Tails of Lipschitz Triangular Flows

Priyank Jaini [*1 2] **Ivan Kobyzev** [3] **Yaoliang Yu** [1 2] **Marcus A. Brubaker** [3 4]

Abstract

We investigate the ability of popular flow based methods to capture tail-properties of a target density by studying the increasing triangular maps used in these flow methods acting on a tractable source density. We show that the density quantile functions of the source and target density provide a precise characterization of the slope of transformation required to capture tails in a target density. We further show that any Lipschitz-continuous transport map acting on a source density will result in a density with similar tail properties as the source, highlighting the trade-off between a complex source density and a sufficiently expressive transformation to capture desirable properties of a target density. Subsequently, we illustrate that flow models like Real-NVP, MAF, and Glow as implemented originally lack the ability to capture a distribution with non-Gaussian tails. We circumvent this problem by proposing tail-adaptive flows consisting of a source distribution that can be learned simultaneously with the triangular map to capture tail-properties of a target density. We perform several synthetic and real-world experiments to compliment our theoretical findings.

1. Introduction

Increasing triangular maps are a recent construct in probability theory that can transform any source density to any target density (Bogachev et al., 2005). The Knothe-Rosenblatt transformation (Rosenblatt, 1952; Knothe et al., 1957) gives an explicit version of an increasing triangular map that does the transformation. These triangular maps provide a unified framework (Jaini et al., 2019) to study popular neural density estimation methods like normalizing flows (Tabak & Vanden-Eijnden, 2010; Tabak & Turner, 2013; Rezende & Mohamed, 2015) and autoregressive models (Papamakarios et al., 2017; Huang et al., 2018; Kingma et al., 2016; Uria et al., 2016; Larochelle & Murray, 2011) which are tractable methods for explicitly modelling densities for high-dimensional datasets. Indeed, these methods have been applied successfully in several domains including natural images, videos, speech and audio synthesis, novelty detection, and natural language.

This work studies the tail properties of a target density by characterizing the properties of the corresponding increasing triangular map required to push a tractable source density with known tails to the desired target density. We begin in §3 by showing that, in one dimension, the density quantile functions of the source and target density characterize the slope of a (unique) increasing transformation. Furthermore, the asymptotic properties of the density quantile function allow us to give a granular characterisation of the *degree of heaviness* of a distribution. We show that the degree of heaviness parameter of the source and target densities characterize the properties of the corresponding triangular map completely. We then give a precise rate at which an increasing transformation must grow in order to capture the tail behaviour of the target density by drawing connections between the degree of heaviness parameter and the existence of higher-order moments of the densities.

We generalize these results for higher dimensions in §4 by showing that a Lipschitz-continuous transport map will always result in a target density with the same tail properties as the source, highlighting the trade-off between choosing an *appropriate* source density and *sufficiently complex* transport map to capture tails in a target density. Additionally, when the source and target densities are from the elliptical family, we show that the increasing triangular map from a light-tailed distribution to a heavy-tailed distribution must have all diagonal entries of the Jacobian unbounded.

In §5, we discuss the implications of these results for a class of flow based models that we call *affine* triangular flows which include NICE (Dinh et al., 2015), Real-NVP (Dinh et al., 2017), MAF (Papamakarios et al., 2017), IAF (Kingma et al., 2016), and Glow (Kingma & Dhariwal, 2018). We show both theoretically and empirically that these models as originally implemented lack the ability to push a fixed source density to a target density with heav-

[*]Equal contribution [1]Univesity of Waterloo, Waterloo, Canada [2]Vector Institute, Toronto, Canada [3]Borealis AI [4]York University, Toronto, Canada. Correspondence to: Priyank Jaini <p.jaini@uwaterloo.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

ier tails. To circumvent these draw-backs of affine flows, we subsequently propose *tail-adaptive flows* in §6, where the source density, instead of being fixed, is endowed with a learnable parameter that controls its tail behaviour and allows affine flows to capture tail properties of the target density. We illustrate these properties of tail-adaptive flows empirically and demonstrate their performance on benchmark datasets.

Contributions. We summarize our main contributions as follows:

- We show that density quantiles precisely capture the properties of a push-forward transformation. We use these to provide asymptotic rates for the slope of maps required to capture heavy-tailed behaviour.
- We show that Lipschitz push-forward maps cannot change the tails of the source density qualitatively. We thus reveal a trade-off between choosing a "complex" source density and an "expressive" transformation for representing heavy-tailed target densities.
- As a consequence, we show that several popular flow models as originally implemented lack the ability to capture heavier tailed density than the fixed source.
- We propose tail-adaptive flows that can be deployed easily in any existing flow based and autoregressive model to better capture tail properties of a target density. We also demonstrate the importance of choosing an appropriate source density.

Due to space constraints, proofs are deferred to Appendix A.

2. Preliminaries and Set-Up

In this section we set up our main problem, introduce key definitions and notations, and formulate the framework of characterizing tail properties of a target probability density through the unique triangular push-forward map.

We call a mapping $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$ *triangular* if its j-th component T_j only depends on the first j variables $z_1, \dots, z_j$. The name "triangular" comes from the fact that the Jacobian $\nabla \mathbf{T}$ is a triangular matrix function. Further, we call $\mathbf{T}$ increasing if for all $j \in [d]$, T_j is an increasing function of z_j. Triangular transformations are appealing due to the following result by Bogachev et al. (2005):

Theorem 1 (Bogachev et al. 2005)**.** *For any two densities p and q over $\mathsf{Z} = \mathsf{X} = \mathbb{R}^d$, there exists a unique (up to null sets of p)* increasing *triangular map $\mathbf{T} : \mathsf{Z} \to \mathsf{X}$ so that if $\mathbf{Z} \sim p$ then $\mathbf{T}(\mathbf{Z}) \sim q$, i.e. q is the push-forward of p, or in symbols $q = \mathbf{T}_{\#}p$.*

Let us give an example to help understand Theorem 1.

Example 1 (Increasing Rearrangement)**.** *Let p and q be univariate probability densities with distribution functions F and G, respectively. One can define the increasing map $T = G^{-1} \circ F$ such that $q = T_{\#}p$, where $G^{-1} : [0,1] \to \mathbb{R}$ is the quantile function of q:*

$$G^{-1}(u) := \inf\{t : G(t) \geq u\}. \tag{1}$$

Indeed, if $Z \sim p$, one has that $F(Z) \sim$ uniform. Also, if $U \sim$ uniform, then $G^{-1}(U) \sim q$. Theorem 1 is a rigorous iteration of this univariate argument by repeatedly conditioning (a construction popularly known as the Knothe-Rosenblatt transformation (Rosenblatt, 1952; Knothe et al., 1957)). Specifically, the j-th component T_j of $\mathbf{T}$ for the Knothe-Rosenblatt transformation is given by $x_j = T_j(z_1, \dots, z_{j-1}, z_j) = F^{-1}_{q,j|<j} \circ F_{p,j|<j}(z_j)$ where $F_{q,j|<j}$ is the cdf of the conditional distribution of X_j given $\mathsf{X}_{<j} := (\mathsf{X}_1, \dots, \mathsf{X}_{j-1})$, and similarly for $F_{p,j|<j}$.

Jaini et al. (2019) showed that several popular normalizing flows and autoregressive models like NICE (Dinh et al., 2015), Real-NVP (Dinh et al., 2017), IAF (Kingma et al., 2016), MAF (Papamakarios et al., 2017), NAF (Huang et al., 2018), and SOS Flows (Jaini et al., 2019) employ increasing triangular transformations as fundamental modules to construct expressive push-forward transformations and are precisely special cases of learning increasing triangular maps.[1]

In this work, we characterize the properties of increasing triangular maps required to capture the tail properties of the target density q given a known source density p and discuss the implications of these results for flow based models that use affine triangular transformations e.g. Real-NVP (Dinh et al., 2017), MAF (Papamakarios et al., 2017), Glow (Kingma & Dhariwal, 2018), etc.

Formally, we characterize the tail properties of the target density q by studying the properties of the induced increasing triangular map $\mathbf{T}$ acting on a known fixed source density p. This approach has been used earlier by Spantini et al. (2018) who studied the Markov properties of the target density and the existence of low-dimensional couplings by characterizing the properties of the induced triangular map and showing that such a map is both sparse and decomposable. Similar studies have been undertaken to characterize the tail properties of "optimal" transport maps by de Valk & Segers (2019) whose results only apply to a related limiting density but not the original ones, and for elliptical distributions by Ghaffari & Walker (2018). In contrast, we focus specifically on *triangular maps* that are used extensively for tractable density estimation in normalizing flows and auto-regressive models (Jaini et al., 2019) and can be learned efficiently using deep neural networks.

Example 1 shows that the increasing triangular map be-

[1] We direct the reader to Section 3 and Table 1 in Jaini et al. (2019) for a comprehensive overview of connecting triangular maps to several models in unsupervised learning.

tween two densities can be constructed iteratively by using the univariate increasing rearrangement repeatedly on the conditional distributions and the quantile functions. We employ the same strategy to characterize the properties of a triangular map $\mathbf{T}$ by characterizing the properties of the univariate maps T_j. Thus, in the next section, we first explore in detail the properties of univariate (increasing) maps.

3. Properties of Univariate Transformations

We define the class of heavy tailed distributions $\mathcal{H}$ as those that have no finite higher-order moments (Foss et al., 2011):

$$\mathcal{H} := \Big\{ p : \forall\, \lambda > 0,\ m_p(\lambda) := \mathop{\mathbf{E}}_{Z\sim p}[e^{\lambda Z}] = \infty \Big\}.$$

otherwise, it is light-tailed i.e. $p \in \mathcal{L}$ if all its higher-order moments are finite[2]. We show that *any* diffeomorphic transformation T that pushes a source density $p \in \mathcal{L}$ to a target density $q \in \mathcal{H}$ cannot have a bounded slope globally.

Theorem 2. *Let $p \in \mathcal{L}$ and $q \in \mathcal{H}$ such that $q = T_{\#}p$, where T is a diffeomorphism. Then, for all $M > 0$ and all $z_0 > 0$ there is $z > z_0$, such that $T'(z) > M$. Conversely, if T is a Lipschitz-continuous map & $p \in \mathcal{L}$, then, $T_{\#}p \in \mathcal{L}$.*

Theorem 2 is mostly a qualitative result, and it provides little knowledge about the map T required to capture a heavy-tailed distribution q given a source density p. Moreover, we would ideally like to characterize the properties of T in terms of the "*degree of heaviness*" of p and q respectively. We will address this problem by proposing a refined definition of tails of a density function in terms of the asymptotic behaviour of the density quantile function as formulated by Parzen (1979) and Andrews et al. (1973).

For a probability density p over a domain $\mathsf{Z} \subseteq \mathbb{R}$, let $F_p : \mathsf{Z} \to [0,1]$ denote the cumulative distribution function of p, and $Q_p : [0,1] \to \mathsf{Z}$ be the quantile function given by $Q_p = F_p^{-1}$. Then, $fQ_p : [0,1] \to \mathbb{R}_+$ is called the density quantile function and is given by $fQ_p = {}^1\!/_{Q_p'}$. Parzen (1979) proved that the limiting behaviour of any density quantile function as $u \to 1^-$ is given by:

$$fQ(u) \sim (1-u)^{\alpha}, \qquad \alpha > 0 \tag{2}$$

where $g(u) \sim h(u)$ implies that $\lim_{u\to 1^-} {}^{g(u)}\!/_{h(u)}$ is a finite constant. We can additionally define the limiting behaviour of the quantile function $Q(u)$ when $u \to 1^-$ as:

$$Q(u) \sim (1-u)^{-\gamma}, \qquad \gamma = \alpha - 1. \tag{3}$$

The parameter α is called the tail-exponent and defines the tail-area of a distribution and acts as a measure of "degree of heaviness." Indeed, for two distributions with tail exponents α_1 and α_2, if $\alpha_1 > \alpha_2$, the former has heavier tails relative to the latter. Thus, the tail exponent α allows us to classify distributions based on their degree of heaviness.

$$\text{Define} \quad \mathcal{H}_\alpha := \Big\{ p \ :\ fQ_p \sim (1-u)^{\alpha} \text{ as } u \to 1^- \Big\}.$$

Following Parzen (1979), if $0 < \alpha < 1$ the distributions are light-tailed, e.g. the Uniform distribution. Here, we further show that a distribution has support bounded from above if and only if the right density quantile function has tail-exponent $0 < \alpha < 1$.

Proposition 1. *Let p be a density with $fQ_p \sim (1-u)^{\alpha}$ as $u \to 1^-$. Then, $0 < \alpha < 1$ iff supp$(p) = [a, b]$ where $b < \infty$ i.e. p has a support bounded from above.*

$\mathcal{H}_1$ corresponds to a family of distributions for which all higher order moments exist. However, these distributions are relatively heavier tailed than short-tailed distributions and were termed as medium tailed distributions by Parzen (1979), e.g. normal and exponential distribution. Additionally, for $\alpha = 1$, a more refined description of the asymptotic behaviour of the quantile function can be given in terms of the shape parameter β:

$$fQ(u) \sim (1-u)\Big(\log\frac{1}{1-u}\Big)^{1-\beta},$$
$$\text{and} \quad Q(u) \sim \Big(\log\frac{1}{1-u}\Big)^{\beta}, \qquad 0 \le \beta \le 1.$$

β determines the degree of heaviness in medium tailed distributions; the smaller the value of β, the heavier the tails of the distribution e.g. exponential distribution has $\beta = 1$, and normal distribution has $\beta = 0.5$. We thus define

$$\mathcal{H}_{1,\beta} = \Big\{ p : fQ_p \sim (1-u)\Big(\log\frac{1}{1-u}\Big)^{1-\beta},\ 0 \le \beta \le 1 \Big\}$$

and we have $\mathcal{H}_1 = \cup_{0\le\beta\le 1}\mathcal{H}_{1,\beta}$. Further, the class of light tailed distributions defined in the beginning of the section is $\mathcal{L} = \cup_{0<\alpha\le 1}\mathcal{H}_\alpha$. Finally, the class of heavy tailed distributions have $\alpha > 1$ i.e. $\mathcal{H} = \cup_{\alpha>1}\mathcal{H}_\alpha$, e.g. student-t distribution t_ν with ν degrees of freedom .

We are now in a position to characterize the map T based on the degree of heaviness of the source and target densities. Following Example 1, the slope of T is given by the ratio of the density quantile function of the source and the target distribution respectively, i.e.

$$T'(z) = \frac{p(z)}{q\circ T(z)} = \frac{p\Big(F_p^{-1}\circ F_p(z)\Big)}{q\Big(F_q^{-1}\circ F_p(z)\Big)}$$
$$i.e. \quad T'(z) = \frac{fQ_p(u)}{fQ_q(u)}, \quad \text{where} \quad u = F_p(z).$$

[2] We note that this definition is restricted to only right-tails. For the sake of simplicity we develop our results for right-tails, but they generalise to left-tails naturally.

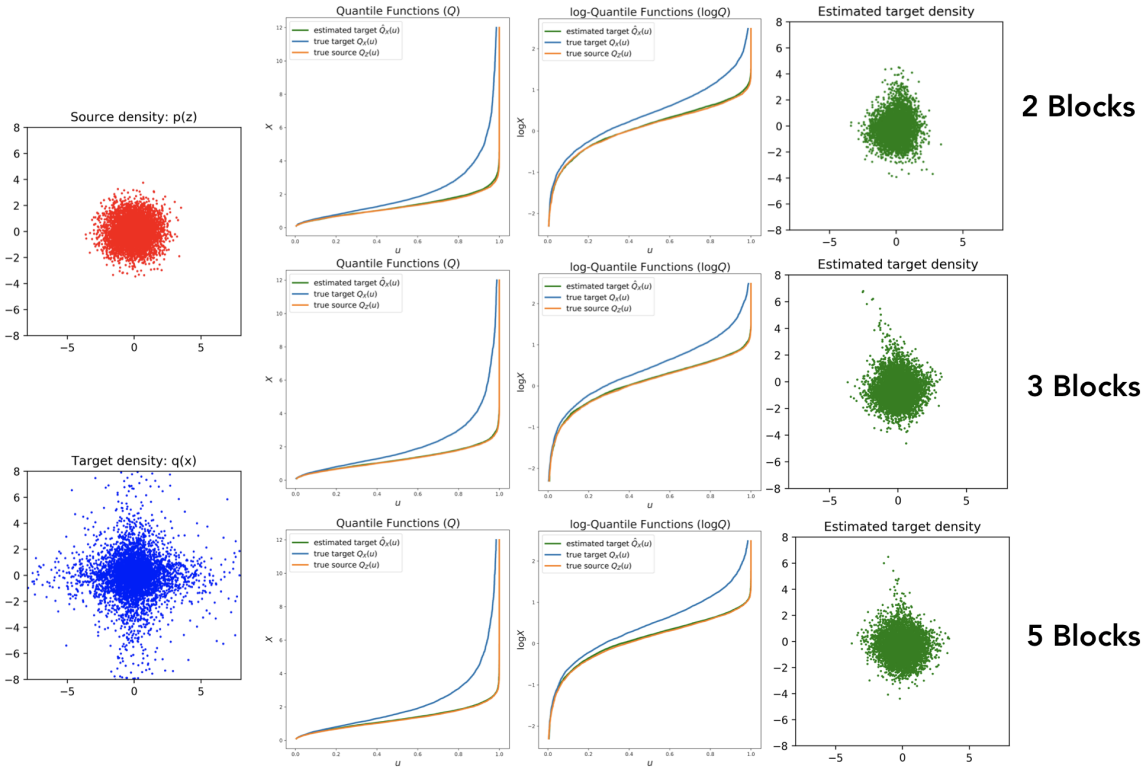

Figure 1. Results for Real-NVP illustrating the inability to capture tails. The second and third column show the quantile and log-quantile plots for the source, target, and estimated target density. The quantile function of the source and the estimated target density are identical depictng the inability to capture heavier tails. This is further explained by the estimated tail-coefficients $\gamma_{\text{source}} = 0.15$, $\gamma_{\text{target}} = 0.81$, and $\gamma_{\text{estimated}-\text{target}} = 0.15$. Best viewed in color. More details in Section 5.

Clearly, the density quantile functions precisely characterizes the slope of an increasing map needed to push a source density p to a target density q.

Proposition 2. *Let p and q be two square integrable univariate densities such that $q := T_{\#}p$. If the density quantile fQ_p of p shrinks to 0 at a rate slower than the density quantile fQ_q of q, then $T'(z)$ is asymptotically unbounded.*

Example 2 in Appendix B helps to illustrate Proposition 2 and the next corollary provides a precise characterization of asymptotic properties of a diffeomorphic transformation between densities with varying tail behaviour.

Corollary 1. *Let $p \in \mathcal{H}_{\alpha_p}$ be a source density, $q \in \mathcal{H}_{\alpha_q}$ be a target density and T be an increasing transformation such that $q = T_{\#}p$. Then, $\lim_{z\to\infty} T'(z) = \lim_{u\to 1^-}(1-u)^{\alpha_q-\alpha_p}$. Further, if $\alpha_p = \alpha_q = 1$, then $\lim_{z\to\infty} T'(z) = \lim_{u\to 1^-}\left(\log {}^1/_{(1-u)}\right)^{\beta_p-\beta_q}$ where $u = F_p(z)$.*

Example 3 in Appendix B further underlines the importance of density quantile functions to study tails of increasing transformations. We now connect the tail-exponent parameter $\alpha(\cdot)$ to the existence of higher-order moments of a random variable. Given a random variable $X \sim p$, the expected value of a function $g(x)$ can be written in terms of the quantile function as: $\mathbb{E}_p[g(x)] = \int_0^1 gQ_p(u)\,\mathrm{d}u$. This allows us to draw a precise connection between the degree of heaviness of a distribution as given by the density quantile functions (and tail exponent α) and the the existence of the number of its higher-order moments (ω).

Proposition 3. *Let p be a distribution with $Q_p(u) \sim (1-u)^{-\gamma}$ as $u \to 1^-$. Then, $\int_{z_0}^{\infty} z^{\omega}p(z)dz$ exists and is finite for some z_0 iff $\omega < \frac{1}{\gamma}$.*

Corollary 2. *If p is a distribution with $Q_p(u) \sim (1-u)^{-\gamma}$ as $u \to 1^-$ and $Q_p(u) \sim u^{-\gamma}$ as $u \to 0^+$.[3] Then, $\mathbb{E}_p[|z|^{\omega}]$ exists and is finite iff $\omega < \frac{1}{\gamma}$.*

Based on these observations, we can equivalently define heavy-tailed distributions as follows:

Definition 1. *A distribution $p(z)$ with compact support i.e. supp$(p) = [a, b]$ where $|a| < \infty$ and $|b| < \infty$ is said to be $\omega-$heavy tailed if for all $0 < \mu < \omega$, $\mathbb{E}_p[|z-b|^{1/\mu}]$ exists and is finite, but for $\mu \geq \omega$, $\mathbb{E}_p[|z-b|^{1/\mu}]$ is infinite or does not exist.*

Definition 2. *A distribution $p(z)$ with tail exponent $\alpha = 1$*

[3]This condition takes the left-tail into account as well. Note that it is not necessary for both tails to have the same behaviour and our analysis extends to such cases.

is said to be $\omega-$heavy tailed if for all $0 < \mu < \omega$, $\mathbb{E}_p[e^{|z|^\mu}]$ exists and is finite, but for $\mu \geq \omega$, $\mathbb{E}_p[e^{|z|^\mu}]$ is infinite or does not exist.

Definition 3 ($\omega^{-1}-$heavy tailed distributions)**.** *A distribution $p(z)$ with tail-exponent $\alpha > 1$ is heavy tailed with degree ω^{-1} with $\omega \in \mathbb{R}_+$ if for all $0 < \mu < \omega$, $\mathbb{E}_p[|z|^\mu]$ exists and is finite, but for all $\mu \geq \omega$, $\mathbb{E}_p[|z|^\mu]$ is infinite or does not exist.*

These definitions allow us to finally give the rate an increasing transformation must emulate to exactly represent tail-properties of a target density given some source density.

Proposition 4. *Let p be a $\omega_p^{-1}-$heavy distribution, q be a $\omega_q^{-1}-$heavy distribution and T be a diffeomorphism such that $q := T_{\#}p$. Then for small $\epsilon > 0$, $T(z) = o(|z|^{\omega_p/\omega_q - \epsilon})$.*

4. Properties of Multivariate Transformations

We now generalize our results to higher dimensions by first fixing the definition of a heavy-tailed distribution in higher dimensions [4]. We say that a random variable $\mathsf{X} \subseteq \mathbb{R}^d$ admits a heavy-tailed density function if the univariate random variable $\|\mathsf{X}\|$ has a heavy tailed density where $\|\cdot\|$ is some norm function. The granular definitions from Section 3 can be extended to the multivariate case through the density function of $\|\mathsf{X}\|$.

Theorem 3. *Let $\mathsf{Z} \subseteq \mathbb{R}^d$ be a random variable with density function p that is light-tailed and $\mathsf{X} \subseteq \mathbb{R}^d$ be a target random variable with density function q that is heavy-tailed. Let $\mathbf{T} : \mathsf{Z} \to \mathsf{X}$ be such that $q = \mathbf{T}_{\#}p$, then $\mathbf{T}$ cannot be a Lipschitz function.*

Corollary 3. *Under the same set-up as in Theorem 3, there exists an index $i \in [d]$ such that $\|\nabla_{\mathbf{z}} T_i\|$ is unbounded.*

Theorem 3 is a general result for any diffeomorphic transformation between two densities and we discuss the implication of this result for flow based models in §5. However, before proceeding further, we also characterize the properties of the triangular map $\mathbf{T}$ such that $q = \mathbf{T}_{\#}p$ by studying the properties of the univariate maps $T_j,\ j \in [d]$ obtained by repeated conditioning when the source and target densities are from the class of elliptical distributions.

Definition 4 (Elliptical distribution, (Cambanis et al., 1981))**.** *A random vector $\mathsf{X} \subseteq \mathbb{R}^d$ is said to be elliptically distributed denoted by $\mathsf{X} \sim \varepsilon_d(\boldsymbol{\mu}, \Sigma, F_R)$ with $\operatorname{rank}(\Sigma) = r$ if and only if there exists a $\boldsymbol{\mu} \in \mathbb{R}^d$, a matrix $\mathbf{A} \in \mathbb{R}^{d \times r}$ with maximal rank r, and a non-negative random variable R, such that $\mathsf{X} \overset{d}{=} \boldsymbol{\mu} + R\mathbf{A}\mathbf{U}^{(d)}$, where the random r-vector $\mathbf{U}$ is independent of R and is uniformly distributed over the unit sphere $\mathcal{B}_{d-1}$, $\Sigma = \mathbf{A}^T\mathbf{A}$ and F_R is the cumulative distribution function of the variate R.*

For ease in developing our results, we consider only full rank elliptical distributions i.e. $\operatorname{rank}(\Sigma) = d$ but the results can be easily extended to the general case. The spherical random vector $U^{(d)}$ produces elliptically contoured density surfaces due to the transformation $\mathbf{A}$. The density function of an elliptical distribution as defined above is given by: $f(x) = |\det \Sigma|^{-\frac{1}{2}} g_R\big((x-\mu)^T\Sigma^{-1}(x-\mu)\big)$, where the function $g_R(t) : [0,\infty) \to [0,\infty)$ is related to f_R, the density function of R, by the equation: $f_R(r) = s_d r^{d-1} g_R(d^2),\ \forall d \geq 0$, here $s_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the area of a unit sphere. Thus, the tail properties of a random variable X with an elliptical distribution $\varepsilon_d(\boldsymbol{\mu}, \Sigma, F_R)$ is determined by the generating random variable R. Indeed, X is heavy-tailed in all directions if the univariate generating random variable R is heavy-tailed.

Define $m_{f_R}(k) = \frac{1}{s_d}\int_0^\infty r^k f_R(r)\,\mathrm{d}r, \quad \forall\ k \in \mathbb{R}_+$. Intuitively, $m_{f_R}(k)$ is the k-th order moment of f_R when k is integer-valued. This allows us to generalize the granular definition of heavy-tailed distributions (§3, Definition 3) to the multivariate elliptical case: the distribution $\varepsilon_d(\boldsymbol{\mu}, \Sigma, F_R)$ is ω^{-1}-heavy iff μ_k is finite for all $k < \omega$ i.e. iff F_R is ω^{-1}-heavy. Similarly, from Definition 2 one has that $\varepsilon_d(\boldsymbol{\mu}, \Sigma, F_R)$ is ω-heavy iff F_R is ω-heavy. Elliptical distributions have certain convenient properties: marginal, conditional and linear transformation of an elliptical distribution are also elliptical (see Appendix B). Furthermore, we derive the degree of heaviness parameter of the conditional distributions of an elliptical distribution.

Proposition 5. *Under the same assumptions as in Lemma 2 (App.B), if $\mathsf{X} \sim \varepsilon_d(0, \mathbf{I}, F_R)$ is ω^{-1}-heavy, then the conditional distribution of $\mathsf{X}_2|(\mathsf{X}_1 = \mathbf{x}_1)$ is $(\omega + d_1)^{-1}$-heavy where $\mathsf{X}_1 \subseteq \mathbb{R}^{d_1}$.*

Equipped with all the necessary results, we now show that an increasing triangular map $\mathbf{T}$ between a light-tailed and a heavy-tailed elliptical distribution has all diagonal entries of $\nabla\mathbf{T}$ unbounded.

Proposition 6. *Let $\mathsf{Z} \sim \varepsilon_d(0, \mathbf{I}, F_S)$ and $\mathsf{X} \sim \varepsilon_d(0, \mathbf{I}, F_R)$ have densities p and q respectively where F_R is heavier tailed than F_S. If $\mathbf{T} : \mathsf{Z} \to \mathsf{X}$ is an increasing triangular map such that $q := \mathbf{T}_{\#}p$, then all diagonal entries of $\nabla\mathbf{T}$ and $\det|\nabla\mathbf{T}|$ are unbounded.*

Remark 1. *Our analysis naturally extends to the case when the target density is lighter tailed by studying the corresponding inverse transformation $\mathbf{T}^{-1}$. Particularly, such a transformation should have a vanishing asymptotic slope to capture lighter-tailed distributions.*

[4] Note that due to the lack of total ordering there is no standard definition of multivariate heavy-tailed distributions.

Table 1. Affine triangular flows

Model	coefficients	$T_j(z_j\,; z_1, \ldots, z_{j-1})$
NICE	$\mu_j(z_{<l})$	$z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$
IAF	$\sigma_j(z_{<j}),\ \mu_j(z_{<j})$	$\sigma_j z_j + (1-\sigma_j)\mu_j$
MAF	$\lambda_j(z_{<j}),\ \mu_j(z_{<j})$	$z_j \cdot \exp(\lambda_j) + \mu_j$
Real-NVP	$\lambda_j(z_{<l}), \mu_j(z_{<l})$	$\exp(\lambda_j \cdot \mathbf{1}_{j \notin [l]}) \cdot z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$
Glow	$\sigma_j(z_{<l}), \mu_j(z_{<l})$	$\sigma_j \cdot z_j + \mu_j \cdot \mathbf{1}_{j \notin [l]}$

5. (Lack of) Tails in Affine Flows

We call a triangular map $\mathbf{T}$ affine if $T_j(z_j; z_1, \cdots, z_{j-1})$ is an affine function of z_j. Several autoregressive and flow models like NICE (Dinh et al., 2015), Real-NVP (Dinh et al., 2017), IAF (Kingma et al., 2016), MAF (Papamakarios et al., 2017), and Glow (Kingma & Dhariwal, 2018) use affine triangular maps as fundamental building blocks to construct expressive transport maps through composition (see Table 1). In the aforementioned models, the coefficients in Table 1 are the output of another network such that $\lambda_j = \mathsf{sigmoid}\big(f(z_1, \cdots, z_{j-1})\big)$ or $\lambda_j = \mathsf{tanh}\big(f(z_1, \cdots, z_{j-1})\big)$[5] and $\mu_j = \mathsf{relu}\big(g(z_1, \cdots, z_{j-1})\big)$, resulting in transformations that lack the ability to learn target densities with heavier tails than the source density. We formalize this result below.

Theorem 4. *Let p be a light-tailed density and $\mathbf{T}$ be a triangular transformation such that $T_j(z_j;\ z_{<j}) = \sigma_j \cdot z_j + \mu_j$. If, $\sigma_j(z_{<j})$ is bounded above and $\mu_j(z_{<j})$ is Lipschitz then the target density $q := \mathbf{T}_{\#}p$ is light-tailed.*

Conversely, if $\sigma_j(z_{<j})$ is bounded and $\mu_j(z_{<j})$ is Lipschitz then the tails of q can not be heavier than the source density p. Moreover, since Lipschitz-continuous affine transformations cannot change tail behaviour, linear maps like permutations and 1×1 convolutions will also lack the ability to capture heavier tails. Thus, through an iterative argument, it is seen easily that composition of several such affine triangular maps (combined with permutations and 1×1 convolutions) will still be unable to push a source density to a target density with heavier tails. We note that most models in Table 1 in their proposed functional form can capture heavier tails due to the exponential term in the coefficient. However, we found that in practice this leads to instability during training while capturing heavy-tailed distributions. Instead, $\mathsf{sigmoid}(\cdot), \mathsf{tanh}(\cdot),$ and $\mathsf{relu}(\cdot)$ are used for modeling these affine flows resulting in models that are unable to capture heavier tails.

We next use synthetic experiments to supplement our findings above and illustrate this inability of certain affine flows to capture tails empirically. In our setup, we choose a bi-variate Gaussian distributed random variable i.e. $\mathsf{Z} \sim \mathcal{N}(0, \mathbf{I})$ as the source and a bi-variate student-t distributed target random variable with two degrees of freedom i.e. $\mathsf{X} \sim t_2(0, \mathbf{I})$. We measure the tail behaviour of a multivariate random variable by measuring the tail-coefficient γ (c.f. Eq.(3)) of the quantile function of the ℓ_2 norm of the random variable. We recall that if the tail-coefficient of a density q is larger than another density p, then q is heavier tailed than p (see §3). We learn an affine triangular flow $\mathbf{T} : \mathsf{Z} \to \mathsf{X}$ where we experiment with architectures the same as Real-NVP, MAF, and Glow. We generated 10,000 samples from the target density and used negative log-likelihood as the training objective. We divided the dataset into training-validation-testing in the ratio 2:1:1. We trained the model using Adam (Kingma & Ba, 2014) for 40 epochs with a batch size of 128 and learning rate of 10^{-3}.

Figure 1 shows the results in detail for Real-NVP with $\lambda(\cdot) = \mathsf{tanh}(\cdot)$. The first column plots the samples from the source (Gaussian, red) and target (student-t, blue) distribution, respectively. The three rows from top to bottom in second to fourth columns correspond to results from transformations learned using two, three, and five compositions (or blocks). The second and third column depict the quantile and log-quantile (for clearer illustration of differences) functions of the source (orange), target (blue), and estimated target (green) and the fourth column plots the samples drawn from the estimated target density. The estimated target quantile function matches exactly with the quantile function of the source distribution illustrating the inability of Real-NVP to capture tails. This is further reinforced by the tail-coefficients $\gamma_{\mathsf{source}} = 0.15$, $\gamma_{\mathsf{target}} = 0.81$, and $\gamma_{\mathsf{estimated-target}} = 0.15$. The negative log-likelihoods for the target, and the estimated target on test data were -3.95 and -3.82 respectively. We also observe that the samples generated from the estimated target density capture only the high density regions of the target but fail to spread to the tail regions of the target density. We show similar results for the quantile and log-quantile plots in Figure 2 when $\lambda(\cdot) = \mathsf{sigmoid}(\cdot)$. In Figure 3 we show the results for architectures using MAF, Glow, and Real-NVP with composition of 5 blocks.

6. Tail-Adaptive Flows

We saw in Sections 3 and 4 that a Lipschitz-continuous map cannot push-forward a light-tailed source density to heavier tailed target density. Subsequently, we illustrated in Section 5 that several flow models that incorporate compositions of triangular affine maps as the function class for the transport map are unable to capture densities that are heavier tailed than the chosen source density. In Figure 6 in Appendix B we also show that for the same experiment set-up where affine triangular flows were unable to capture

[5] IAF and Glow use $\sigma_j = \mathsf{sigmoid}(\cdot)$

Table 2. Average test log-likelihoods and standard deviation for Tail-adaptive flows (TAF) over 5 trials (higher is better). The numbers in the parenthesis indicate the number of compositions used. Results for other models are from (Huang et al., 2018; Jaini et al., 2019).

Method	Power	Gas	Hepmass	MiniBoone	BSDS300
MADE	0.40 ± 0.01	8.47 ± 0.02	-15.15 ± 0.02	-12.24 ± 0.47	153.71 ± 0.28
MAF affine (5)	0.14 ± 0.01	9.07 ± 0.02	-17.70 ± 0.02	-11.75 ± 0.44	155.69 ± 0.28
MAF affine (10)	0.24 ± 0.01	10.08 ± 0.02	-17.73 ± 0.02	-12.24 ± 0.45	154.93 ± 0.28
MAF MoG (5)	0.30 ± 0.01	9.59 ± 0.02	-17.39 ± 0.02	-11.68 ± 0.44	156.36 ± 0.28
TAN	0.60 ± 0.01	12.06 ± 0.02	-13.78 ± 0.02	-11.01 ± 0.48	159.80 ± 0.07
NAF DDSF (5)	0.62 ± 0.01	11.91 ± 0.13	-15.09 ± 0.40	-8.86 ± 0.15	157.73 ± 0.04
NAF DDSF (10)	0.60 ± 0.02	11.96 ± 0.33	-15.32 ± 0.23	-9.01 ± 0.01	157.43 ± 0.30
SOS (7)	0.60 ± 0.01	11.99 ± 0.41	-15.15 ± 0.10	-8.90 ± 0.11	157.48 ± 0.41
TAF affine (5)	0.28 ± 0.01	9.87 ± 0.23	-17.41 ± 0.20	-11.71 ± 0.09	156.53 ± 0.52
TAF SOS (7)	0.59 ± 0.01	11.99 ± 0.34	-15.11 ± 0.18	-8.94 ± 0.23	157.52 ± 0.22

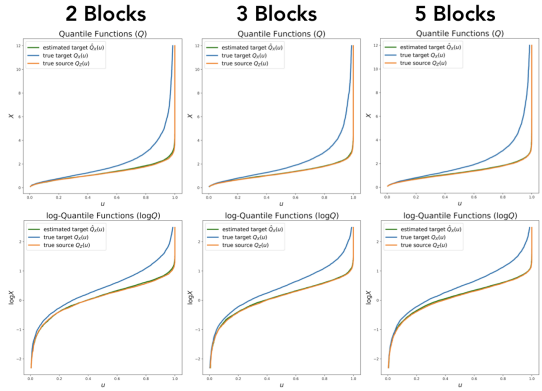

Figure 2. Same as Figure 1 with $\lambda(\cdot) = \text{sigmoid}(\cdot)$. The three columns correspond to two, three, and five compositions (blocks). $\gamma_{\text{source}} = 0.15$, $\gamma_{\text{target}} = 0.81$, and $\gamma_{\text{estimated}-\text{target}} = 0.15$.

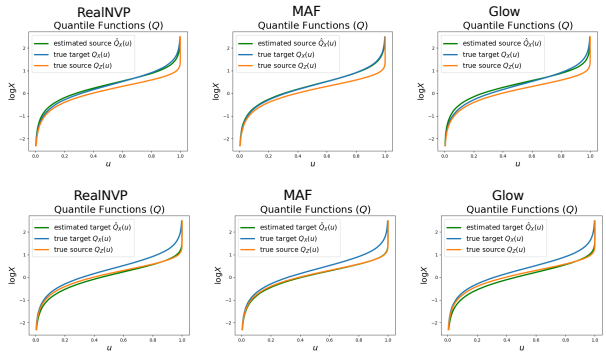

Figure 3. Quantile functions of distributions. Source is i.i.d student-t with 3 degree of freedom, true target is Gaussian, estimated source is a distribution modelled by an affine flow (in generative direction). Left is RealNVP, Middle is MAF, Right is Glow. Bottom row corresponds to the setting of $\mathbf{T} : \mathsf{Z} \to \mathsf{X}$ and top row corresponds when $\mathbf{T}^{-1} : \mathsf{X} \to \mathsf{Z}$

heavier tails of a density, SOS flows (Jaini et al., 2019) that use higher-order polynomial maps were able to learn the heavy-tail properties. These findings demonstrate a trade-off between choosing a *complex source density* vs. *expressive transformations*. Intuitively, following Corollary 1 it is clear that a Lipschitz map is appropriate to learn tails of a target density if both the source distribution and the target distribution belong to a family of densities that have equally heavy tails. However, if the two densities are from families with differing degree of heaviness then the transformation needs to be more expressive than a Lipschitz-continuous function. This choice of either using source densities with the same heaviness as the target, or deploying more expressive transformations than Lipschitz functions is what we refer to as the trade-off between choosing a *complex source density* vs. *expressive transformations*.

In practice, however, we do not know a priori the degree of heaviness of a target distribution to guide the choice of the source density accordingly. We circumvent this problem by proposing tail-adaptive flows (TAFs) wherein the tail property of the source density can be adapted during training such that simpler transformations like Lipschitz maps are able to capture heavy-tailed target distributions. In our approach, we propose to fix the source density as a standard student-t distribution with its degrees of freedom being a learnable parameter i.e. $\mathsf{Z} \sim t_\nu(0, \mathbf{I})$ where $\nu \in (1, \infty)$ is a learnable parameter. The source density becomes lighter tailed as ν increases and approaches a Gaussian distribution as $\nu \to \infty$. The source density is still tractable and hence we can learn the transport map $\mathbf{T}$ and degrees of freedom ν by maximizing the likelihood of the target density.

We thus formulate the density estimation paradigm for tail-adaptive flows as follows: Suppose we have access to an i.i.d. sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim q$ and our interest lies in estimating q and capturing its tail behaviour. Let $\mathcal{F}$ be a class of mappings and p_ν be the source density which is a

standard student-t distribution with ν degrees of freedom i.e. $p_\nu := t_\nu(0, \mathbf{I})$. The log-likelihood objective is:

$$\max_{\substack{\mathbf{T}\in\mathcal{F}\\ \nu\in(1,\infty)}} \frac{1}{n}\sum_{i=1}^{n}\Big[-\log|\mathbf{T}'(\mathbf{T}^{-1}\mathbf{x}_i)| + \log p_\nu(\mathbf{T}^{-1}\mathbf{x}_i)\Big],$$

where $\log p_\nu(\mathbf{T}^{-1}\mathbf{x}_i) = d\log\Gamma(\frac{\nu+1}{2}) - d\log\Gamma(\frac{\nu}{2}) - \frac{d}{2}\log\nu - \sum_{j=1}^{d}\frac{\nu+1}{2}\cdot\log\Big(1+\frac{z_{ij}^2}{\nu}\Big)$, $z_{ij} = T_j^{-1}(x_{ij})$ and $\Gamma(\cdot)$ is the gamma function. Tail-adaptive flows are easy to implement as they can be easily optimized using automatic differentiation. Further, they can be plugged-in any existing flow based learning framework to substitute the Gaussian density since the transformation $\mathbf{T}$ in the objective above can be from any family of functions. In our experiments we used tail-adaptive flows with Real-NVP, MAF, and SOS flows to illustrate its performance on inference tasks on real datasets and ability to capture tails with affine flows on synthetic datasets.

We first show that affine tail-adaptive flows can capture heavier tailed distributions. In Figure 4, we give the results for tail-adaptive flows using Real-NVP on the synthetic experiment we used in Section 5. We kept the set-up of the experiment exactly the same as before with the source distribution to be $t_\nu(0, \mathbf{I})$ and initialised $\nu = 30$. It is evident from the figure that tail-adaptive flows are able to capture the heavy-tails since the density quantiles of the target and estimated target overlap with $\gamma_{\text{source}} = 0.15$, $\gamma_{\text{target}} = 0.81$, and $\gamma_{\text{estimated}-\text{target}} = 0.80$.

Next, we considered another setting to test the performance of tail-adaptive flows where we fixed the target density to be a bi-variate Neal's funnel distribution given by $\mathbf{x}_i = (x_{1,i}, x_{2,i})$ where $x_{1,i} \sim \mathcal{N}(0, 1)$ and $x_{2,i} \sim \mathcal{N}(0, \exp(0.5x_{1,i}))$ and generated 10,000 samples from this distribution. We fixed the flow architecture to follow Real-NVP with $\lambda(\cdot) = \tanh(\cdot)$ and trained the model using Adam for 40 epochs with a batch size of 128 and learning rate of 10^{-3}. We learned tail-adaptive flows with two, three, and five blocks respectively and the results are given in Figure 5. Here we noticed that as the number of blocks increased, the estimated target density approximated the true target density more faithfully. Furthermore, we also noticed that the tails became heavier as the number of stacked blocks increased with $\gamma_{\text{source}} = 0.15$, $\gamma_{\text{target}} = 0.63$, and $\gamma_{\text{estimated}-\text{target},2} = 0.36$, $\gamma_{\text{estimated}-\text{target},3} = 0.56$, and $\gamma_{\text{estimated}-\text{target},5} = 0.61$.

Lastly, we replicate density estimation experiments on benchmark datasets popularly used to measure performance of flows and autoregressive models. Here we illustrate that tail-adaptive flows can be incorporated easily in existing architectures and achieve comparable performance on inference tasks. In Table 2, we report the performance of tail-adaptive flows using MAF (Papamakarios et al., 2017) and SOS (Jaini et al., 2019) keeping the architecture fixed as reported in the original papers but changing the source to tail adaptive ones. We compare the results to original implementations using Gaussian source density and other models like NAF (Huang et al., 2018), TAN (Oliva et al., 2018), and MADE (Germain et al., 2015).

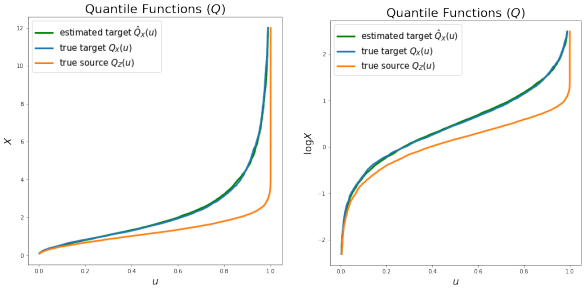

Figure 4. Quantile and log-Quantile plots for 5 block Real-NVP with tail adaptive flows on the same setup as in Figure 1. Best viewed in color.

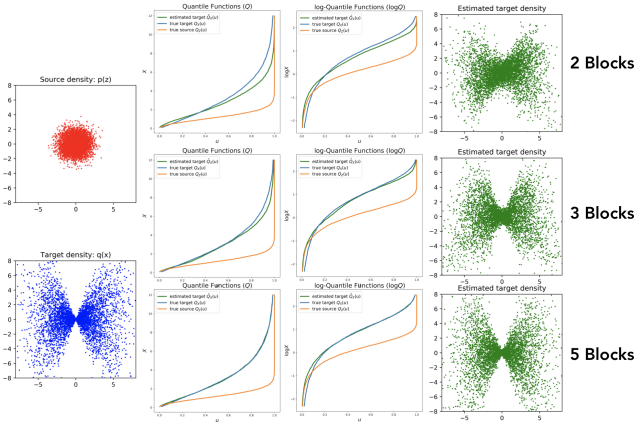

Figure 5. Results for tail-adaptive Real-NVP on Neal's funnel distribution as target density. Figure organization is same as Figure 1.

7. Conclusion

We studied the ability of popular flow models to capture tail-properties of a target density by studying the corresponding increasing triangular map approximated by these flow methods acting on a tractable source density with known fixed tails. We showed that any Lipschitz-continuous transport map cannot push a source density to a heavier target density, implying that affine flow models like Real-NVP, NICE, MAF, Glow etc. cannot capture heavier tails than the source density. We then propose tail-adaptive flows (TAFs) where the tails of the source density can be adapted during training. TAFs are appealing because they can be substituted easily in existing flow architectures and optimized using automatic differentiation. Further, their ability to adapt tails of the source density allows affine TAFs to learn heavier tailed distributions. In future work, we will be interesting to

explore the applications of TAFs for extreme value theory and financial risk analysis.

Acknowledgement

We thank Andy Keller, Didrik Nielsen, Jorn Peters, and Patrick Forre for discussions and feedback. We also thank the anonymous reviewers for their valuable feedback. We would also like to acknowledge NSERC, the Canada CIFAR AI Chairs Program, and MITACS Accelerate for financial support. We thank NVIDIA Corporation (the data science grant) for donating two Titan V GPUs that enabled in part the computation in this work. PJ was additionally supported by a Borealis AI fellowship and Huawei Graduate fellowship.

References

Andrews, D. et al. A general method for the approximation of tail areas. *The Annals of Statistics*, 1(2):367–372, 1973.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335, 2005.

Cambanis, S., Huang, S., and Simons, G. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

de Valk, C. and Segers, J. Tails of optimal transport plans for regularly varying probability measures. *arXiv preprint arXiv:1811.12061*, 2019.

Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear independent components estimation. In *ICLR workshop*, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using Real NVP. In *ICLR*, 2017.

Foss, S., Korshunov, D., Zachary, S., et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Frahm, G. *Generalized elliptical distributions: theory and applications*. PhD thesis, Universität zu Köln, 2004.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. MADE: Masked autoencoder for distribution estimation. In *ICML*, pp. 881–889, 2015.

Ghaffari, N. and Walker, S. On Multivariate Optimal Transportation. *arXiv preprint arXiv:1801.03516*, 2018.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural Autoregressive Flows. In *ICML*, 2018.

Jaini, P., Selby, K. A., and Yu, Y. Sum-of-Squares Polynomial Flow. *International Conference of Machine Learning (ICML)*, 2019.

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

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, pp. 4743–4751, 2016.

Knothe, H. et al. Contributions to the theory of convex bodies. *The Michigan Mathematical Journal*, 4(1):39–52, 1957.

Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *AIStats*, pp. 29–37, 2011.

Oliva, J., Dubey, A., Zaheer, M., Poczos, B., Salakhutdinov, R., Xing, E., and Schneider, J. Transformation Autoregressive Networks. In *ICML*, pp. 3898–3907, 2018.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. In *NeurIPS*, pp. 2338–2347, 2017.

Parzen, E. Nonparametric Statistical Data Modeling. *Journal of the American statistical association*, 74(365):105–121, 1979.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Rosenblatt, M. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

Spantini, A., Bigoni, D., and Marzouk, Y. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19:1–71, 2018.

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

Tabak, E. G. and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Uria, B., Côté, M.-A., Gregor, K., Murray, I., and Larochelle, H. Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17 (1):7184–7220, 2016.