# Copula-Based Normalizing Flows

Mike Laszkiewicz [1] Johannes Lederer [1] Asja Fischer [1]

## Abstract

Normalizing flows, which learn a distribution by transforming the data to samples from a Gaussian base distribution, have proven powerful density approximators. But their expressive power is limited by this choice of the base distribution. We, therefore, propose to generalize the base distribution to a more elaborate copula distribution to capture the properties of the target distribution more accurately. In a first empirical analysis, we demonstrate that this replacement can dramatically improve the vanilla normalizing flows in terms of flexibility, stability, and effectivity for heavy-tailed data. Our results suggest that the improvements are related to an increased local Lipschitz-stability of the learned flow.

## 1. Introduction

*Normalizing Flows* (NFs) are a recently developed class of density estimators, which aim to transform the distribution of interest $P_{\mathbf{x}}$ to some tractable base distribution $P_{\mathbf{z}}$. Using the change of variables formula, this allows for exact likelihood computation, which is in contrast to other deep generative models such as Variational Autoencoders (Kingma & Welling, 2014) or Generative Adversarial Networks (Goodfellow et al., 2014). Impressive estimation results, especially in the field of natural image generation, have lead to great popularity of these deep generative models. Motivated by this success, much effort has been put into the development of new parametric classes of NFs to make them even more performant (Dinh et al., 2015; 2016; Chen et al., 2019; Papamakarios et al., 2017; Grathwohl et al., 2019). However, our theoretical understanding has not developed at the same speed, which, so we claim, slows down further progress in the development of powerful NF architectures.

Fortunately, very recent works have addressed theoretical limitations of these methods: One important limitation is the expressive power of NFs. Because they are based on the change of variables formula, the learned transformations are required to be diffeomorphisms. As a consequence, a NF with bounded Lipschitz constant is unable to map one distribution $P_{\mathbf{x}}$ to a lighter-tailed distribution $P_{\mathbf{z}}$ (Wiese et al., 2019; Jaini et al., 2020). Therefore, since vanilla NFs are implemented using an isotropic Gaussian base distribution, they are unable to learn heavy-tailed distributions, which are known to appear in natural image data (Zhu et al., 2014; Horn & Perona, 2017; Zhang et al., 2017). This is in conflict with observations recently made by Behrmann et al. (2021): Even though NFs are typically designed to obey invertibility, this property is often violated in practice. This is due to numerical inaccuracies, which are promoted by a large Bi-Lipschitz constant. Bounding the Bi-Lipschitz constant, however, conflicts with the previously mentioned theoretical requirements needed to avoid a limited expressiveness of the NF.

These findings emphasize the high importance of choosing an appropriate base distribution for NFs. We therefore propose to generalize the isotropic Gaussian to a much broader class of distributions—the class of *copula* distributions. Copulae are a well-known concept in statistical theory and are being used to model complex distributions in finance, actuarial science, and extreme-value theory (Genest et al., 2009; Joe, 2014; Hofert et al., 2018). Broadly speaking, a copula is some function that couples marginal distributions to a multivariate joint distribution. Hence, copulae allow for flexible multivariate modeling with marginals stemming from a huge range of suitable classes. This allows for example to formulate NF base distributions that combine heavy-tailed marginals—as proposed by Wiese et al. (2019); Jaini et al. (2020); Alexanderson & Henter (2020)—with light-tailed marginals. This paper presents a novel framework for choosing the base distribution of NFs building on the well-studied theory of copulae. A first empirical investigation demonstrates the benefits brought by this approach. Our experimental analysis on toy data reveals that using even the most simple copula model—the Independence Copula—we are able to outperform the vanilla NF approach, which uses an isotropic Gaussian base distribution. The resulting NF converges faster, is more robust, and

[*] Equal contribution [1] Department of Mathematics, Ruhr University, Bochum, Germany. Correspondence to: Mike Laszkiewicz <Mike.Laszkiewicz@rub.de>.

Third workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2021). Copyright 2021 by the author(s).

achieves an overall better test performance. In addition, we show that the learned transformation has a better-behaved functional form in the sense of a more stable local Lipschitz continuity.

## 2. Background

In this section, we quickly review some background knowledge about NFs (Section 2.1), followed by an introduction to copula theory (Section 2.2).

### 2.1. Normalizing Flows

Density estimation via NFs revolve around learning a diffeomorphic transformation $T_\theta$ that maps some unknown target distribution $P_\mathbf{x}$ to a known and tractable base distribution $P_\mathbf{z}$. At the cornerstone of NFs is the *change of variables formula*

$$p_\theta(x) = p_\mathbf{z}\big(T_\theta(x)\big)\big|\det J_{T_\theta}(x)\big| \quad \text{for } x \in \mathbb{R}^D \quad , \qquad (1)$$

which relates the evaluation of the estimated density $p_\theta$ of $\mathbf{x} \sim P_\mathbf{x}$ to the evaluation of the base density $p_\mathbf{z}$, of $T_\theta(x)$, and of $\det J_{T_\theta}(x)$. By composing simple diffeomorphic building blocks $T_\theta := T_{\theta,l} \circ \cdots \circ T_{\theta,1}$, we are able to obtain expressive transformations, while presuming diffeomorphy and computational tractablity of the building blocks. Due to the tractable PDF in (1), we are able to train the model via maximum likelihood estimation (MLE)

$$\hat\theta \in \arg\min_\theta \mathbb{E}_{p_{\mathrm{data}}}\big[-\log p_\theta(\mathbf{x})\big] \quad , \qquad (2)$$

where $p_{\mathrm{data}}$ is the PDF of the empirical distribution of $\mathbf{x}$. A comprehensive overview of NFs, including the exact parameterizations of certain flow models $T_\theta$, computational aspects, and more, can be found in Kobyzev et al. (2020) and Papamakarios et al. (2021).

### 2.2. Copulae

A completely different approach of density estimation, which has mostly been left unrelated to NFs, is the idea of copulae.

**Definition 2.1** (Copula)**.** *A copula is a multivariate distribution with CDF $C : [0,1]^D \to [0,1]$ that has standard uniform marginals, i.e. the marginals $C_j$ of $C$ satisfy $C_j \sim U[0,1]$.*

The fundamental idea behind copula theory is that we can associate every distribution with a uniquely defined copula $C$. Vice versa, given $D$ marginal distributions, each copula $C$ defines a multivariate distribution with the given marginals. Formally, this is known as *Sklar's Theorem* (Sklar, 1959; 1996).

**Theorem 2.2** (Sklar's Theorem)**.** *Taken from Hofert et al. (2018).*

1. *For any $D$-dimensional CDF $F_\mathbf{z}$ with marginal CDFs $F_{\mathbf{z}_1}, \ldots, F_{\mathbf{z}_D}$, there exists a copula $C$ such that*

$$F_\mathbf{z}(z) = C\big(F_{\mathbf{z}_1}(z_1), \ldots, F_{\mathbf{z}_D}(z_D)\big) \qquad (3)$$

*for all $z \in \mathbb{R}^D$. The copula is uniquely defined on $\mathcal{U} := \prod_{j=1}^D \mathrm{Im}(F_{\mathbf{z}_j})$, where $\mathrm{Im}(F_{\mathbf{z}_j})$ is the image of $F_{\mathbf{z}_j}$. For all $u \in \mathcal{U}$ it is given by*

$$C(u) = F_\mathbf{z}\big(F^{\leftarrow}_{\mathbf{z}_1}(u_1), \ldots, F^{\leftarrow}_{\mathbf{z}_D}(u_D)\big) \quad , \qquad (4)$$

*where $F^{\leftarrow}_{\mathbf{z}_j}$ are the right-inverses of $F_{\mathbf{z}_j}$.*

2. *Conversely, given any $D$-dimensional copula $C$ and marginal CDFs $F_{\mathbf{z}_1}, \ldots F_{\mathbf{z}_D}$, a function $F_\mathbf{z}$ as defined in (3) is a $D$-dimensional CDF with marginals $F_{\mathbf{z}_1}, \ldots, F_{\mathbf{z}_D}$.*

Part 1 of Sklar's Theorem finds much application in statistical dependency analysis (Joe, 2014). In contrast to classical dependency measures, such as Pearson correlation, copulae are a more flexible tool that allow the decoupling of the marginals and the dependency structure. Part 2 of Sklar's Theorem is of relevance for statistical modeling, and more precisely, to define multivariate distributions. Given marginal distributions, which are typically much easier to estimate than the full joint distribution, and a copula $C$ we can "couple" the marginals and the dependency structure to a multivariate joint distribution. This perspective finds various applications in the context of finance and related disciplines that need to take heavy tails and tail dependencies into account, see Genest et al. (2009) for an overview. In Section A of the Appendix we give some illustrative examples and further details about properties of copula distributions.

By differentiating (3), we obtain the PDF of $\mathbf{z}$ as

$$p_\mathbf{z}(z) = c\big(F_{\mathbf{z}_1}(z_1), \ldots, F_{\mathbf{z}_D}(z_D)\big) \prod_{j=1}^D p_{\mathbf{z}_j}(z_j) \quad , \qquad (5)$$

where $c$, $p_{\mathbf{z}_1}, \ldots, p_{\mathbf{z}_D}$ are the PDFs of $C$, $F_{\mathbf{z}_1}, \ldots, F_{\mathbf{z}_D}$, respectively.

## 3. NFs With Copula-Base Distributions

In this paper, we propose to employ copulae to model a flexible, yet appropriate base distribution with the goal of gaining a NF that is able to solve the limitations of NFs discussed in Section 1. We expect to gain powerful and robust PDF approximators by combining different marginals and properties of theoretical sound copulae (see for instance Chapter 8 in Joe (2014)) with NFs, which allow for the estimation of complex densities.

### 3.1. A General Framework

We propose to replace the isotropic Gaussian base distribution in the vanilla NF framework by a more flexible copula distribution. Importantly, we want to learn a base distribution that is able to represent the tail behavior of the distribution of $\mathbf{x}$. For training a NF with a copula base distribution we build on the fact that we can write the PDF of the latent variables as written in (5). This requires two estimation steps: First, we need to estimate the marginal distributions $F_{\mathbf{z}_1}, \dots, F_{\mathbf{z}_D}$, which can further be used to calculate the marginal densities $p_{\mathbf{z}_1}, \dots, p_{\mathbf{z}_D}$. Secondly, we need to estimate the copula density $c$. A popular approach for estimating the density in (5) is to employ the method of inference functions for margins (IFM) (Joe & Xu, 1996), which sequentially estimates the marginals using MLE first, and then employs these marginals to estimate the copula using MLE.

It is important to note that in contrast to standard applications of copula theory, we do not aim at estimating the full data generating distribution based on (5). Instead, following the investigations by Jaini et al. (2020), our goal is to capture the tail-behavior of $\mathbf{x}$. Hence, we propose to learn surrogate marginals $\mathbf{z}_1, \dots, \mathbf{z}_D$ that are able to represent the tailedness of the marginals $\mathbf{x}_1, \dots, \mathbf{x}_D$. By combining these marginals with some simple copula structure, such as the Gaussian Copula or the *Independence Copula* (see (6) below), we are able to create a joint distribution that represents the marginal tail behavior of $\mathbf{x}$.

The proposed adjustment can be applied to any existing NF architecture as long as (5) remains tractable. However, as the main goal of the base distribution is not to fully estimate the target but to represent the tail behavior of $\mathbf{x}$, we can restrict ourselves to tractable parametric marginal distributions and copulae.

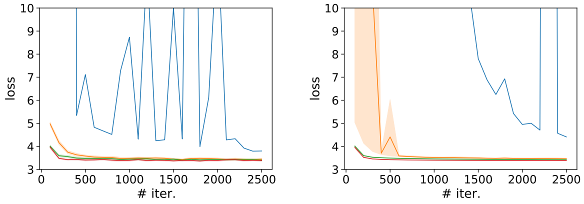
### 3.2. Experimental Analysis

In this section, we investigate the benefits of the proposed approach by analyzing a toy problem. In the following experiments, we employ the framework proposed in Section 3.1 using the most simple copula: the Independence Copula, i.e. we consider a base distribution $\mathbf{z}$ with PDF

$$p_{\mathbf{z}}(z) = \prod_{j=1}^{D} p_{\mathbf{z}_j}(z_j), \quad z \in \prod_{j=1}^{D} \operatorname{supp}(\mathbf{z}_j) \ . \tag{6}$$

Note that by plugging Gaussian marginals in (6), we would obtain the vanilla NF. We consider a training set generated from a 2-dimensional heavy-tailed distribution,[1] which has standardized t-distributed marginals $\mathbf{x}_1, \mathbf{x}_2 \sim t_2(0, 1)$ with 2 degrees of freedom. The corresponding copula is a *Gumbel Copula* with parameter $\rho = 2.5$, i.e.

$$C(u) := \exp\big(-((-\log(u_1))^{\rho} + (-log(u_2))^{\rho})^{1/\rho}\big) \ .$$



(a) Training loss (b) Test loss

*Figure 1.* Mean training and test loss over 100 trails for NFs with different base distributions: *normal* (blue), *heavierTails* (orange), *correctFamily* (green), and *exactMarginals* (red). The shaded area depicts the 95% confidence interval, which was computed using bootstrapping. We excluded *normal* runs that achieved a final loss larger than 25, which happened in 17 out of 100 runs.

As a proof of concept we compare the estimation of this heavy-tailed distribution using a NF[2] with an isotropic Gaussian base distribution, and with 3 different heavy-tailed base distributions constructed via (6). We consider the following heavy-tailed marginals:

1. Laplace$(0, 4)$ and $t_5(0, 2)$. We call this case *heavierTails* because one marginal is heavy-tailed;
2. $t_5(0, 1)$ and $t_5(0, 1)$. We call this case *correctFamily* since both marginals stem from the same parametric class as the exact marginals;
3. $t_2(0, 1)$ and $t_2(0, 1)$. We call this case *exactMarginals*.

Samples from the target distribution and from the different base distributions are visualized in Figure 6 and 7 in the Appendix.

**Training and test loss** In Figure 1, we plot the average training and test performance over 100 trails. It is apparent that training using a base distribution with the correct type of tail behavior is beneficial. First of all, we observe a significant gap between the test performance of the vanilla NF and the NFs with a heavy-tailed base distribution. Notice that in Figure 1 we excluded all runs with a final test loss of above 25, which happened in 17 of the *normal* runs and not once in the other cases. Furthermore, we clearly observe a much faster convergence and a more stable training procedure. The fluctuations and instabilities in the vanilla NF are due to tail-samples that have a massive effect on the likelihood in (2), which can be reduced by choosing base distributions with slower decaying tails (Alexanderson & Henter, 2020).

**Learning the tails** To illustrate the ability to model the tails, we compared the estimated empirical quantile func-

[1] all computational details can be found in Section C of the Appendix

[2] We are using a 3-layered MAF (Papamakarios et al., 2017). Further computational details can be found in Section C of the Appendix.

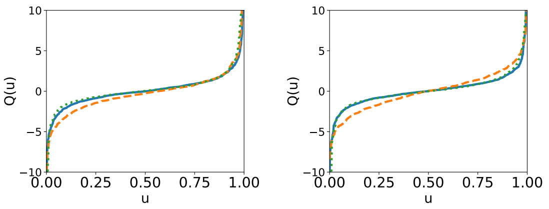


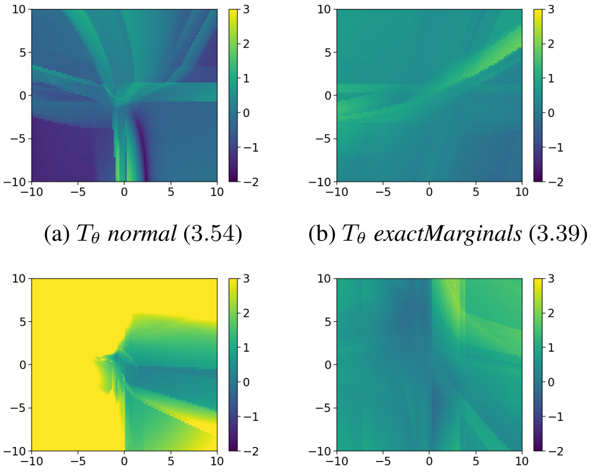
*Figure 2.* Estimated marginal quantiles in the case of *normal* (orange, dashed) and *exactMarginals* (green, dotted). The corresponding negative log-likelihoods are 4.00 and 3.39, respectively.

tions. We did so for both marginal distributions (Figure 2) but also for the distribution of $\|\mathbf{x}\|_2$ (Figure 9 in the Appendix). In line with the findings by Jaini et al. (2020), we notice that the vanilla NF is not capable of modeling the quantiles of the target distribution. More precisely, we observe that the corresponding quantile function is steeper around its center and has shorter tails. This means that the distribution learned by the NF does not account for the heavy tails by directly modeling them, but instead covers samples from the tails of the data distribution by being more widespread. In contrast, the base distributions that took the tailedness of $\mathbf{x}$ into account, could achieve a much better fit to the quantiles, see Figure 8 in the Appendix for further results.

**Invertibility and numerical stability** As investigated by Behrmann et al. (2021), the Bi-Lipschitz constant plays a fundamental role in the practical invertibility and numerical stability of NFs. To understand the learned transformation $T$ in terms of its Lipschitz continuity, we propose to study the *Lipschitz surface* of $T$. Note that if $T$ is differentiable and $L$-Lipschitz, we can follow the derivation by Behrmann et al. (2021) (in equation (5) and (6) therein) to approximate

$$\begin{aligned} L &= \sup_{x\in\operatorname{supp}(\mathbf{x})} \|J_T(x)\|_2 = \sup_{x\in\operatorname{supp}(\mathbf{x})} \sup_{\|v\|_2=1} \|J_T(x)v\|_2 \\ &\approx \sup_{x\in\operatorname{supp}(\mathbf{x})} \sup_{\|v\|_2=1} \frac{1}{\varepsilon}\|T(x)-T(x+\varepsilon v)\|_2 \ , \end{aligned}$$

where $\varepsilon$ is some small constant. This motivates to consider an estimate[3] of $\sup_{\|v\|_2=1} \|T(x)-T(x+\varepsilon v)\|_2/\varepsilon$ for $x \in \operatorname{supp}(\mathbf{x})$ as a local surrogate for the Lipschitz-continuity of $T$. Plotting these quantities for both, $T$ and $T^{-1}$, and $x \in [-10, 10]^2$ we obtain the Lipschitz surfaces, which are depicted in Figure 3. We notice that the vanilla NF has many fluctuations in the local Lipschitz-continuity, while the proposed copula method leads to a well-behaved transformation. The inverse transformation $T_\theta^{-1}$ in the vanilla NF has exploding local Lipschitz constants, while—again—the proposed method results in a stable inverse transformation.



(a) $T_\theta$ *normal* (3.54)

(b) $T_\theta$ *exactMarginals* (3.39)

(c) $T_\theta^{-1}$*normal* (3.54)

(d) $T_\theta^{-1}$ *exactMarginals* (3.39)

*Figure 3.* Examples for the Lipschitz surfaces of $T_\theta$ and $T_\theta^{-1}$ on a log-scale. The corresponding negative log-likelihood is shown in brackets.

## 4. Discussion

In this work, we paved the way toward a general extension of NF architectures using copulae. Synthetic experiments revealed that the modeling performance of NFs can be improved substantially by replacing the vanilla Gaussian base distribution by a base distribution that reflects basic properties of the data distribution more accurately. Of course, we have just scraped the surface of the underlying potential of the proposed approach: While we concentrate on the tail behavior of the marginals in this work, the general idea can potentially also be applied to incorporate other types of inductive bias, such as multimodality by choosing multimodal marginals, or symmetries and tail dependencies by selecting appropriate marginals and copulae.

Our experiments suggest that it is sufficient to have only a broad estimate of the marginals. As mentioned in Section 3.1, one could also employ IFM to learn these before training the NF. However, the question of what is the best technique for choosing or estimating an appropriate marginal distributions and copula still requires further investigation. Nevertheless, we think that this flexibility brings additional improvement over the methods proposed by Jaini et al. (2020) and Alexanderson & Henter (2020). Of course, we have yet only gained preliminary results with our empirical study, which we plan to verify on real-world data and for different models in the future.

We believe that our analysis of the base functions can help to popularize NFs in a wide spectrum of domains. One such application might be financial risk analysis, where it is essential to model tail dependencies.

[3] see Section C in the Appendix for details

## Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC- 2092 CASA – 390781972. We also thank the anonymous reviewers for their careful reading and their useful suggestions, which will help us extending this work to a full paper.

## References

Alexanderson, S. and Henter, G. E. Robust model training and generalisation with studentising flows. *arXiv preprint arXiv:2006.06599*, 2020.

Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. Understanding and mitigating exploding inverses in invertible neural networks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1792–1800, Virtual Event, 2021.

Chen, R. T. Q., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9913–9923, Vancouver, BC, Canada, 2019.

Dinh, L., Krueger, D., and Bengio, Y. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR Workshop Track Proceedings*, San Diego, CA, USA, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. nflows: normalizing flows in PyTorch, 2020.

Genest, C., Gendron, M., and Bourdeau-Brien, M. The advent of copulas in finance. In *The European Journal of Finance*, volume 15, pp. 609–618. Routledge, 2009.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 27, pp. 2672–2680, Cambridge, MA, USA, 2014.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. *Elements of Copula Modeling with R*. Springer Use R! Series, 2018.

Horn, G. V. and Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

Jaini, P., Kobyzev, I., Yu, Y., and Brubaker, M. Tails of Lipschitz triangular flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4673–4681, Virtual Event, 2020.

Joe, H. *Dependence modeling with copulas*. CRC press, 2014.

Joe, H. and Xu, J. J. The estimation method of inference functions for margins for multivariate models. In *Faculty Research and Publications*, Faculty Research and Publications, 1996.

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

Kobyzev, I., Prince, S., and Brubaker, M. Normalizing flows: An introduction and review of current methods. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2020.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2338–2347, Long Beach, CA, USA, 2017.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. In *Journal of Machine Learning Research*, volume 22, pp. 1–64, 2021.

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

Sklar, A. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pp. 1–14, 1996.

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

Zhang, X., Fang, Z., Wen, Y., Li, Z., and Qiao, Y. Range loss for deep face recognition with long-tailed training data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5419–5428, Los Alamitos, CA, USA, 2017.

Zhu, X., Anguelov, D., and Ramanan, D. Capturing long-tail distributions of object subcategories. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 915–922, Columbus, OH, USA, 2014.

# A. Examples of Copula Distributions

There is a wealth of different copula distributions, ranging from parametric distributions to semi-parametric and completely non-parametric copula models. In this Section, we give some insights into the construction of one class of copula distributions and illustrate some basic properties.

**Example A.1** (Construction using Sklar's Theorem)**.** *There is a generic way to construct a copula according to Definition 2.1. Consider any multivariate continuous and invertbile CDF $\Phi$ with marginal CDFs $\Phi_1, \ldots, \Phi_D$. Then, following (4) in Sklar's Theorem 2.2, we know that*

$$C(u) := \Phi\big(\Phi_1^{-1}(u_1), \ldots, \Phi_D^{-1}(u_D)\big) \qquad (7)$$

*defines a valid copula. Setting $\Phi$ to be the multivariate Gaussian distribution with correlation matrix $R$, we obtain the* Gaussian Copula*. If $R$ is the identity matrix we obtain the* Independence Copula*, which is simply the product of independent uniform distributions. Both copulae are visualized in Figure 4. Following the construction in (7) and replacing the CDF, we can obtain other copulae,such as the t-Copula. Figure 5 visualizes one example for a distribution based on the Independence Copula and based on the Gaussian Copula.*

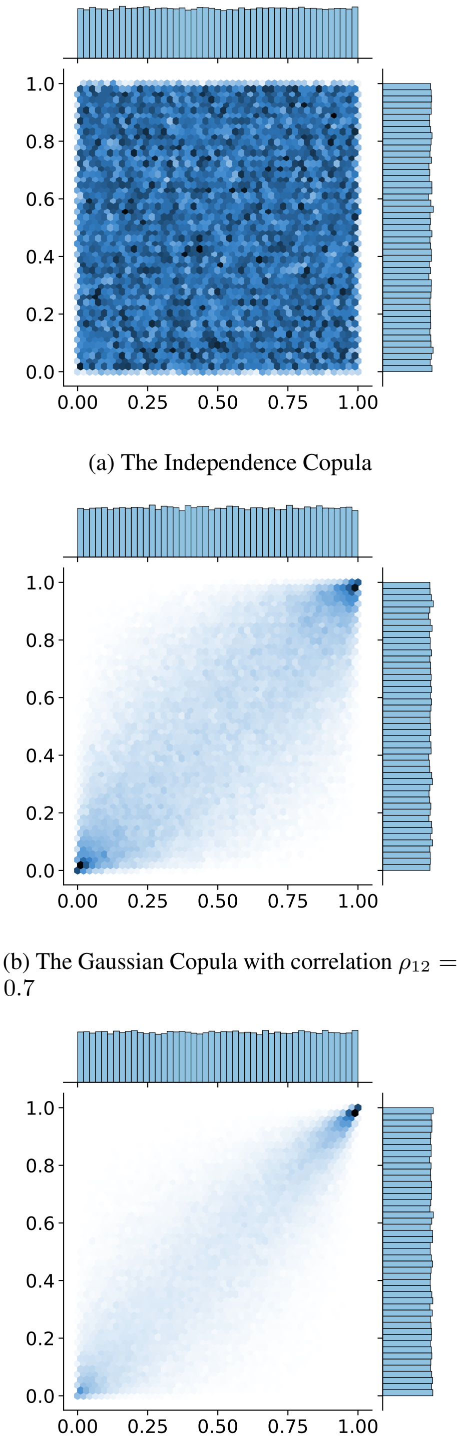
**Example A.2** (Copulae that induce Tail-Dependency)**.** *A crucial property of models in financial risk analysis is* tail dependency*. Roughly said, a tail dependency casts marginals to be dependent in a tail event. To give an example for a tail dependency, consider two essentially independent marginal distributions, such as $\mathbf{x}_1 = \{$debt status of your bank$\}$ and $\mathbf{x}_2 = \{$trust for your bank$\}$. Usually, our trust for the bank is not essentially determined by the amount of debts it has. However, in a marginal tail event $\{\mathbf{x} =$ bankrupt$\}$, of course, our trust drops rapidly. Mathematically, the upper tail dependency, i.e. the dependency in upper-tail events, for a random variable $(\mathbf{x}_1, \mathbf{x}_2)$ is given by*

$$\lambda_U = \lim_{u \to 1^-} \mathbb{P}\big(\mathbf{x}_2 > F_{\mathbf{x}_2}^{-1}(u) \,|\, \mathbf{x}_1 > F_{\mathbf{x}_1}^{-1}(u)\big) \ .$$

*Similarly, we can define the lower tail dependency $\lambda_L$. One such copula that accounts for upper tail dependency is the* Gumbel Copula*, which is given by*

$$C(u) := \exp\big(-((-\log(u_1))^\rho + (-log(u_2))^\rho)^{1/\rho}\big) \qquad (8)$$

*for $\rho \geq 1$. Figure 4 shows a visualization of the Gumbel copula. One can show that $\lambda_U = 2 - 2^{1/\theta}$ (Joe, 2014). Figure 5 visualizes the Gumbel copula distribution with Gaussian and Gamma marginals. We can observe a decent dependency—indicated by the peak pointing to the upper right—in the upper-tail events, which is mathematically described by $\lambda_U$.*



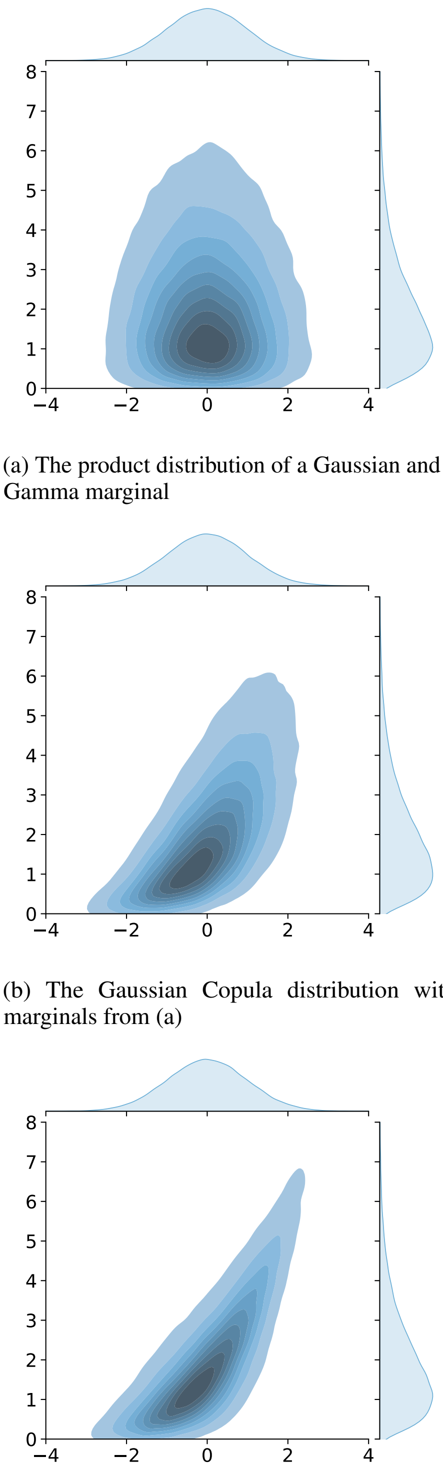
(a) The Independence Copula

(b) The Gaussian Copula with correlation $\rho_{12} = 0.7$

(c) The Gumbel Copula with parameter $\rho = 2.5$

*Figure 4.* Some popular Copulae.

One could think of many more different properties that might be included using our copula approach, such as multimodality and symmetry. However, it is still to be researched

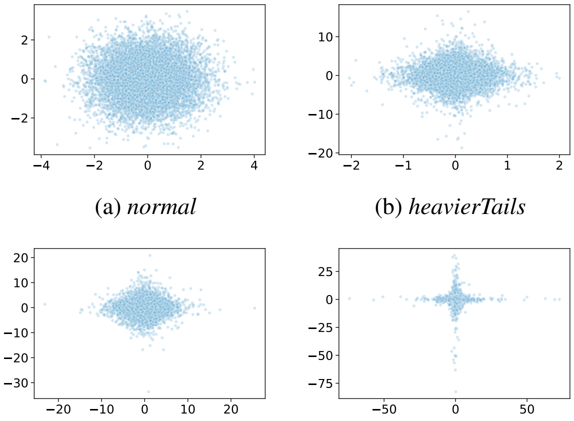


(a) The product distribution of a Gaussian and a Gamma marginal

(b) The Gaussian Copula distribution with marginals from (a)

(c) The Gumbel Copula distribution with marginals from (a)

*Figure 5.* Some examples of distributions constructed via the copula approach.



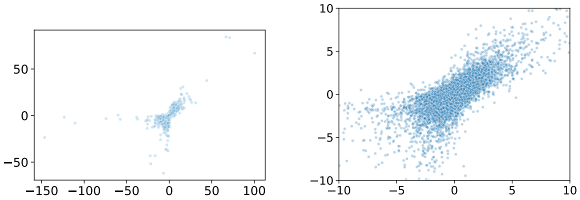
(a) *normal*

(b) *heavierTails*

(c) *correctFamily*

(d) *exactMarginals*

*Figure 6.* Samples from the 4 different base distributions that we used in our experiments.



(a) Gumbel Copula with $t_2(0,1)$-distributed marginals.

(b) The distribution from 7a zoomed in.

*Figure 7.* Samples from the target distribution.

whether properties, such as the tail-dependency, are being preserved by the NF. Nonetheless, we think that fixing specific known properties in the base distribution facilitates training in acting as a type of regularization towards these given properties.

## B. Supplementary Experiments

In the following, we give some further empirical results that underpin our findings from the main paper.

Figure 6 shows the investigated base distributions and Figure 7 visualizes the target distribution.

Figure 8 and Figure 9 supplement the findings about the learned quantiles: Even the other heavy-tailed distributions—the cases *heavierTails* and *correctFamily*—are able to successfully approximate the quantiles. This observation suggests that it is sufficient to use a broad surrogate of the base distribution that captures the true tail behavior.

To address the invertibility and numerical stability of the learned transformation, we investigate the local Lipschitz constant as derived in Section 3.2 in the main text. Fig-

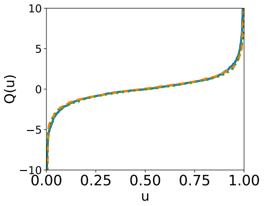
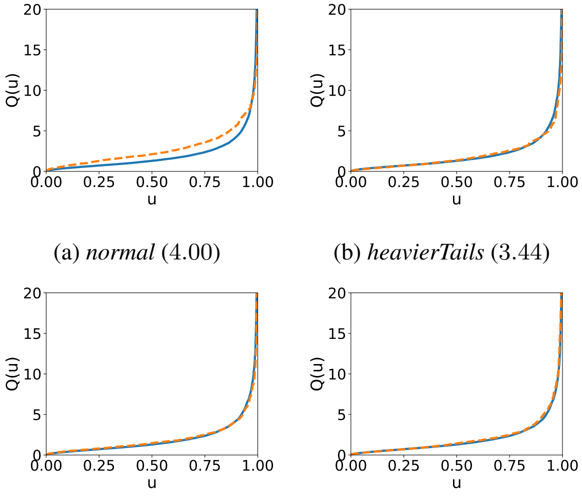


*Figure 8.* Estimated marginal CDFs in the case of *heavierTails* (orange, dashed) and *correctFamily* (green, dotted). The corresponding negative log-likelihoods are $3.44$ and $3.43$, respectively.



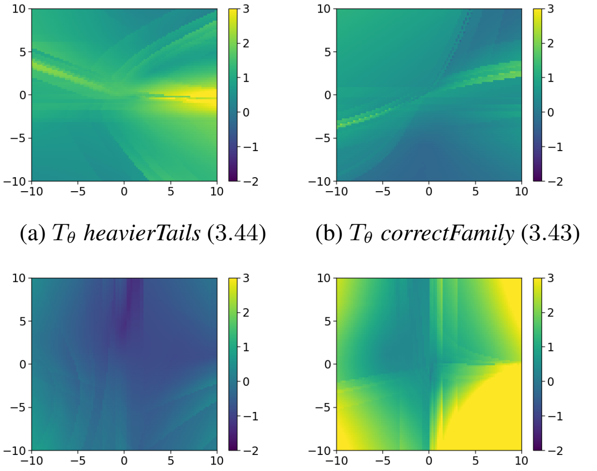
(a) *normal* $(4.00)$

(b) *heavierTails* $(3.44)$

(c) *correctFamily* $(3.43)$

(d) *exactMarginals* $(3.39)$

*Figure 9.* Estimated quantiles of $\|\mathbf{x}\|_2$ using the different base distributions. The corresponding negative log-likelihood is shown in brackets.



(a) $T_\theta$ *heavierTails* $(3.44)$

(b) $T_\theta$ *correctFamily* $(3.43)$

(c) $T_\theta^{-1}$*heavierTails* $(3.44)$

(d) $T_\theta^{-1}$ *correctFamily* $(3.43)$

*Figure 10.* Examples of the Lipschitz surfaces of $T_\theta$ and $T_\theta^{-1}$ on a log-scale. The corresponding negative log-likelihood is shown in brackets.

ure 10 shows the supplementary Lipschitz surfaces of the cases *heavierTails* and *correctFamily*, which indicate a more regular transformation than the *normal* case. Still, in the case *correctFamily* we find large local Lipschitz constants in $T_\theta^{-1}$, which, however, are structured in accordance to the base distribution. In the vanilla NF this irregularity is less structured and, since these irregularities occur in high-probability areas of the base distribution, more relevant.

These results support our claim that choosing appropriately tailed base distribution can help in learning a numerically robust transformation.

## C. Computational Details

In all of our experiments, we employed *Masked Autoregressive Flows* (Papamakarios et al., 2017) with 3 layers. Each layer contains a reverse permutation, followed by an autoregressive transformation with 4 hidden features. The code is based on the *nflows* package (Durkan et al., 2020).

We trained the NFs on a sample of size $10\,000$. Optimization was carried out using the Adam optimizer with the *PyTorch* default settings and a batch size of $128$. Test losses are evaluated based on $10\,000$ test samples.

The reported training and test losses (Figure 1) have been averaged over $100$ runs and the depicted confidence intervals correspond to a confidence of $95\%$.

To estimate the Lipschitz surface we rely on an estimation of

$$\sup_{\|v\|_2=1} \frac{1}{\varepsilon}\|T(x) - T(x+\varepsilon v)\|_2 \ , \tag{9}$$

for some small constant $\varepsilon$. We chose $\varepsilon := 10^{-3}$. Further, we approximate (9) by

$$\max_{j\in\{1,\ldots,100\}} \frac{1}{\varepsilon}\Big\|T(x) - T(x - \varepsilon\frac{v^{(j)}}{\|v^{(j)}\|_2}\Big\|_2,$$

where $v^{(1)}, \ldots v^{(100)}$ are i.i.d. samples from $\mathcal{N}(0, I)$, see Behrmann et al. (2021) for further details.

All code is provided with the submission and can further be accesses through `https://github.com/MikeLasz/Copula-Based-Normalizing-Flows`.