epsilon))$ and number of weights $\mathcal{O}(\epsilon^{-1/2}\log(1/\epsilon))$. Up to the log factor, this is similar to the above result (Theorem 4) by setting $\epsilon = J^{-\alpha-1}$. From the numerical point of view, identifying in which circumstances a deep ReLU NN could be useful is part of our further investigations. Let us highlight that Lemma 13, and thus the whole analysis, can be adapted to any other non-polynomial activation function in view of the Universal approximation theorem (Pinkus, 1999).

3. Implementation

3.1 Experimental design

The neural network training is done by alternating generator and discriminator steps. The ranges of hyperparameters that are explored in order to find the best model for each data configuration are reported in Table 5. Note that, in order to respect the architecture (6), the generator is restricted to be a one hidden layer NN. Additionally, we use the optimizer Adam (Kingma and Ba, 2014) with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all tests performed during $1,000$ iterations. No additional normalization techniques are used. Every 5 iterations, two metrics (see Section 3.2 below) are computed and, for each metric, the NN parameters associated with the best results among the 200 checkpoints are selected.

3.2 Performance assessment

Recall that from (7), in the heavy-tail model, for all $j \in \{1, \dots, d\}$, $\log q_{X^{(j)}}(u)$ is approximately proportional to $\log(1/(1-u))$ when u is close to 1, with the tail-index γ as

1. His result does not apply to our possible case $\alpha \in (0, 1]$.

proportionality factor. It is therefore common practice to check the heavy-tail assumption on each margin $j \in \{1, \dots, d\}$ by drawing a log quantile-quantile plot, namely the points $(\log((n+1)/i), \log X^{(j)}_{n-i+1,n})$, for $i \in \{1, \dots, \lceil (1-\xi)n \rceil\}$, where $\xi \in [0, 1)$ is a given probability level. The performance of a generator can then be visually assessed by comparing the pairs $(\log((n+1)/i), \log X^{(j)}_{n-i+1,n})$ and $(\log((n+1)/i), \log \tilde{X}^{(j)}_{n-i+1,n})$. Here, and in the sequel, $\{\tilde{X}_1, \dots, \tilde{X}_n\}$ denotes the outputs generated either by the EV-GAN model (20) or by the classic GAN. To further quantify the fit on the tails of the marginal distributions, we define the Mean squared logarithmic error (MSLE) as the squared distance between the logarithm of the original and generated data:

$$\mathrm{MSLE}(\xi) = \frac{1}{d\lceil (1-\xi)n \rceil} \sum_{j=1}^{d} \sum_{i=1}^{\lceil (1-\xi)n \rceil} \left(\log(X^{(j)}_{n-i+1,n}) - \log(\tilde{X}^{(j)}_{n-i+1,n}) \right)^2,$$

with $\xi \in \{0.90, 0.95, 0.99\}$. Thus, a 100% relative error on the marginals corresponds to $\mathrm{MSLE}(\xi) = (\log(2))^2 \simeq 0.48$. Considering the dependence structure, one may also graphically compare the estimated Kendall's dependence functions K (or equivalently the $t \mapsto \lambda(t) := t - K(t)$ functions) on the n observations associated with the original sample and the generated one. From the quantitative point of view, the fit of the dependence structure is assessed by the 1-Wasserstein distance between these two Kendall's dependence functions which indeed are cumulative distribution functions. The distance can be computed as a L^1 norm referred to as the Absolute Kendall error (AKE) in the sequel:

$$\mathrm{AKE} = \frac{1}{n} \sum_{i=1}^{n} \left| Z_{i,n} - \tilde{Z}_{i,n} \right|,$$

where $Z_{1,n} \leq \cdots \leq Z_{n,n}$ (resp. $\tilde{Z}_{1,n} \leq \cdots \leq \tilde{Z}_{n,n}$) are the order statistics associated with $\{Z_1, \dots, Z_n\}$ (resp. $\{\tilde{Z}_1, \dots, \tilde{Z}_n\}$) and the $\tilde{Z}_i$ are computed similarly to (21) in Appendix A on the generated sample. We shall also compare Kendall's tau estimated on the original sample $\hat{\tau}_n$, on the generated sample $\tilde{\tau}_n$ and the theoretical value $\tau_{C^G_\mu}$.

3.3 Computational aspects

The numerical experiments presented in the next two sections have been conducted on the Cholesky computing cluster from Ecole Polytechnique `http://meso-ipp.gitlab.labos.polytechnique.fr/user_doc/`. It is composed by 2 nodes, where each one includes 2 CPU Intel Xeon Gold 6230 @ 2.1GHz, 20 cores and 4 Nvidia Tesla v100 graphics card. All the code was implemented in Python 3.8.2 and using the library PyTorch 1.7.1 for the GANs' training.

4. Validation on simulated data

The data simulation is based on the use of copulas, which allow to model separately the dependence structure and the margins, see Appendix A for a short overview. We focus on the Gumbel copula, denoted by C^G_μ which has been proved to be the only max-stable Archimedean copula (Genest and Rivest, 1989). The associated generating function is

$\psi_\mu^G(t) = \exp(-t^{1/\mu})$ defined for all $\mu \geq 1$ and $t \geq 0$. It is easily seen that Kendall's dependence function is given by $K_{C_\mu^G}(t) = t - t\log(t)/\mu$ for all $t \in (0,1]$ and Kendall's tau is $\tau_{C_\mu^G} = 1 - 1/\mu$. These two above quantities respectively provide a local and global characterization of the dependence structure induced by the copula. Besides, $C_1^G = \Pi$ the independence copula, and $C_\mu^G \to M$, the comotonic copula, as $\mu \to \infty$. In the following Section 4.1 we restrict ourselves to the dimension $d = 2$, while, in Section 4.2, we provide illustrations in higher dimensions.

4.1 Bivariate case

Three values of the dependence parameter are investigated: $\mu \in \{1.1, 2, 10\}$ leading to $\tau_{C_\mu^G} \in \{0.1, 0.5, 0.9\}$. The two margins are chosen to be Burr distributed, with common tail-index $\gamma := \gamma_1 = \gamma_2 \in \{0.1, 0.5, 0.9\}$ and second-order parameters $(\rho_1, \rho_2) \in \{(-1,-2), (-1,-3), (-2,-3)\}$, see Table 4 for the parametrization of the Burr distribution. Finally, $n = 10,000$ i.i.d data $\{X_1, \ldots, X_n\}$ are simulated from the resulting bivariate model for the above $3 \times 3 \times 3 = 27$ combinations of parameters. Results are reported on Table 1 in terms of MSLE(0.99), AKE and Kendall's tau.

When the tail-index γ increases, the tails of the marginal distributions of the simulated get heavier and the MSLE criteria of GAN and EV-GAN methods increase for all considered values of (ρ_1, ρ_2, μ) with a clear soaring when $\gamma = 1$. In this latter case, the expectation of the simulated distribution does not exist. However, from this marginal point of view, EV-GAN outperforms GAN in terms of MSLE for all considered configurations of $(\gamma, \rho_1, \rho_2, \mu)$. This conclusion remains true from the dependence point of view: EV-GAN outperforms GAN in terms of AKE for all the considered configurations of (ρ_1, ρ_2, μ) when $\gamma \in \{0.5, 0.9\}$. This phenomenon is illustrated in Figure 5 in the case where $\gamma = 0.9$, $\rho_1 = -1$, $\rho_2 = -3$ and $\mu = 10$. The log quantile-quantile plots associated with both margins are displayed on the top panel. It is easily seen that GAN method is not able to generate data in the distribution tail since the tail heaviness is strongly underestimated. At the opposite, EV-GAN method yields realistic data generation in both marginal tail distributions. Note that, in this case, the dependence structure is well captured by both NNs, see the estimated $\lambda(\cdot)$ functions on the bottom panel.

Finally, it appears on Table 1 that Kendall's tau is not a sufficient summary of the dependence structure: All estimated Kendall's tau are close to the theoretical ones even though the AKE is large. This criterion is thus dropped in the real data analysis hereafter since it might yield misleading conclusions.

4.2 Multivariate case

The ability of EV-GAN to properly scale in high dimension is now investigated. Using the `R` package `copulas` (Kojadinovic et al., 2010), $n = 10,000$ samples are simulated from a d-variate Gumbel copula for increasing dimensions $d \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, with a unique dependence parameter $\mu = 2$ and where all margins are Burr distributed with parameters $\gamma = 0.5$ and $\rho = -1$. MSLE results at level $\xi \in \{0.90, 0.95, 0.99\}$ are reported in Table 2 for both GAN and EV-GAN methods. Here again, EV-GAN clearly outperforms GAN for all dimensions and levels considered. Indeed, EV-GAN method yields

MSLE(0.99)

γ	(ρ_1, ρ_2) \ μ	1.1		2		10	
0.1	$(-1,-2)$	0.895	**0.116**	0.545	**0.097**	0.232	**0.037**
	$(-1,-3)$	0.923	**0.103**	0.732	**0.082**	0.553	**0.143**
	$(-2,-3)$	0.677	**0.190**	0.836	**0.083**	0.700	**0.174**
0.5	$(-1,-2)$	3.576	**1.058**	10.673	**1.006**	2.567	**1.321**
	$(-1,-3)$	1.943	**0.958**	6.913	**1.569**	3.812	**3.252**
	$(-2,-3)$	10.809	**1.707**	10.157	**1.201**	1.306	**1.195**
0.9	$(-1,-2)$	47.742	**4.966**	-	**6.473**	-	**8.651**
	$(-1,-3)$	44.949	**3.129**	-	**4.573**	45.900	**3.205**
	$(-2,-3)$	-	**3.860**	36.304	**6.390**	44.814	**5.922**

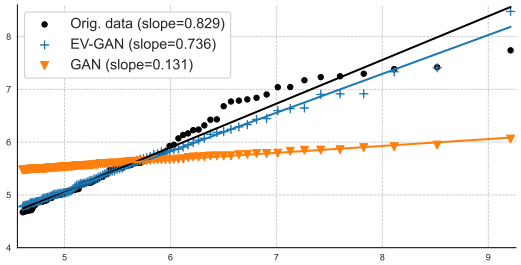
AKE

γ	(ρ_1, ρ_2) \ μ	1.1		2		10	
0.1	$(-1,-2)$	3.122	**2.865**	**7.385**	8.322	**2.807**	2.855
	$(-1,-3)$	3.102	**2.293**	**5.671**	6.958	**1.585**	2.415
	$(-2,-3)$	**2.519**	3.244	**4.596**	7.125	**1.823**	2.340
0.5	$(-1,-2)$	3.052	**2.234**	4.857	**1.772**	2.265	**2.015**
	$(-1,-3)$	6.261	**2.342**	4.538	**1.665**	2.616	**1.301**
	$(-2,-3)$	5.772	**2.134**	12.277	**1.408**	4.245	**1.531**
0.9	$(-1,-2)$	2.555	**2.103**	-	**1.990**	-	**1.932**
	$(-1,-3)$	3.788	**1.861**	-	**1.700**	1.623	**1.429**
	$(-2,-3)$	-	**1.696**	5.632	**1.788**	1.991	**1.181**

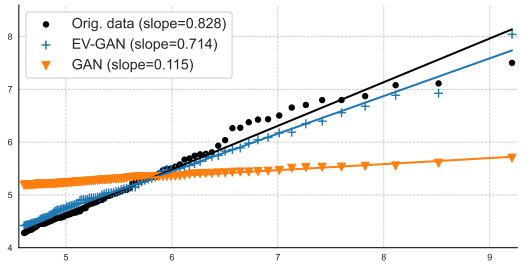
Kendall's tau

γ	(ρ_1, ρ_2) \ μ $(\tau_{C_\mu^\theta})$	1.1 (0.1)		2 (0.5)		10 (0.9)	
0.1	$(-1,-2)$	0.092	**0.091**	**0.514**	0.531	0.905	**0.895**
	$(-1,-3)$	**0.093**	0.083	0.477	**0.500**	**0.900**	0.905
	$(-2,-3)$	**0.086**	0.083	0.511	**0.480**	**0.899**	0.903
0.5	$(-1,-2)$	**0.090**	0.088	0.493	**0.500**	0.903	**0.900**
	$(-1,-3)$	0.106	**0.096**	0.506	**0.502**	0.901	**0.900**
	$(-2,-3)$	0.093	**0.087**	0.473	**0.502**	0.885	**0.898**
0.9	$(-1,-2)$	0.088	**0.090**	-	**0.503**	-	**0.901**
	$(-1,-3)$	**0.091**	0.088	-	**0.499**	**0.899**	0.897
	$(-2,-3)$	-	**0.089**	0.487	**0.498**	0.900	**0.900**

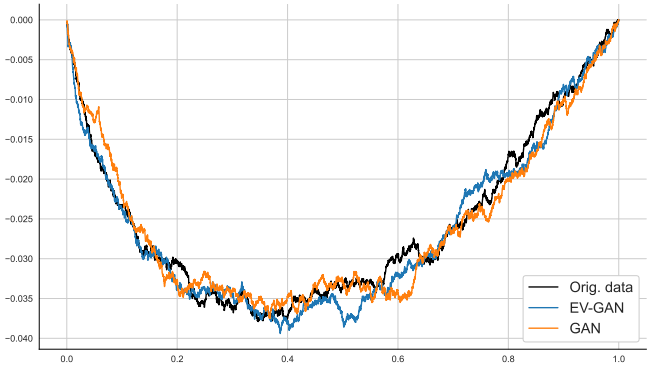
Table 1: Comparison between the best GAN (left column) and EV-GAN (right column) results on simulated data for the 27 combinations of parameters using two model selection criteria. Top: MSLE criterion at level $\xi = 0.99$, MSLE(ξ) ≥ 0.48 are not reported, all results are scaled by 10^2. Center: AKE criterion, the results are scaled by 10^3. Bottom: Kendall's tau (using the same models as the ones based on the AKE criterion). Best results are emphasized in bold.



(a) First margin ($\gamma = 0.9, \rho_1 = -1$)



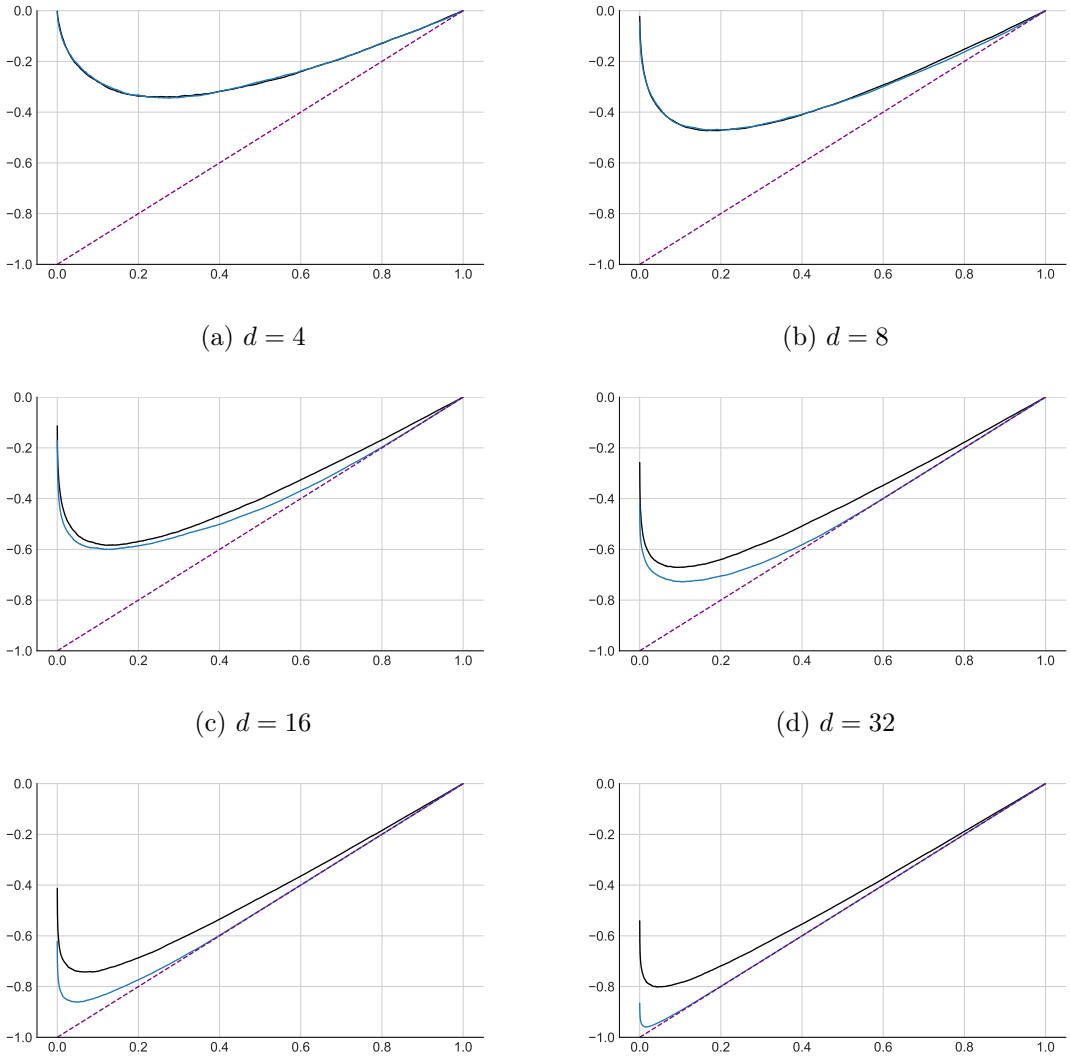
(b) Second margin ($\gamma = 0.9, \rho_2 = -3$)



(c) Estimated $\lambda(\cdot)$ functions

Figure 5: Top: log quantile-quantile plots on each margin $\log((n+1)/i) \mapsto \log X^{(j)}_{n-i+1,n}$, for $i \in \{1, \ldots, \lceil (1-\xi)n \rceil\}$ and $j \in \{1, 2\}$ on simulated data at probability level $\xi = 0.99$. The estimated regression lines are superimposed to each scatter plot. The associated slope is an estimation of the tail-index γ. Bottom: estimated $t \in [0, 1] \mapsto \lambda(t)$ functions. Black: original simulated data ($\gamma = 0.9$, $\rho_1 = -1$, $\rho_2 = -3$ and $\mu = 10$), blue: data generated with EV-GAN model, orange: data generated with classic GAN model.

realistic margins up to dimension 512 for high levels of quantiles $\xi \in \{0.90, 0.95\}$. In the case of higher levels ($\xi \in \{0.99\}$) the dimension is limited to 128. In contrast, the classic GAN model is limited more or less to dimension 8 for all levels. Figure 6 illustrates the dependence associated with samples in dimension $d \in \{4, 8, 16, 32, 64, 128\}$. First, remark that $\lambda(\cdot)$ associated with the original data tends toward the independence function $t \mapsto t - 1$ as d increases, accordingly to (Garcin et al., 2018, Section 3.3). Second, it appears that EV-GAN manages to reproduce very well the dependence structure of the original data up to $d = 16$, but tends faster to the independence between the margins for higher dimensions. Removing this trend is part of our future work, see Section 6.



(a) $d = 4$ (b) $d = 8$

(c) $d = 16$ (d) $d = 32$

(e) $d = 64$ (f) $d = 128$

Figure 6: Estimated $t \in [0,1] \mapsto \lambda(t)$ functions on multivariate simulated data in dimension $d \in \{4, 8, 16, 32, 64, 128\}$. Black: original simulated data (Burr distribution, $\gamma = 0.5$, $\rho = -1$ and $\mu = 2$), blue: data generated with EV-GAN model, dashed purple: independence case $t \mapsto \lambda_\Pi(t) = t - 1$.

dimension d	MSLE(0.90)		MSLE(0.95)		MSLE(0.99)		AKE	
4	3.134	**0.946**	5.627	**1.726**	16.961	**5.990**	14.	**2.**
8	11.399	**3.391**	17.613	**5.262**	-	**12.682**	32.	**5.**
16	47.632	**11.294**	-	**12.288**	-	**9.515**	-	**28.**
32	47.466	**12.519**	43.610	**10.052**	-	**32.022**	-	**49.**
64	-	**13.455**	-	**12.746**	-	**15.278**	-	**53.**
128	-	**19.365**	-	**14.751**	-	**28.977**	-	**48.**
256	-	**18.073**	-	**30.824**	-	-	-	-
512	-	**19.390**	-	**18.863**	-	-	-	-
1024	-	-	-	-	-	-	-	-

Table 2: Performance comparison between the best GAN (left column) and EV-GAN (right column) results on simulated d-variate data with respect to four model selection criteria. First three columns: MSLE(ξ) criterion computed at levels $\xi \in \{0.90, 0.95, 0.99\}$, MSLE($\xi$) ≥ 0.48 are not reported, all results are scaled by 10^2. Last column: AKE criterion, results are scaled by 10^3. Best results are emphasized in bold.

5. Illustration on real financial data

Our approach is tested on closing prices of daily financial stock market indices taken from `https://stooq.com/db/h/` on the October 1st, 2020. This database includes 61 world indices from their first day of quotation. Here, we selected six indices: NKX (Nikkei, Japan), KOSPI (Korea), HSI (Hong-Kong), CAC (France), AMX (Amsterdam Exchange, Netherlands), Nasdaq (USA) from three market zones: Asia, Europe, USA.

As a pre-processing step, the daily log-returns are computed for each ticker index. In case of missing data at a given business day, the next available day is removed from the dataset. Also, since we are interested in the modeling of synchronous indices, we kept only the data available at the same date for all selected tickers. Finally, positive returns were discarded since we focus on the generation of losses.

Figure 7 proposes a graphical summary of the tail and dependence properties associated with this dataset. First, the log quantile-quantile plots computed on all indices at level $\xi = 0.95$ are approximately linear which provides a graphical evidence of the tail heaviness of all six marginal distributions, with, for all indices, estimated slopes pointing towards a tail-index $\hat{\gamma} \simeq 0.3$ and an estimated second order parameter $\hat{\rho} \simeq -0.7$ using the estimator implemented in the `R` package `evt0` (Manjunath et al., 2013). Second, the $\lambda(\cdot)$ associated with all 15 pairs of indices are also displayed together with the two extreme cases $\lambda_\Pi(\cdot)$ and $\lambda_M(\cdot)$. The strongest dependence is found within the European market zone (pair AMX,CAC) while weakest dependencies are located between US and Asian market zones. Let us however note that the dependence between Asian, European and US markets may be under-estimated due to different time zones.

In the following, the performance of GAN and EV-GAN approaches are compared on four datasets of increasing dimensions: NKX ($d = 1$), Europe (AEX, CAC, $d = 2$), Asia (NKX, KOSPI, HSI, $d = 3$) and world (AEX, CAC, NKX, KOSPI, HSI, NDQ, $d = 6$). The training procedure described in Section 4 is adopted and results are reported in Table 3. EV-GAN outperforms GAN both on tail criteria MSLE(ξ), $\xi \in \{0.90, 0.95, 0.99\}$ and on

dependence criterion AKE, even though the condition $\rho < -1$ may not be fulfilled on this dataset. These results are illustrated on Figure 8 where it appears that EV-GAN is able to generate financial indices with realistic marginal tail behaviors. Finally, Figure 9 provides a comparison of dependence results obtained either using the MSLE or the AKE criteria. Unsurprisingly, the latter yields better results. Here again, the results associated with EV-GAN are visually more satisfying than those of the classic GAN. Information on the selected hyperparameters is provided in Table 5.

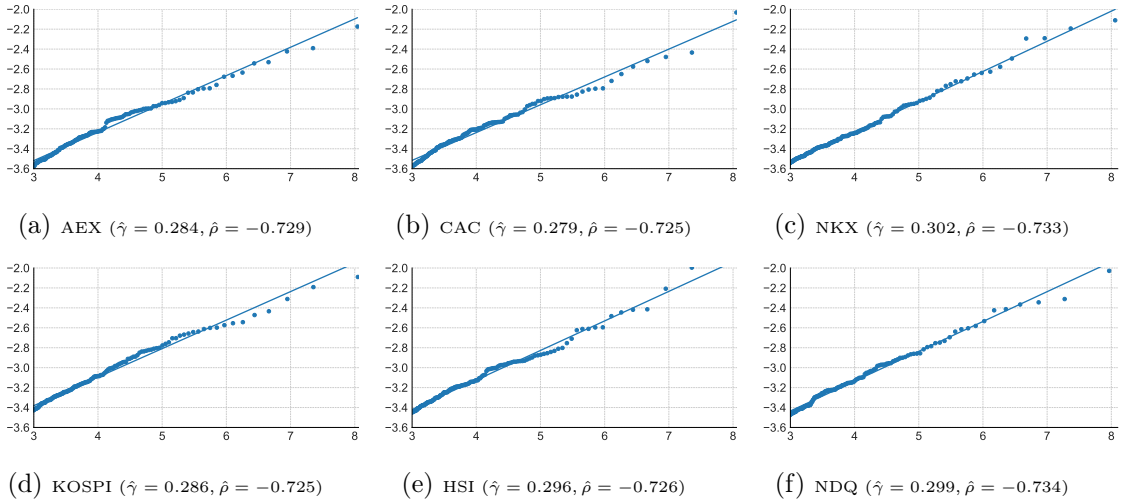
ticker	NKX		Europe		Asia		World	
dimension d	1		2		3		6	
sample size n	3173		2504		1378		548	
MSLE(0.90)	0.473	**0.133**	3.860	**0.132**	2.353	**0.677**	3.306	**0.874**
MSLE(0.95)	0.742	**0.103**	4.925	**0.178**	1.481	**0.579**	4.467	**1.219**
MSLE(0.99)	1.381	**0.200**	2.792	**0.320**	1.023	**0.538**	5.000	**1.960**
AKE	–	–	16.807	**4.697**	9.760	**4.872**	24.781	**3.533**

Table 3: Performance comparison between the best GAN (left column) and the EV-GAN (right column) results on real data with respect to four model selection criteria: using the MSLE(ξ) criterion computed at levels $\xi \in \{0.90, 0.95, 0.99\}$ and the AKE criteria (results are multiplied by 10^3 for the sake of readability).

6. Conclusion

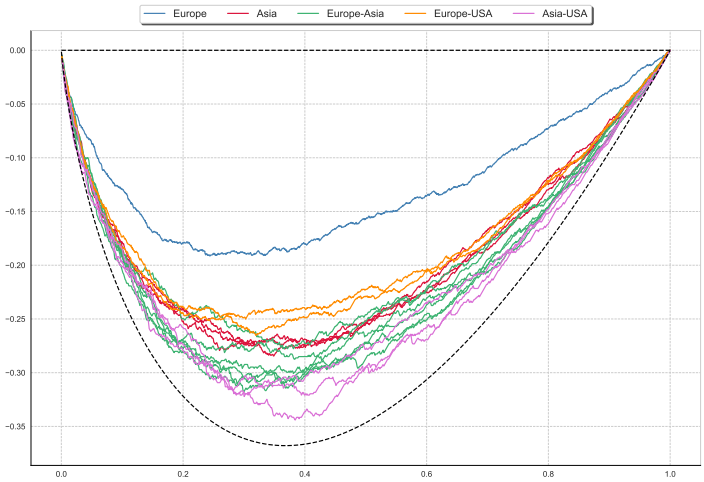
In this work, we have introduced a new generative method called EV-GAN dedicated to tail events. It relies on a new parametrization of GANs allowing to generate data coming from a heavy-tailed distribution. From the theoretical point of view, the uniform convergence rate of the proposed transformed quantile function f^{TIF} by a one hidden-layer ReLU NN is established within an extreme-value framework. From the practical point of view, we have illustrated on real and simulated data that EV-GAN outperforms classic GAN both in terms of tail behavior of the marginal distributions and in terms of dependence structure.

To complete the current theoretical analysis which ensures accurate approximation of marginals using NN, our further work will be dedicated to investigate mathematically how dependence structure is preserved, leveraging multivariate extreme-value theory. The analysis goes far beyond this work since it is known that dependence structure in the tails can be quite different from one case to another (Coles et al., 1999).
Finally, we shall investigate the behavior of the proposed EV-GAN corrections in other GAN architectures, using different distances and alternative criteria to MSLE and AKE.



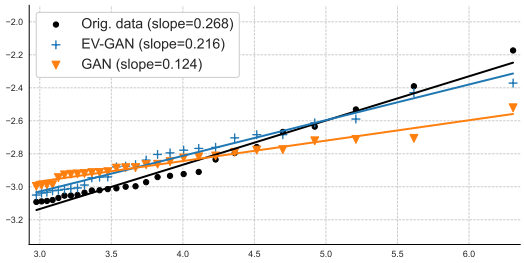
(a) AEX ($\hat{\gamma} = 0.284, \hat{\rho} = -0.729$) (b) CAC ($\hat{\gamma} = 0.279, \hat{\rho} = -0.725$) (c) NKX ($\hat{\gamma} = 0.302, \hat{\rho} = -0.733$)

(d) KOSPI ($\hat{\gamma} = 0.286, \hat{\rho} = -0.725$) (e) HSI ($\hat{\gamma} = 0.296, \hat{\rho} = -0.726$) (f) NDQ ($\hat{\gamma} = 0.299, \hat{\rho} = -0.734$)

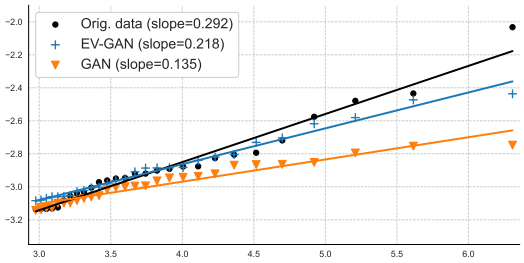


(g) Estimated $\lambda(\cdot)$ functions

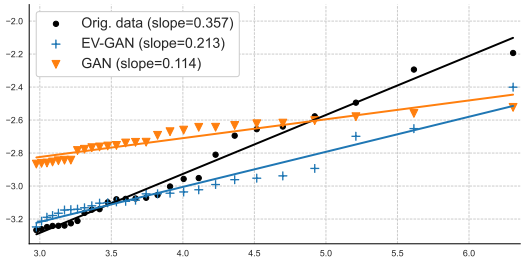
Figure 7: Top panels: Log quantile-quantile plots $\log((n+1)/i) \mapsto \log X^{(j)}_{n-i+1,n}$, for $i \in \{1, \ldots, \lceil (1-\xi)n \rceil\}$ on the selected financial indices $j \in \{1, \ldots, 6\}$ at probability level $\xi = 0.95$. The estimated regression line is superimposed to each scatter plot. The associated slope is an estimation of the tail-index. Bottom panel: Estimated $t \in [0,1] \mapsto \lambda(t)$ functions for all 15 pairs of indices. Functions $t \in [0,1] \mapsto \lambda_\Pi(t) = t \log t$ and $t \in [0,1] \mapsto \lambda_M(t) = 0$ respectively associated with independence and comotonic dependence in the bivariate case ($d = 2$) are depicted by black dashed lines.



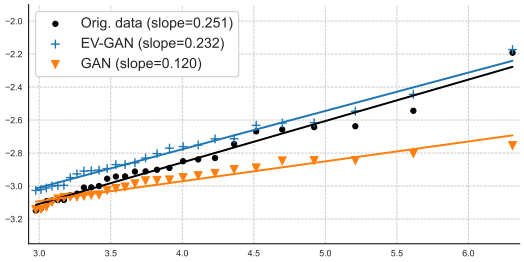
(a) AEX



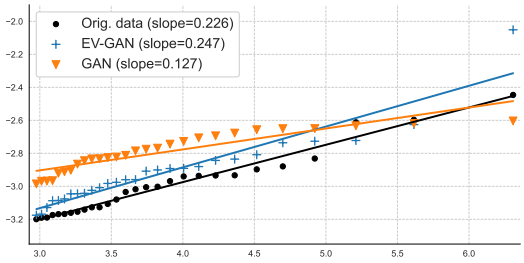
(b) CAC



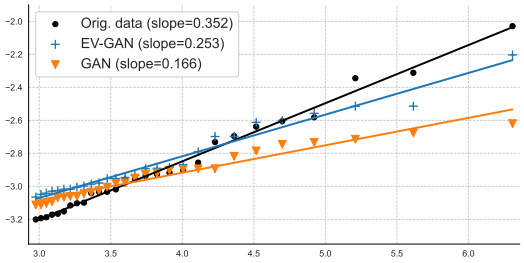
(c) NKX



(d) KOSPI

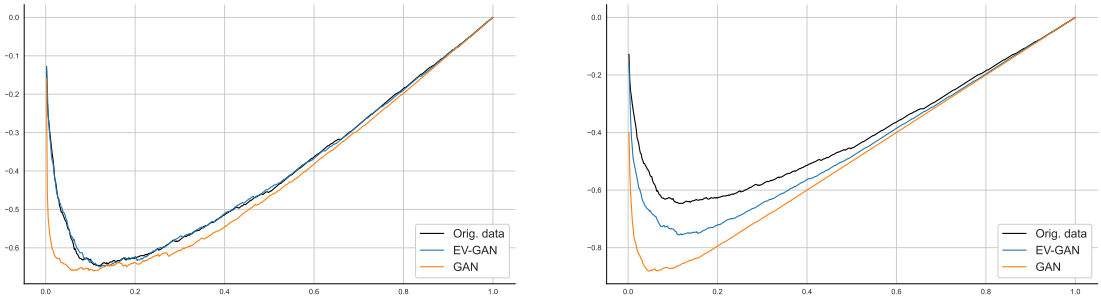


(e) HSI



(f) NDQ

Figure 8: Log quantile-quantile plots $\log((n+1)/i) \mapsto \log X^{(j)}_{n-i+1,n}$, for $i \in \{1, \ldots, \lceil (1-\xi)n \rceil\}$ and $j \in \{1, \ldots, 6\}$ associated with the world market zone at probability level $\xi = 0.95$.



(a) (b)

Figure 9: Estimated $t \in [0,1] \mapsto \lambda(t)$ functions associated with the World market zone ($d = 6$). Black: original real data, blue: data generated with EV-GAN model, orange: data generated with classic GAN model. (a) AKE criterion, (b) MSLE(0.95) criterion.

Distribution (parameters)	Density function	γ	ρ
Pareto ($\alpha > 0$)	$\alpha t^{-\alpha-1}$ $(t > 1)$	$1/\alpha$	$-\infty$
Burr ($\alpha, \beta > 0$)	$\alpha\beta t^{\alpha-1}\left(1+t^{\alpha}\right)^{-\beta-1}$ $(t > 0)$	$1/(\alpha\beta)$	$-1/\beta$
Fréchet ($\alpha > 0$)	$\alpha t^{-\alpha-1}\exp\left(-t^{-\alpha}\right)$ $(t > 0)$	$1/\alpha$	-1
Fisher ($\nu_1, \nu_2 > 0$)	$\frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2,\nu_2/2)} t^{\nu_1/2-1}(1+\nu_1 t/\nu_2)^{-(\nu_1+\nu_2)/2}$ $(t > 0)$	$2/\nu_2$	$-2/\nu_2$
Inverse-Gamma ($\alpha, \beta > 0$)	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{-\alpha-1}\exp(-\beta/t)$ $(t > 0)$	$1/\alpha$	$-1/\alpha$
Cauchy ($\sigma > 0$)	$\frac{\sigma}{\pi(\sigma^2+t^2)}$	1	-2
Student ($\nu > 0$)	$\frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$1/\nu$	$-2/\nu$

Table 4: A list of heavy-tailed distributions with the associated values of γ and ρ.

Hyperparameters ranges								
	latent dimension	batch size	neurons G.	learning rate G.	hidden layers D.	neurons D.	learning rate D.	training loop D.
A	$[10, 100]$	$[5-64]$	$[10, 500]$	$[0.0001, 0.01]$	$[1, 4]$	$[10, 500]$	$[0.0001, 0.01]$	$[1, 5]$
B	$[10, 2000]$	$[5-256]$	$[10, 500]$	$[0.0001, 0.01]$	$[1, 4]$	$[10, 500]$	$[0.0001, 0.01]$	$[1, 5]$
1. Selected hyperparameters on bivariate simulated data (setting A)								
MSLE(0.90)								
EV-GAN	31 (18)	32 (22)	47 (22)	0.0005 (0.0004)	2 (0)	28 (27)	0.0005 (0.0004)	1 (0)
GAN	30 (8)	28 (21)	43 (24)	0.0007 (0.0004)	2 (0)	29 (26)	0.0007 (0.0004)	1 (0)
MSLE(0.95)								
EV-GAN	36 (13)	31 (18)	68 (54)	0.0004 (0.0004)	2 (0)	45 (39)	0.0004 (0.0004)	1 (0)
GAN	30 (8)	26 (20)	44 (24)	0.001 (0.0004)	2 (0)	30 (26)	0.0007 (0.0004)	1 (0)
MSLE(0.99)								
EV-GAN	31 (13)	33 (22)	46 (22)	0.0005 (0.0005)	2 (0)	26 (27)	0.0005 (0.0005)	1 (0)
GAN	31 (8)	29 (21)	5426 (24)	0.0007 (0.0004)	3 (1)	27 (27)	0.0007 (0.0004)	1 (0)
AKE								
EV-GAN	40 (14)	30 (15)	90 (67)	0.0003 (0.0003)	2 (0)	63 (41)	0.0003 (0.0004)	1 (0)
GAN	36 (14)	24 (9)	70 (31)	0.002 (0.0003)	2 (0)	57 (42)	0.003 (0.0003)	1 (0)
2. Selected hyperparameters on multivariate simulated data (setting B)								
MSLE(0.90)								
EV-GAN	700 (523)	114 (88)	214 (172)	0.0007 (0.0004)	3 (1)	57 (31)	0.0007 (0.0004)	1 (1)
GAN	306 (412)	61 (37)	138 (152)	0.0006 (0.0005)	3 (1)	34 (22)	0.0006 (0.0005)	1 (0)
MSLE(0.95)								
EV-GAN	1065 (509)	171 (105)	167 (132)	0.0007 (0.0004)	2 (1)	75 (30)	0.0007 (0.0004)	2 (1)
GAN	306 (412)	61 (37)	138 (152)	0.0006 (0.0005)	3 (1)	34 (22)	0.0006 (0.0005)	1 (0)
MSLE(0.99)								
EV-GAN	1065 (509)	171 (105)	167 (132)	0.0007 (0.0004)	2 (1)	75 (30)	0.0007 (0.0004)	2 (1)
GAN	306 (412)	61 (37)	138 (152)	0.0006 (0.0005)	3 (1)	34 (22)	0.0006 (0.0005)	1 (0)
AKE								
EV-GAN	345 (394)	51 (33)	120 (65)	0.0003 (0.0004)	3 (1)	35 (12)	0.0003 (0.0004)	1 (0)
GAN	411 (464)	121 (107)	95 (50)	0.0003 (0.0004)	2 (0)	30 (13)	0.0003 (0.0004)	2 (1)
3. Selected hyperparameters on real data (setting A)								
MSLE(0.90)								
EV-GAN	21 (9)	9 (5)	18 (10)	0.0001 (0)	3 (0)	12 (2)	0.0006 (0.0005)	1 (0)
GAN	28 (5)	8 (0)	28 (5)	0.0001 (0.)	3 (0)	10 (0)	0.0008 (0.0005)	2 (1)
MSLE(0.95)								
EV-GAN	19 (9)	15 (13)	13 (5)	0.0003 (0.0005)	3 (1)	14 (4)	0.0006 (0.0005)	1 (0)
GAN	45 (37)	10 (4)	70 (87)	0.0001 (0.)	3 (0)	58 (95)	0.0006 (0.0005)	2 (1)
MSLE(0.99)								
EV-GAN	18 (10)	5 (0)	10 (0)	0.0001 (0.)	2 (1)	13 (2)	0.0001 (0.)	1 (0)
GAN	45 (37)	10 (4)	70 (87)	0.0001 (0.)	3 (1)	58 (95)	0.0006 (0.0005)	2 (1)
AKE								
EV-GAN	20 (12)	7 (2)	20 (12)	0.0003 (0.0005)	3 (0)	11 (2)	0.0006 (0.0005)	1 (0)
GAN	46 (36)	11 (4)	68 (89)	0.0001 (0.)	3 (1)	58 (94)	0.0006 (0.0005)	1 (0)

Table 5: Hyperparameters ranges used for tuning GANs across the experiments and mean (standard deviation) of selected hyperparameters in three situations: 1. simulated bivariate data, selection according to the MSLE(0.99) and AKE criteria, 2. simulated multivariate data, selection according to MSLE(ξ) for $\xi \in \{0.90, 0.95, 0.99\}$, 3. real data, selection according to the MSLE(0.95) and AKE criteria.

Acknowledgments

The authors would like to thank the three referees and the Associate Editor for their valuable suggestions, which have significantly improved the paper. This action benefited from the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole polytechnique and its foundation and sponsored by BNP Paribas. This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003).

Appendix A collects some statistical tools based on copulas used in experiments (Section 4 and Section 5). Appendix B provides auxiliary results used in Appendix C to prove the main results of Section 2.

Appendix A. Copulas

Let us consider a $d-$ variate distribution function F_X with continuous margins denoted by $F_X^{(j)}$, $j \in \{1, \dots, d\}$. From Sklar's Theorem (Sklar, 1959), there exists a unique function C such that

$$F_X\left(x^{(1)}, \dots, x^{(d)}\right) = C\left(F_X^{(1)}(x^{(1)}), \dots, F_X^{(d)}(x^{(d)})\right),$$

with $\left(x^{(1)}, \dots, x^{(d)}\right) \in \mathbb{R}^d$. The function C is called the copula of F_X. Introducing the uniform random variables $U^{(j)} = F^{(j)}(X^{(j)})$ for all $j \in \{1, \dots, d\}$, the copula C is the $d-$ dimensional distribution function of the random vector $\left(U^{(1)}, \dots, U^{(d)}\right)$ with uniform margins on $[0, 1]$. Copulas are a flexible tool to impose a given dependence structure on the marginal distributions of interest, see (Nelsen, 2006) for a detailed account on copulas. The independence between margins corresponds to the product copula $\Pi(u^{(1)}, \dots, u^{(d)}) = u^{(1)} \dots u^{(d)}$ while comotonic dependence corresponds to the Fréchet copula $M(u^{(1)}, \dots, u^{(d)}) = \min(u^{(1)}, \dots, u^{(d)})$.

Archimedean copulas. An Archimedean copula C_μ is defined for all $(u^{(1)}, \dots, u^{(d)}) \in [0, 1]^d$ by

$$C_\mu\left(u^{(1)}, \dots, u^{(d)}\right) = \psi_\mu\left(\psi_\mu^{-1}(u^{(1)}) + \dots + \psi_\mu^{-1}(u^{(d)})\right),$$

where $\psi_\mu : [0, \infty) \to [0, 1]$ is a parametric function which has to verify certain properties listed for instance in (McNeil and Nešlehová, 2009).

Kendall's dependence function. Kendall's dependence function (Genest and Rivest, 1993) characterizes the dependence structure associated with a copula C and is the univariate cumulative distribution function defined by $K_C(t) = \mathbb{P}\left(C\left(U^{(1)}, \dots, U^{(d)}\right) \leq t\right)$ for all $t \in [0, 1]$. In the case of an Archimedean copula C_μ, it can be derived as (Garcin et al., 2018):

$$K_{C_\mu}(t) = t + \sum_{j=1}^{d-1} \frac{(-\psi_\mu^{-1}(t))^j}{j!} \psi_\mu^{(j)}(\psi_\mu^{-1}(t)),$$

and we shall thus consider $\lambda_{C_\mu}(t) := t - K_{C_\mu}(t)$. It is then easily seen that $\lambda_M(t) = 0$ and

$$\lambda_\Pi(t) = t \sum_{j=1}^{d-1} \frac{(-\log(t))^j}{j!}$$

for all $t \in (0, 1]$.

Kendall's tau (bivariate case). Kendall's tau (Kendall, 1938) is a measure of dependence between two random variables. Let us then assume $d = 2$ and let X and $\tilde{X}$ be two bivariate random vectors from F_X. Kendall's tau is defined as the probability of concordance minus the probability of discordance of $X = (X^{(1)}, X^{(2)})$ and $\tilde{X} = (\tilde{X}^{(1)}, \tilde{X}^{(2)})$. It can be shown (Nelsen, 2006, Theorem 5.1.3) that this quantity only depends on the copula C of F_X and is given by

$$\tau_C = 4\mathbb{E}\left[C(U^{(1)}, U^{(2)})\right] - 1 = 4\int_0^1 \int_0^1 C(u,v)\,\mathrm{d}C(u,v) - 1,$$

with $\tau_M = 1$ and $\tau_\Pi = 0$ as special cases. In case of an Archimedean copula C_μ, Kendall's tau and Kendall's dependence functions are linked (Genest and MacKay, 1986):

$$\tau_{C_\mu} = 1 + 4\int_0^1 \lambda_{C_\mu}(v)\,\mathrm{d}v,$$

meaning that τ_{C_μ} can be interpreted as a summary of the dependence information encoded in $\lambda_{C_\mu}(\cdot)$.

Sampling (bivariate case). Sampling a random pair (U, V) from a bivariate copula C can be achieved by first simulating independently $(U, W) \sim \mathcal{U}([0,1]^2)$ and then letting $V = C_u^{-1}(W)$ where C_u is the conditional copula defined by

$$C_u(v) = \mathbb{P}(V \leq v | U = u) = \partial_u C(u, v).$$

In the case of bivariate Archimedean copulas, the conditional copula and its inverse are given by (Bernard and Czado, 2015):

$$C_{\mu,u}(v) = \frac{\partial_u\left(\psi_\mu^{-1}\right)(u)}{\partial_u\left(\psi_\mu^{-1}\right)(C(u,v))},$$
$$C_{\mu,u}^{-1}(y) = \psi_\mu\left((\partial_u\psi_\mu)^{-1}\left(\frac{y}{\partial_u\left(\psi_\mu^{-1}\right)(u)}\right) - \psi_\mu^{-1}(u)\right).$$

We also refer to Wu et al. (2007) and Hofert (2008) for alternative methods based on Kendall's dependence function and Laplace transform respectively.

Inference. The estimation of Kendall's dependence function is based on the pseudo-observations $\{Z_1, \ldots, Z_n\}$ from the cumulative distribution function K and computed as

$$Z_i = \frac{1}{n-1}\sum_{j\neq i}^{n} \mathbb{1}\left\{X_j^{(1)} < X_i^{(1)}, \ldots, X_j^{(d)} < X_i^{(d)}\right\}, \tag{21}$$

for all $i \in \{1, \ldots, n\}$, see (Genest and Rivest, 1993). The estimator of K is computed using the associated empirical cumulative distribution function:

$$\hat{K}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{Z_i \leq t\right\},$$

and we set $\hat{\lambda}_n(t) = t - \hat{K}_n(t)$, for all $t \in [0,1]$. Similarly, Kendall's tau is estimated by

$$\hat{\tau}_n = \frac{4}{n}\sum_{i=1}^{n} Z_i - 1.$$

Appendix B. Auxiliary results

We begin with a constructive proof of a particular case of Kuratowski Theorem (Bertsekas and Shreve, 1978, Chapter 7)-(Villani, 2009, p.8).

Lemma 7 *Let X be a random variable on $\mathbb{R}^d$. There exists a measurable function $G : (0,1) \to \mathbb{R}^d$ such that $X \overset{\mathrm{d}}{=} G(U)$ with $U \sim \mathcal{U}([0,1])$.*

Proof Let $Q : \mathbb{R}^d \to (0,1)^d$ be the component-wise logistic bijective function defined as $Q^{(m)}(x) = 1/(1+\exp(-x^{(m)}))$ for all $m \in \{1,\dots,d\}$. Let us also consider a continuous surjection $S : [0,1] \to [0,1]^d$ associated with a Space filling curve (like Peano or Hilbert curves, see (Sagan, 2014)). Define the inverse function $S^{-1}(x) := \inf\{t \in [0,1] : S(t) = x\}$, for any $x \in [0,1]^d$: it is measurable and is such that $S(S^{-1}(x)) = x$ since S is continuous. Then, $k := S^{-1} \circ Q$ is a measurable function from $\mathbb{R}^d$ to $(0,1)$, $k^{-1} = Q^{-1} \circ S$ is measurable too and satisfies $k^{-1}(k(x)) = x$ for any x. Additionally, let $Y := k(X)$ be a random variable on $(0,1)$ with cumulative distribution function F_Y so that $F_Y^{-1}(U) \overset{\mathrm{d}}{=} Y$, set $G(u) = k^{-1}\left(F_Y^{-1}(u)\right)$, for all $u \in (0,1)$. Then, for any bounded test function $\varphi : (0,1) \to \mathbb{R}^d$ we get

$$\mathbb{E}\left[\varphi(G(U))\right] = \mathbb{E}\left[\varphi\left(k^{-1}\left(F_Y^{-1}(U)\right)\right)\right] = \mathbb{E}\left[\varphi\left(k^{-1}(Y)\right)\right] = \mathbb{E}\left[\varphi(X)\right],$$

which proves that $X \overset{\mathrm{d}}{=} G(U)$. ■

The following three lemmas provide asymptotic expansions that will reveal useful to establish the behavior of the TIF as well as its derivatives in the neighborhood of $u = 0$ and $u = 1$.

Lemma 8

(i) The following asymptotic expansions hold, as $u \to 1$:

$$\frac{1}{\log\left(\frac{1-u^2}{2}\right)} = \frac{1}{\log(1-u)} + \frac{1-u}{2\left(\log(1-u)\right)^2} + \mathcal{O}\left(\frac{(1-u)^2}{\left(\log(1-u)\right)^2}\right), \tag{22}$$

$$\partial_u\left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right] = \frac{1}{(1-u)\left(\log(1-u)\right)^2} - \frac{1}{2\left(\log(1-u)\right)^2} + \frac{1}{\left(\log(1-u)\right)^3} + \mathcal{O}\left(\frac{(1-u)}{\left(\log(1-u)\right)^2}\right), \tag{23}$$

$$\partial^2_{uu}\left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right] = \frac{1}{(1-u)^2\left(\log(1-u)\right)^2} + \frac{2}{(1-u)^2\left(\log(1-u)\right)^3} - \frac{1}{(1-u)\left(\log(1-u)\right)^3} + \frac{3}{(1-u)\left(\log(1-u)\right)^4} + \frac{1}{4\left(\log(1-u)\right)^2} + \mathcal{O}\left(\frac{1}{\left(\log(1-u)\right)^3}\right). \tag{24}$$

(ii) Assume $(\mathbf{H_1})$ and $(\mathbf{H_2})$ hold. Then,

$$q_Y(u) = \eta^{-\gamma}(1-u)^{-\gamma} L\left(\frac{1}{(1-u)\eta}\right), \tag{25}$$

$$\partial_u q_Y(u) = \eta^{-\gamma}(1-u)^{-(\gamma+1)} L\left(\frac{1}{(1-u)\eta}\right)\left(\gamma + \varepsilon\left(\frac{1}{(1-u)\eta}\right)\right), \tag{26}$$

$$\log q_Y(u) = -\gamma\log(1-u) - \beta + \frac{1}{\rho}\varepsilon\left(\frac{1}{(1-u)\eta}\right)\big(1+o(1)\big), \ \text{as } u\to 1, \tag{27}$$

$$\partial_u \log q_Y(u) = (1-u)^{-1}\left(\gamma + \varepsilon\left(\frac{1}{(1-u)\eta}\right)\right). \tag{28}$$

(iii) Assume $(\mathbf{H_1})$, $(\mathbf{H_2})$ *and* $(\mathbf{H_3})$ *hold. Then, as* $u\to 1$,

$$\partial^2_{uu} q_Y(u) = \eta^{-\gamma}(1-u)^{-(\gamma+2)} L\left(\frac{1}{(1-u)\eta}\right) \times\left[\gamma^2+\gamma+(1+2\gamma+\rho+o(1))\varepsilon\left(\frac{1}{(1-u)\eta}\right)\right], \tag{29}$$

$$\partial^2_{uu}\log q_Y(u) = (1-u)^{-2}\left(\gamma+\varepsilon\left(\frac{1}{(1-u)\eta}\right)(1+\rho+o(1))\right). \tag{30}$$

Proof *(i)* The proof of (22)–(24) is straightforward but requires tedious calculations which can be checked by a formal calculation software (using `sympy` in `Python` for instance, see below). Details are omitted here.

```
import sympy as spy
u = spy.symbols('u')

f = 1 / spy.log((1 - u ** 2) / 2)

# series as u->1
f.series(u, 1, 2, dir="-")

f_first = spy.diff(f, u)
f_first.series(u, 1, 1, dir="-")

f_second = spy.diff(f_first, u)
f_second.series(u, 1, 1, dir="-")
```

(ii) Under $(\mathbf{H_1})$, Equations (7), (8) and (10) entail

$$q_Y(u) = \eta^{-\gamma}(1-u)^{-\gamma} L\left(\frac{1}{(1-u)\eta}\right),$$

which proves (25) and moreover, owing to $(\mathbf{H_2})$,

$$\log q_Y(u) = -\gamma\log(1-u) + \log(c_\infty) - \gamma\log\eta + \int_1^{\frac{1}{(1-u)\eta}} \frac{\varepsilon(t)}{t}\,\mathrm{d}t. \tag{31}$$

By differentiating, we get

$$\partial_u q_Y(u) = q_Y(u)\times\partial_u(\log q_Y(u)) = \eta^{-\gamma}(1-u)^{-(\gamma+1)} L\left(\frac{1}{(1-u)\eta}\right)\left(\gamma+\varepsilon\left(\frac{1}{(1-u)\eta}\right)\right),$$

and (26) is proved. Now, $t \mapsto \varepsilon(t)/t$ is regularly varying with index $\rho - 1 < -1$ and thus, $\int_1^\infty \varepsilon(t)/t \, \mathrm{d}t$ is finite leading to:

$$\log L_\infty = \log c_\infty + \int_1^\infty \frac{\varepsilon(t)}{t} \, \mathrm{d}t.$$

Replacing in (31) yields:

$$\log q_Y(u) = -\gamma \log(1-u) - \beta - \int_{\frac{1}{(1-u)\eta}}^\infty \frac{\varepsilon(t)}{t} \, \mathrm{d}t.$$

Moreover, Karamata's theorem (de Haan and Ferreira, 2006, Equation (B.1.9)) states that

$$\int_x^\infty \frac{\varepsilon(t)}{t} \, \mathrm{d}t = -\frac{1}{\rho}\varepsilon(x)(1 + o(1)),$$

as $x \to \infty$ so that (27) is proved. Finally, (28) is a direct consequence of (25) and (26).

(iii) From (26), letting $U(u) = \eta^{-\gamma}(1-u)^{-(\gamma+1)} L\left(\frac{1}{(1-u)\eta}\right)$, one has

$$\partial^2_{uu} q_Y(u) = \partial_u \left[U(u) \left(\gamma + \varepsilon \left(\frac{1}{(1-u)\eta} \right) \right) \right]. \tag{32}$$

Using the form of L under $(\mathbf{H_2})$ and $x\frac{\partial_x L(x)}{L(x)} = \varepsilon(x)$, we obtain

$$\partial_u U(u) = \eta^{-\gamma}(1-u)^{-(\gamma+2)} L\left(\frac{1}{(1-u)\eta}\right) \left(\gamma + 1 + \varepsilon \left(\frac{1}{(1-u)\eta} \right) \right). \tag{33}$$

In addition, recalling that ε is differentiable under $(\mathbf{H_3})$ yields

$$\begin{aligned}
\partial_u \left[\varepsilon \left(\frac{1}{(1-u)\eta} \right) \right] &= \eta^{-\rho}(1-u)^{-(\rho+1)} \ell \left(\frac{1}{(1-u)\eta} \right) \left(\rho + \frac{1}{(1-u)\eta} \frac{\partial \ell \left(\frac{1}{(1-u)\eta} \right)}{\ell \left(\frac{1}{(1-u)\eta} \right)} \right) \\
&= \frac{1}{(1-u)} \varepsilon \left(\frac{1}{(1-u)\eta} \right) \left(\rho + \frac{1}{(1-u)\eta} \frac{\partial \ell \left(\frac{1}{(1-u)\eta} \right)}{\ell \left(\frac{1}{(1-u)\eta} \right)} \right) \qquad (34) \\
&= \frac{1}{(1-u)} \varepsilon \left(\frac{1}{(1-u)\eta} \right) (\rho + o(1)). \qquad (35)
\end{aligned}$$

Collecting (32), (33) and (35) entails

$$\begin{aligned}
\partial^2_{uu} q_Y(u) &= \eta^{-\gamma}(1-u)^{-(\gamma+2)} L\left(\frac{1}{(1-u)\eta}\right) \\
&\quad \times \left[\left(\gamma + \varepsilon \left(\frac{1}{(1-u)\eta} \right) \right)^2 + \gamma + (1 + \rho + o(1)) \varepsilon \left(\frac{1}{(1-u)\eta} \right) \right] \\
&= \eta^{-\gamma}(1-u)^{-(\gamma+2)} L\left(\frac{1}{(1-u)\eta}\right)
\end{aligned}$$

$$\times \left[\gamma^2 + \gamma + (1 + 2\gamma + \rho + o(1))\varepsilon\left(\frac{1}{(1-u)\eta}\right) \right]$$

which proves (29). Finally, (28) and (34) entail

$$\partial^2_{uu} \log q_Y(u) = (1-u)^{-2}\left(\gamma + \varepsilon\left(\frac{1}{(1-u)\eta}\right)\right) + (1-u)^{-2}\varepsilon\left(\frac{1}{(1-u)\eta}\right)\left(\rho + \frac{1}{(1-u)\eta}\frac{\partial \ell\left(\frac{1}{(1-u)\eta}\right)}{\ell\left(\frac{1}{(1-u)\eta}\right)}\right) \tag{36}$$

and (30) is proved owing to $(\mathbf{H_3})$. ■

Lemma 9 *Let* li *be the logarithmic integral function defined for all* $u \in (0,1)$ *as*

$$\mathrm{li}(u) = \int_0^u \frac{1}{\log(t)}\,\mathrm{d}t.$$

Then, for any $p > 0$, $u^p \,\mathrm{li}(1-u) \to 0$ *as* $u \to 0$.

Proof This stems from the convexity inequality $\log(1/t) \geq 1 - t$ for $t \in (0,1]$. ■

Lemma 10 *For all* $u \in (0,1)$, *let* $\Phi(u) = \sum_{j=0}^3 c_j \Phi_j(u)$. *One has:*

$$\partial^2_{uu}\left[g(u)\left(\gamma + \Phi(u)\right)\right] = -20\gamma + \frac{c_0}{(1-u)^2\left(\log(1-u)\right)^2} + \frac{2c_0}{(1-u)^2\left(\log(1-u)\right)^3} + \frac{c_1}{(1-u)\left(\log(1-u)\right)^2} + \frac{2c_2}{(1-u)\left(\log(1-u)\right)^3} + \frac{3c_3}{(1-u)\left(\log(1-u)\right)^4} + \mathcal{O}\left(\frac{1}{\log(1-u)}\right), \qquad \text{as } u \to 1, \tag{37}$$

$$\partial^2_{uu}\left[g(u)\left(\gamma + \Phi(u)\right)\right] \to 5\beta, \qquad \text{as } u \to 0. \tag{38}$$

Proof Differentiating Φ yields for all $u \in (0,1)$,

$$\partial_u \Phi(u) = \sum_{j=0}^3 c_j \varphi_j(u),$$

$$\partial^2_{uu}\Phi(u) = \frac{c_0}{(1-u)^2\left(\log(1-u)\right)^2} + \frac{2c_0}{(1-u)^2\left(\log(1-u)\right)^3} + \frac{c_1}{(1-u)\left(\log(1-u)\right)^2} + \frac{2c_2}{(1-u)\left(\log(1-u)\right)^3} + \frac{3c_3}{(1-u)\left(\log(1-u)\right)^4}.$$

Besides, for all $u \in (0,1)$,

$$\partial^2_{uu}\left[g(u)\left(\gamma + \Phi(u)\right)\right] = 20u^2(3-4u)\left(\gamma + \Phi(u)\right) + 40u^3(1-u)\,\partial_u\Phi(u)$$

$$+ u^4 (5 - 4u) \partial^2_{uu} \Phi(u). \tag{39}$$

Remarking that $\Phi(u) = \mathcal{O}(1/\log(1-u))$ and $(1-u)\partial_u \Phi(u) = \mathcal{O}\left(1/(\log(1-u))^2\right)$ as $u \to 1$ proves (37). Similarly, Lemma 9 entails that $\mathrm{li}(1-u) = \mathcal{O}(1/u)$ as $u \to 0$ and thus $\Phi(u) = c_3/(2u^2)(1+o(1))$, $\partial_u \Phi(u) = -c_3/u^3(1+o(1))$ and $\partial^2_{uu}\Phi(u) = 3c_3/u^4(1+o(1))$ as $u \to 0$. Replacing in (39) and taking the limit as $u \to 0$ gives (38). ■

The next Lemma provides a sufficient condition for a given function to belong to $\mathcal{C}^{0,\alpha}([0,1])$.

Lemma 11 *Let $g : [0,1] \to \mathbb{R}$ be a continuous function on $[0,1]$ and differentiable on $(0,1)$ such that $|\partial_u g(u)| \leq C u^{\alpha-1}$ for all $u \in (0,1)$ with $0 < \alpha \leq 1$ and $C > 0$. Then, $g \in \mathcal{C}^{0,\alpha}([0,1])$.*

Proof Let $0 \leq a < b \leq 1$, then

$$|g(b) - g(a)| \leq \int_a^b C x^{\alpha-1} \, \mathrm{d}x \leq \int_a^b C(x-a)^{\alpha-1} \, \mathrm{d}x = \frac{C}{\alpha}(b-a)^\alpha,$$

and the conclusion follows. ■

Our goal here is to study the uniform convergence rate of the approximation error of a $\mathcal{C}^{1,\alpha}([0,1])$ or $C^2([0,1])$ function f by a NN. To this end, consider a triangular function $\widehat{\sigma} : \mathbb{R} \to [-1,1]$ built using three translated ReLU functions $x \in \mathbb{R} \mapsto \sigma(x) := \max(0, x)$:

$$\widehat{\sigma}(t) := \sigma(t+1) - 2\sigma(t) + \sigma(t-1) = \begin{cases} 1, & \text{if } t = 0, \\ 1 + t, & \text{if } -1 < t < 0, \\ 1 - t, & \text{if } 0 < t < 1, \\ 0, & \text{otherwise.} \end{cases}$$

It is then possible to control the uniform error between the function f and its piecewise linear approximation based on triangular functions, depending on the regularity of f.

Lemma 12 *Let $\widehat{\sigma}$ be a triangular function and $f : [0,1] \to \mathbb{R}$. For all $M \in \mathbb{N}\backslash\{0\}$, let $\delta = 1/M$ and $t_j = j/M$ for $j = 0, \ldots, M$. If $f \in \mathcal{C}^{1,\alpha}([0,1])$ with $\alpha \in (0,1]$, then*

$$\sup_{t \in [0,1]} \left| f(t) - \sum_{j=0}^{M} f(t_j) \widehat{\sigma}\left(\frac{t - t_j}{\delta}\right) \right| \leq \frac{[\partial_t f]_\alpha}{4} M^{-\alpha-1}. \tag{40}$$

Proof Clearly,

$$\sup_{t \in [0,1]} \left| f(t) - \sum_{j=0}^{M} f(t_j) \widehat{\sigma}\left(\frac{t - t_j}{\delta}\right) \right| =: \max_{i=0,\ldots,M-1} \sup_{t \in [t_i, t_{i+1}]} |\Delta_i(t)|,$$

where

$$\Delta_i(t) := f(t) - \left(f(t_i) \left(\frac{t_{i+1} - t}{\delta}\right) + f(t_{i+1}) \left(\frac{t - t_i}{\delta}\right) \right).$$

Two first order Taylor expansions yield that there exist $t_i' \in (t_i, t)$ and $t_i'' \in (t, t_{i+1})$ such that

$$\begin{aligned}\Delta_i(t) &= f(t) - \left[\big(f(t) + \partial_t f(t_i')(t_i - t)\big)\left(\frac{t_{i+1}-t}{\delta}\right) + \big(f(t) + \partial_t f(t_i'')(t_{i+1} - t)\big)\left(\frac{t-t_i}{\delta}\right)\right] \\ &= \frac{(t_{i+1}-t)(t-t_i)}{\delta}\left(\partial_t f(t_i') - \partial_t f(t_i'')\right).\end{aligned}$$

Remarking $(t_{i+1} - t)(t - t_i)$ is maximum on $[t_i, t_{i+1}]$ at $t = (t_{i+1} + t_i)/2$ entails

$$|\Delta_i(t)| \le \frac{\delta}{4}\left|\partial_t f(t_i') - \partial_t f(t_i'')\right| \le \frac{\delta}{4}[\partial_t f]_\alpha (t_i'' - t_i')^\alpha \le \frac{1}{4}[\partial_t f]_\alpha \delta^{\alpha+1},$$

and the result is proved. ■

Finally, one can determine the minimum number $J(\epsilon)$ of ReLU functions to approximate f with a given precision ϵ. The above construction in Lemma 12 involves $(M+1)$ triangular functions corresponding to $J = 3(M+1)$ ReLU functions. Fixing bound (40) to ϵ provides M as a function of ϵ, and we obtain:

Lemma 13 *Let σ be a ReLU function and $f \in \mathcal{C}^{1,\alpha}([0,1])$ with $\alpha \in (0,1]$. For all $\epsilon > 0$, let $J(\epsilon) = 3(M(\epsilon)+1)$ with $M(\epsilon) \in \mathbb{N}$ such that*

$$M(\epsilon) \ge \left(\frac{[\partial_t f]_\alpha}{4\epsilon}\right)^{1/(\alpha+1)}.$$

Then, there exist $(a_j, w_j, b_j) \in \mathbb{R}^3$, $j = 1, \dots, J(\epsilon)$ such that

$$\sup_{t\in[0,1]}\left|f(t) - \sum_{j=1}^{J(\epsilon)} a_j \sigma\left(w_j t + b_j\right)\right| \le \epsilon.$$

The above lemma is not that surprising, a similar result is stated in (Yarotsky, 2017, Theorem 1) up to a log factor but under the condition that $1 + \alpha$ is an integer.

Appendix C. Proof of main results

Proof of Proposition 1. (i) The continuity of f^{TIF} on $(0,1)$ is a consequence of the assumptions on F_X. Besides, $q_X(1-\eta) = 1$ and thus

$$f^{\mathrm{TIF}}(0) = \log(q_X(1-\eta))/\log 2 = 0.$$

From $(\mathbf{H_1})$, the cumulative distribution function F_X has an unbounded right-hand support, and thus, from (8), $q_Y(u) \to \infty$ as $u \to 1$. Thus, replacing in (7) and taking the log yields

$$\log q_Y(u) = -\gamma \log\big((1-u)\eta\big)\left(1 - \frac{\log L\left(\frac{1}{(1-u)\eta}\right)}{\gamma \log\left((1-u)\eta\right)}\right).$$

Since L is slowly varying, $\log L(v)/\log v \to 0$ as $v \to \infty$ (Bingham et al., 1987, Proposition 1.3.6) and then,

$$\log q_Y(u) = -\gamma \log(1-u)(1+o(1)), \text{ as } u \to 1.$$

Similarly, as $u \to 1$, $\log\left(\frac{1-u^2}{2}\right) = \log(1-u)(1+o(1))$, which leads to $f^{\text{TIF}}(u) \to \gamma$ as $u \to 1$. Finally, f^{TIF} is bounded on $[0,1]$ and the conclusion follows.
(ii) First, $q_Y(0) = 1$ directly yields

$$\partial_u f^{\text{TIF}}(0) = \frac{\gamma + \varepsilon\left(1/\eta\right)}{\log(2)}.$$

Second, collecting (22) and (28), it follows, as $u \to 1$,

$$\begin{aligned}\frac{\partial_u \log q_Y(u)}{\log\left(\frac{1-u^2}{2}\right)} &= \frac{\gamma}{(1-u)\log(1-u)} + \frac{\gamma}{2\left(\log(1-u)\right)^2} + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)\log(1-u)} + \mathcal{O}\left(\frac{(1-u)}{\left(\log(1-u)\right)^2}\right) \\ &= \frac{\gamma}{(1-u)\log(1-u)} + \frac{\gamma}{2}\varphi_2(u) + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)\log(1-u)} + \mathcal{O}\left(\frac{(1-u)}{\left(\log(1-u)\right)^2}\right).\end{aligned}$$

In addition, from (23) and (27), we have, as $u \to 1$,

$$\begin{aligned}\log q_Y(u)\partial_u\left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right] &= \frac{-\gamma}{(1-u)\log(1-u)} - \frac{\beta}{(1-u)\left(\log(1-u)\right)^2} + \frac{\gamma}{2\log(1-u)} \\ &\quad - \frac{(2\gamma-\beta)}{2\left(\log(1-u)\right)^2} - \frac{\beta}{\left(\log(1-u)\right)^3} + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)(1+o(1))}{\rho(1-u)\left(\log(1-u)\right)^2} \\ &\quad + \mathcal{O}\left(\frac{(1-u)}{\log(1-u)}\right) \\ &= \frac{-\gamma}{(1-u)\log(1-u)} - \beta\varphi_0(u) + \frac{\gamma}{2}\varphi_1(u) - \frac{(2\gamma-\beta)}{2}\varphi_2(u) + \beta\varphi_3(u) \\ &\quad + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)(1+o(1))}{\rho(1-u)\left(\log(1-u)\right)^2} + \mathcal{O}\left(\frac{(1-u)}{\log(1-u)}\right).\end{aligned}$$

Summing up the two above expansions and inverting the signs yield

$$\begin{aligned}\partial_u f^{\text{TIF}}(u) &= \beta\varphi_0(u) - \frac{\gamma}{2}\varphi_1(u) + \frac{\gamma-\beta}{2}\varphi_2(u) + \beta\varphi_3(u) \\ &\quad - \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)\log(1-u)}\left(1 + \frac{1}{\rho\log(1-u)}(1+o(1))\right) + \mathcal{O}\left(\frac{(1-u)}{\log(1-u)}\right),\end{aligned}$$

which proves the result. ■

Proof of Proposition 2. For all $u \in (0,1)$, let $\Phi(u) = \sum_{j=0}^{3} c_j \Phi_j(u)$.
(i) First, note that $\Phi(u) \to 0$ as $u \to 1$, $h(1) = 0$ and $g(1) = 1$. Besides, Proposition 1(i) shows that $f^{\mathrm{TIF}}(u) \to \gamma$ as $u \to 1$ and therefore $f^{\mathrm{CTIF}}(u) \to 0$ as $u \to 1$. Second, Lemma 9 entails that $\mathrm{li}(1-u) = \mathcal{O}(1/u)$ as $u \to 0$ and thus $\Phi(u) = c_3/(2u^2)(1+o(1))$. It follows that $g(u)\Phi(u) \to 0$ as $u \to 0$. Clearly, one also has $g(0) = h(0) = 0$. Besides, Proposition 1(i) shows that $f^{\mathrm{TIF}}(0) = 0$ and therefore $f^{\mathrm{CTIF}}(u) \to 0$ as $u \to 0$.
(ii) First, differentiating (12) and taking account of $g'(1) = h'(1) = 0$, $\Phi(u) \to 0$ as $u \to 1$ yields

$$\partial_u f^{\mathrm{CTIF}}(u) = \partial_u f^{\mathrm{TIF}}(u) - \partial_u \Phi(u) g(u) + o(1) = \partial_u \Phi(u)\,(1 - g(u)) + o(1),$$

as $u \to 1$, since $\partial_u f(u) = \partial_u \Phi(u) + o(1)$ when $\rho < -1$, in view of (11) in Proposition 1(ii). Remarking that $1 - g(u) = o(1-u)$ and recalling from the proof of Lemma 10 that $(1-u)\partial_u \Phi(u) = \mathcal{O}\left(1/(\log(1-u))^2\right)$ as $u \to 1$ prove that $\partial_u f^{\mathrm{CTIF}}(u) \to 0$ as $u \to 1$. Second, taking account of $g'(0) = 0$ and $h'(0) = 1$ yields

$$\partial_u f^{\mathrm{CTIF}}(u) = -g(u)\partial_u \Phi(u) - \Phi(u)\partial_u g(u) + o(1),$$

as $u \to 0$. Recall from the proof of Lemma 10 that $\Phi(u) = c_3/(2u^2)(1+o(1))$ and $\partial_u \Phi(u) = -c_3/u^3(1+o(1))$ as $u \to 0$. Since $g(u) = o(u^3)$ and $\partial_u g(u) = o(u^2)$ as $u \to 0$, it follows that $\partial_u f^{\mathrm{CTIF}}(u) \to 0$ as $u \to 0$ and (15) is proved.
(iii) The first part of the proof is based on successive applications of Lemma 8. From (22) and (30), one has, as $u \to 1$:

$$\partial^2_{uu}\left[\log q_Y(u)\right] \frac{1}{\log\left(\frac{1-u^2}{2}\right)} = \frac{\gamma}{(1-u)^2 \log(1-u)} + \frac{\gamma}{2(1-u)\left(\log(1-u)\right)^2} + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)^2\log(1-u)}(1+\rho+o(1)) + \mathcal{O}\left(\frac{1}{\left(\log(1-u)\right)^2}\right).$$

Similarly, from (23) and (28), as $u \to 1$,

$$\partial_u\left[\log q_Y(u)\right]\partial_u\left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right] = \frac{\gamma}{(1-u)^2\left(\log(1-u)\right)^2} - \frac{\gamma}{2(1-u)\left(\log(1-u)\right)^2} + \frac{\gamma}{(1-u)\left(\log(1-u)\right)^3} + \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)^2\left(\log(1-u)\right)^2} + \mathcal{O}\left(\frac{1}{\left(\log(1-u)\right)^2}\right),$$

and, from (24) and (27),

$$\log(q_Y(u))\partial^2_{uu}\left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right] = -\frac{\gamma}{(1-u)^2\log(1-u)} - \frac{2\gamma+\beta}{(1-u)^2\left(\log(1-u)\right)^2}$$

$$
\begin{aligned}
&- \frac{2\beta}{(1-u)^2\left(\log(1-u)\right)^3} + \frac{\gamma}{(1-u)\left(\log(1-u)\right)^2} \\
&- \frac{3\gamma-\beta}{(1-u)\left(\log(1-u)\right)^3} - \frac{3\beta}{(1-u)\left(\log(1-u)\right)^4} \\
&+ \frac{\varepsilon\left(\frac{1}{(1-u)\eta}\right)(1+o(1))}{\rho(1-u)^2(\log(1-u))^2} + \frac{2\varepsilon\left(\frac{1}{(1-u)\eta}\right)(1+o(1))}{\rho(1-u)^2(\log(1-u))^3} \\
&- \frac{\gamma}{4\log(1-u)} + \mathcal{O}\left(\frac{1}{\left(\log(1-u)\right)^2}\right).
\end{aligned}
$$

Collecting the above three asymptotic expansions yields, as $u \to 1$,

$$
\begin{aligned}
\partial^2_{uu} f^{\mathrm{TIF}}(u) = &\frac{\beta}{(1-u)^2\left(\log(1-u)\right)^2} + \frac{2\beta}{(1-u)^2\left(\log(1-u)\right)^3} - \frac{\gamma}{2(1-u)\left(\log(1-u)\right)^2} \\
&+ \frac{\gamma-\beta}{(1-u)\left(\log(1-u)\right)^3} + \frac{3\beta}{(1-u)\left(\log(1-u)\right)^4} + \frac{\gamma}{4\log(1-u)} \\
&- \frac{(1+\rho)\varepsilon\left(\frac{1}{(1-u)\eta}\right)}{(1-u)^2\log(1-u)}(1+o(1)) + \mathcal{O}\left(\frac{1}{\left(\log(1-u)\right)^2}\right). \qquad (41)
\end{aligned}
$$

In addition, note that $h''(1) = 2$ and $\partial_u f^{\mathrm{TIF}}(0) = (\gamma + \varepsilon(1/\eta))/\log(2)$ in view of Proposition 1(*ii*), so that collecting (37) in Lemma 10 with (41) proves (16). The second part of the proof consists in remarking that $\log q_Y(0) = 0$ by construction and $\partial_u \left[\frac{1}{\log\left(\frac{1-u^2}{2}\right)}\right](0) = 0$. Therefore, taking account of (36), it follows:

$$
\partial^2_{uu} f^{\mathrm{TIF}}(0) = \frac{\partial^2_{uu}\left[\log\left(q_Y(u)\right)\right](0)}{\log(2)} = \frac{\gamma + \varepsilon(1/\eta)\left(1 + \rho + \frac{1}{\eta}\frac{\partial \ell(1/\eta)}{\ell(1/\eta)}\right)}{\log(2)}. \qquad (42)
$$

Finally, note that $h''(0) = -4$ and $\partial_u f^{\mathrm{TIF}}(0) = (\gamma + \varepsilon(1/\eta))/\log(2)$ in view of Proposition 1(*ii*), so that collecting (38) in Lemma 10 with (42) proves (17).
(iv) is a direct consequence of *(iii)*. ■

Proof of Corollary 5. Theorem 4 yields, uniformly on $u \in [0,1]$:

$$
\left|\frac{\log q_Y(u) - \log \widetilde{q}_Y(u)}{\log((1-u^2)/2)}\right| \leq c(J),
$$

with $c(J) := \frac{[\partial_t f^{\mathrm{CTIF}}]_\alpha}{4}\left\lceil\frac{J-3}{3}\right\rceil^{-\alpha-1} \to 0$ as $J \to \infty$. It follows, for all $u \in [0,1]$:

$$
q_Y(u)\left(\frac{1-u^2}{2}\right)^{c(J)} \leq \widetilde{q}_Y(u) \leq q_Y(u)\left(\frac{1-u^2}{2}\right)^{-c(J)}.
$$

Substracting $q_Y(u)$ and integrating, we obtain $W_1(q, \widetilde{q}_Y) \leq \max\{D(-c(J)), -D(c(J))\}$ where

$$
D(t) := \int_0^1 q_Y(u)\left(\left(\frac{1-u^2}{2}\right)^t - 1\right)\mathrm{d}u \qquad (43)
$$

is defined for all (γ, t) such that $\gamma - t < 1$. Recall that $\gamma < 1$ and let us thus consider J large enough so that $\gamma + c(J) < 1$. Expanding (43) as $t \to 0$ yields

$$D(t) = t \int_0^1 q_Y(u) \log((1-u^2)/2)\, \mathrm{d}u\, (1 + o(1)),$$

and the conclusion follows. ■

Proof of Corollary 3. *(i)* When $-2 \leq \rho \leq -1$, Proposition 2(iii) implies $f^{\mathrm{CTIF}} \in C^2([0,1))$ and

$$\left|\partial^2_{uu} f^{\mathrm{CTIF}}(u)\right| \leq C(1-u)^{\alpha-1}, \quad \forall u \in (0,1),$$

for any fixed $\alpha \in (0, -\rho - 1)$. Thus, applying Lemma 11 to $\partial_u f^{\mathrm{CTIF}}$ yields $f^{\mathrm{CTIF}} \in \mathcal{C}^{1,\alpha}([0,1])$.
(ii) is a direct consequence of Proposition 2(iv). ■

References

M. Allouche, S. Girard, and E. Gobet. Generative model for fBm with deep ReLU neural networks. preprint, 2021. URL `https://hal.archives-ouvertes.fr/hal-03237854`.

J. Alm. Signs of dependence and heavy tails in non-life insurance data. *Scandinavian Actuarial Journal*, 2016(10):859–875, 2016.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

S. Asmussen and H. Albrecher. *Ruin probabilities*. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition, 2010.

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

C. Bernard and C. Czado. Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, 138:104–126, 2015.

D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, Inc., New York-London, 1978.

S. Bhatia, A. Jain, and B. Hooi. ExGAN: Adversarial generation of extreme samples. *arXiv preprint arXiv:2009.08454*, 2020.

G. Biau, B. Cadre, M. Sangnier, and U. Tanielian. Some theoretical properties of GANs. *The Annals of Statistics*, 48(3):1539–1566, 2020a.

G. Biau, M. Sangnier, and U. Tanielian. Some theoretical insights into Wasserstein GANs. *arXiv preprint arXiv:2006.02682*, 2020b.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987.

J. A. Bucklew. *Introduction to rare event simulation*. Springer Series in Statistics. Springer-Verlag, New York, 2004.

V. Chavez-Demoulin, P. Embrechts, and S. Sardy. Extreme-quantile tracking for financial time series. *Journal of Econometrics*, 181(1):44–52, 2014.

S. Coles, J. Heffernan, and J. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365, 1999.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006.

P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *The Annals of Applied Probability*, 15(4):2496–2534, 2005.

N. Dionelis, M. Yaghoobi, and S. A Tsaftaris. Tail of distribution GAN (TailGAN): Generative adversarial-network-based boundary formation. In *2020 Sensor Signal Processing for Defence Conference (SSPD)*, pages 1–5. IEEE, 2020.

R. Eckhardt. Stam Ulam, John Von Neumann and the Monte-Carlo method. *Los Alamos Science*, Special Issue:131–143, 1987.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997.

European Banking Authority. Guidelines on the revised common procedures and methodologies for the supervisory review and evaluation process (SREP) and supervisory stress testing. *Available at* *https://eba.europa.eu/regulation-and-policy/supervisory-review-and-evaluation-srep-and-pillar-2/guidelines-for-common-procedures-and-methodologies-for-the-supervisory-review-and-evaluation-process-srep-and-supervisory-stress-testing*, EBA/GL/2014/13, 2014.

R. M. Feder, P. Berger, and G. Stein. Nonlinear 3D cosmic web simulation with heavy-tailed generative adversarial networks. *Physical Review D*, 102(10):103504, 18, 2020.

D. Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media, 2019.

M. Garcin, D. Guegan, and B. Hassani. A novel multivariate risk measure: the Kendall VaR. Technical report, 2018. URL `https://halshs.archives-ouvertes.fr/halshs-01467857`.

L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204, 2010.

L. Gardes and S. Girard. Functional kernel estimators of large conditional quantiles. *Electronic Journal of Statistics*, 6:1715–1744, 2012.

C. Genest and J. MacKay. The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283, 1986.

C. Genest and L.-P. Rivest. A characterization of Gumbel's family of extreme value distributions. *Statistics & Probability Letters*, 8(3):207–211, 1989.

C. Genest and L.-P. Rivest. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043, 1993.

E. Gobet and G. Liu. Rare event simulation using reversible shaking transformations. *SIAM Journal on Scientific Computing*, 37(5):A2295–A2316, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

M. Haas and S. Richter. Statistical analysis of Wasserstein GANs with applications to time series forecasting. *arXiv preprint arXiv:2011.03074*, 2020.

M. Hofert. Sampling Archimedean copulas. *Computational Statistics & Data Analysis*, 52 (12):5163–5174, 2008.

T. Huster, J. EJ Cohen, Z. Lin, K. Chan, C. Kamhoua, N. Leslie, CY. J. Chiang, and V. Sekar. Pareto GAN: Extending the representational power of GANs to heavy-tailed distributions. *arXiv preprint arXiv:2101.09113*, 2021.

M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

E. Kohlbecker. Weak asymptotic properties of partitions. *Transactions of The American Mathematical Society*, 88(2):346–365, 1958.

I. Kojadinovic and J. Yan. Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20, 2010.

B. G. Manjunath and F. Caeiro. evt0: Mean of order p, peaks over random threshold, Hill and high quantile estimates. *R package version 1.1-3.* 2013.

A. McNeil and J. Nešlehová. Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097, 2009.

R. Nelsen. *An introduction to copulas.* Springer Series in Statistics. Springer, New York, second edition, 2006.

A. Pinkus. Approximation theory of the MLP model in neural networks. In *Acta numerica*, volume 8, pages 143–195. Cambridge Univ. Press, Cambridge, 1999.

M. Prandini and O.J. Watkins. Probabilistic aircraft conflict detection. *HYBRIDGE WP3: Reachability analysis for probabilistic hybrid systems*, 2005.

C. Remlinger, J. Mikael, and R. Elie. Conditional versus adversarial Euler-based generators for time series. *arXiv preprint arXiv:2102.05313*, 2021.

S. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.* Springer, 2007.

P. Robert. *Stochastic networks and queues*, volume 52 of *Applications of Mathematics.* Springer-Verlag, Berlin, 2003.

H. Sagan. *Space-filling curves.* Springer Science & Business Media, 2014.

M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, 2009.

M. Vladimirova, J. Arbel, and P. Mesejo. Bayesian neural networks become heavier-tailed with depth. In *NeurIPS 2018-Thirty-second Conference on Neural Information Processing Systems*, pages 1–7, 2018.

M. Wiese, R. Knobloch, and R. Korn. Copula & marginal flows: Disentangling the marginal from its joint. *arXiv preprint arXiv:1907.03361*, 2019.

M. Wiese, R. Knobloch, R. Korn, and P. Kretschmer. Quant GANs: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020.

F. Wu, E. Valdez, and M. Sherris. Simulating from exchangeable Archimedean copulas. *Communications in Statistics-Simulation and Computation*, 36(5):1019–1034, 2007.

D. Yarotsky. Error bounds for approximations with deep reLu networks. *Neural Networks*, 94:103–114, 2017.