Appendix

1 Preliminaries

In this section, we describe the general set up of triangular normalizing flow models.

1.1 Normalizing flows and triangular maps

NFs learn an invertible mapping between a prior and a more complex distribution (the target) in the same dimension. Typically, the prior is chosen to be a Gaussian with identity covariance or uniform on the unit cube, and the target is the one we intend to learn. Below, we present a summary of related ideas and refer the readers to Jaini et al. [5] and Kobyzev et al. [6] for a comprehensive discussion.

More formally, let $\mathbf{z}$ and $\mathbf{x}$ be sampled data from the prior with density P_z and the target distribution with density P_x, respectively. Then, NFs learn a transformation f such that $f(\mathbf{z}) = \mathbf{x}$ which is differentiable and invertible with a differentiable inverse. Such transformations are called diffeomorphisms and they allow the estimation of the probability density $P_x(\mathbf{x})$ via the change of variables formula $P_x(\mathbf{x}) = P_z(f^{-1}\mathbf{x})|\mathbf{J}_f(f^{-1}\mathbf{x})|^{-1}$ where $\mathbf{J}_f$ is the Jacobian determinant of f.

Given an independent and identically distributed (i.i.d.) sample $\{x_1,\dots,x_n\}$ with law P_x, learning the target density P_x and the transformation f (within an expressive function class $\mathfrak{F}$) is done simultaneously via minimizing the Kullback-Leibler (KL) divergence between P_x and the pushforward of P_z under f denoted by f_*P_z,

$$\min_{f\in\mathfrak{F}} \mathrm{KL}(P_x\|f_*P_z) = -\max_{f\in\mathfrak{F}} \int \log \frac{P_z(f^{-1}\mathbf{x})}{|\mathbf{J}_f(f^{-1}\mathbf{x})|} \cdot P_x(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

Density estimation using (1) requires efficient calculation of the Jacobian as well as f^{-1}. Both can be achieved via constraining f to be an *increasing triangular map*. That is, taking $P_x(\mathbf{x})$ to be a multivariate distribution where $\mathbf{x} = (x_1, x_2, \dots, x_d)$, and the prior $P_z(\mathbf{z})$ where $\mathbf{z} = (z_1, z_2, \dots z_d)$, the components of $\mathbf{x}$ are expressed as $x_j = f_j(z_1, z_2, \dots, z_j)$ for suitably defined transformations $f_j, j = 1,2,\dots,d$ where f_j is increasing with respect to z_j. From now on, we denote $(z_1, z_2, \dots, z_j)$ by $z_{<j+1}$. In this case, the Jacobian determinant is the product $\prod_{j=1}^{d} \partial_{z_j} f_j$. Also, because f_j is increasing in z_j, inversion can be done recursively starting from f_1^{-1}.

2 Proofs

Proof of Theorem 1. To illustrate the relevance of the theorem to our setting, we write down the details of the proof assuming that the learnable class $\mathfrak{F}$ is Bernstein-type polynomials. The same proof is true for any class of functions.

Take $I_j = [a_j, b_j]$ for $j = 1,2,3$ with $a_j < b_j$ with the possibility that $a_1 = -\infty$ or $b_1 = \infty$ whence the interval is understood to be open on the infinite end and the target may have non-compact support. Let $B_n : I_2 \to I_3$ be the learnable Bernstein-type polynomial with coeffecients $\{\alpha_j\}_{j=0}^{n}$. Let $h : I_3 \to I_1$ be a fixed invertible transformation so that h^{-1} transforms the target density P_x to P_y supported on I_3, i.e., $h_*P_y = P_x$, and

$$P_x(\mathbf{x}) = \frac{P_y(h^{-1}\mathbf{x})}{|\mathbf{J}_h(h^{-1}\mathbf{x})|}. \tag{2}$$

Fix $\alpha_0 = a_3$, $\alpha_n = b_3$ and let $I = \{(\alpha_1, \ldots, \alpha_{n-1}) \,|\, a_3 < \alpha_1 < \cdots < \alpha_{n-1} < b_3\}$. Then the optimization problem is

$$\min_I \mathrm{KL}(P_x \| (h \circ B_n)_* P_z) \tag{3}$$

$$= -\max_I \int \log \frac{P_z(B_n^{-1}(h^{-1}\mathbf{x}))}{|\mathbf{J}_{h\circ B_n}(B_n^{-1}(h^{-1}\mathbf{x}))|} \cdot P_x(\mathbf{x})\, d\mathbf{x} \tag{4}$$

$$= -\max_I \int \log \frac{P_z(B_n^{-1}(h^{-1}\mathbf{x}))}{|\mathbf{J}_h(h^{-1}\mathbf{x})\mathbf{J}_{B_n}(B_n^{-1}(h^{-1}\mathbf{x}))|} \cdot P_x(\mathbf{x})\, d\mathbf{x} \tag{5}$$

$$= -\max_I \left\{ \int \log \frac{P_z(B_n^{-1}(h^{-1}\mathbf{x}))}{|\mathbf{J}_{B_n}(B_n^{-1}(h^{-1}\mathbf{x}))|} \cdot P_x(\mathbf{x})\, d\mathbf{x} + \int \log |\mathbf{J}_h(h^{-1}\mathbf{x})| \cdot P_x(\mathbf{x})\, d\mathbf{x}. \right\} \tag{6}$$

$$= -\max_I \left\{ \int \log \frac{P_z(B_n^{-1}(h^{-1}\mathbf{x}))}{|\mathbf{J}_{B_n}(B_n^{-1}(h^{-1}\mathbf{x}))|} \cdot \frac{P_y(h^{-1}\mathbf{x})}{|\mathbf{J}_h(h^{-1}\mathbf{x})|}\, d\mathbf{x} \right\} + \int \log |\mathbf{J}_h(h^{-1}\mathbf{x})| \cdot \frac{P_y(h^{-1}\mathbf{x})}{|\mathbf{J}_h(h^{-1}\mathbf{x})|} d\mathbf{x}. \tag{7}$$

$$= -\max_I \left\{ \int \log \frac{P_z(B_n^{-1}(\mathbf{y}))}{|\mathbf{J}_{B_n}(B_n^{-1}(\mathbf{y}))|} \cdot P_y(\mathbf{y})\, d\mathbf{y} \right\} + \int \log |\mathbf{J}_h(\mathbf{y})| \cdot P_y(\mathbf{y})\, d\mathbf{x}. \tag{8}$$

$$= \min_I \mathrm{KL}(P_y \| (B_n)_* P_z) + \int \log |\mathbf{J}_h(\mathbf{y})| \cdot P_y(\mathbf{y})\, d\mathbf{x} \tag{9}$$

Note that the second integral in (6) can be taken outside the max because it is independent of B_n, and hence, it becomes a constant that is irrelevant for the optimization. From (9), it follows that the minimum of $\mathrm{KL}(P_x \| (h \circ B_n)_* P_z)$ is achieved if and only if the minimum of $\mathrm{KL}(P_y \| (B_n)_* P_z)$ is achieved. Hence,

$$\arg\min_I \mathrm{KL}(P_x \| (h \circ B_n)_* P_z) = \arg\min_I \mathrm{KL}(P_y \| (B_n)_* P_z) \tag{10}$$

as required. It is easy to see that this argument remains unchanged when B_n is replaced by f and I is replaced by $f \in \mathfrak{F}$. □

Proof of Theorem 2. There is a probabilistic interpretation of Bernstein polynomials that makes the analysis easier. Let Z_i^x, $0 \le i \le n$ be i.i.d. Bernoulli(x) random variables. Then

$$B_n(x) = \mathsf{E}\left(f\left(\sum_{i=0}^n Z_i^x / n \right) \right). \tag{11}$$

See, for example, Chapter 2 of Bustamante [3]. We will use this definition in the proof.

Let $f : [0,1] \to \mathsf{R}$ be a strictly increasing continuous function such that $f(k/n) = \alpha_k$. Let $s < t$ and let Z_i^x, $0 \le i \le n$ and be a sequence of iid Bernoulli(x) for $x = s, t$, defined on the same probability space such that $Z_i^s \le Z_i^t$ via monotone coupling. That is, let $Z_i^s = \mathbf{1}_{U \le s}$ and $Z_i^t = \mathbf{1}_{U \le t}$ where U is a uniform random variables on $[0,1]$ and couple them as follows.

$$\mathsf{P}((Z_i^s, Z_i^t) = (j,k))_{j,k \in \{0,1\}} = \begin{pmatrix} 1-t & t-s \\ 0 & s \end{pmatrix} \tag{12}$$

and $\mathsf{P}(Z_i^s > Z_i^t) = \mathsf{P}(Z_i^s = 1, Z_i^t = 0) = 0$. So, $Z_i^t \ge Z_i^s$ as required.

Then

$$f\left(\sum_{i=0}^n Z_i^s / n \right) \le f\left(\sum_{i=0}^n Z_i^t / n \right). \tag{13}$$

Consequently,

$$\mathsf{E}\left(f\left(\sum_{i=0}^{n} Z_i^s/n\right)\right) \leq \mathsf{E}\left(f\left(\sum_{i=0}^{n} Z_i^t/n\right)\right). \tag{14}$$

Due to (11), this is equivalent to $B_n(s) \leq B_n(t)$.

If (14) is not strict, then $f(\sum_{i=0}^{n} Z_i^t/n) = f(\sum_{i=0}^{n} Z_i^s/n)$ almost surely, and therefore, $\sum_{i=0}^{n} Z_i^t = \sum_{i=0}^{n} Z_i^s$ almost surely. But this is impossible due to monotone coupling. Therefore, by contradiction, (14) is strict as required. □

Proof of Theorem 3. Recall From Bernstein [2] that B_ns are uniformly dense in the space of continuous function on $[0,1]$ because $B_n(f) \to f$ uniformly. By rescaling, this is true on any interval $[a,b]$. Moreover, by construction, whenever f is increasing, $B_n(f)$ is increasing. So, it is automatic that increasing Bernstein polynomials on $[a,b]$ are uniformly dense in the space of increasing continuous functions on $[a,b]$. Finally, to show true universality, we have to show that any increasing continuous function $f : \mathsf{R} \to \mathsf{R}$ is well-approximated by B_ns.

Given $f : \mathsf{R} \to \mathsf{R}$ continuous and increasing, choose two positive sequences $\{M_n\}$ and $\{\varepsilon_n\}$ such that $M_n \to \infty$ and $\varepsilon_n \to 0$. Let $I_n = [-M_n, M_n]$. Then, there exists a Bernstein approximation of f, say q_n, which is increasing on I_n (which can be monotonically extended to R) such that

$$\max_{I_n} |f - q_n| \leq \varepsilon_n. \tag{15}$$

Then the sequence of Bernstein approximations $\{q_n\}$ converges point-wise to f on R, and this convergence is uniform on each compact interval. □

Remark 1. *We can write down a sequence q_n explicitly when f is regular. For example, when f is C^3 with bounded derivatives and $M_n = \log n$, choosing the degree of q_n to be n is sufficient because it follows from the error estimate in Section 4.3 that $\varepsilon_n \sim (\log n)/n$ works. That is, choose q_n to be the degree n Bernstein approximation of f on $[-M_n, M_n]$.*

Remark 2. *Note that this result is not a restatement of the original result in [2]. The latter is about Bernstein-type polynomials being uniformly dense in the space of continuous functions on a compact interval. It uses the fact that such functions have a maximum. For the universality of NFs, we need that given an increasing continuous function on the real line (which is noncompact and hence, no guarantee of a maximum) there is a sequence of Bernstein-type polynomials that converge (at least, pointwise) to it.*

3 Universality and the explicit rate of convergence

The basis of all the universality proofs of NFs in the existing literature is that the learnable class of functions is dense in the class of increasing continuous functions. In contrast, the argument we present here is constructive. As a result, we can write down sequences of approximations for (known) transformations between densities; see Section 4.

In the case of cubic-spline NFs of Durkan et al. [4], it is known that for $k = 1, 2, 3$ and 4, when the transformation is k times continuously differentiable and the bin size is h, the error is $O(h^k)$ [1, Chapter 2]. However, we are not aware of any other instance where an error bound is available. Fortunately for us, the error of approximation of a function f by its Bernstein polynomials has been extensively studied. We recall from Voronovskaya [9] the following error bound: for $f : [0,1] \to \mathsf{R}$ twice continuously differentiable

$$B_n(f) - f = \frac{x(1-x)}{2n} f''(x) + o(n^{-1}). \tag{16}$$

and this holds for an arbitrary interval $[a,b]$ with $x(1-x)$ replaced by $(x-a)(b-x)$.

Since the error estimate is given in terms of the degree of the polynomials used, we can improve the optimality of our NF by avoiding unnecessarily high degree polynomials. This allows us to keep the number of trainable parameters under control in our NF model. The following example shows that the error $O(n^{-1})$ above does not necessarily improve when SOS polynomials are used instead.

Example 1. Uniform$[0,1]$ *to the* Normal$(0,1)$ *: There is bounded* $\{c_k\}_{k\geq 0} \subset \mathsf{R}_+$ *such that*

$$f(z) = \mathrm{Erf}^{-1}(2z-1) = \sum_{k=0}^{\infty} \frac{\sqrt{2}\pi^{k+\frac{1}{2}} c_k}{2k+1} \left(z-\frac{1}{2}\right)^{2k+1}; \tag{17}$$

see Jaini et al. [5]. *This is the power series expansion of* f *at* $z = 1/2$*, and hence, it is unique. The SOS approximation of* f *(the series above truncated at* $k = n$*) is only* $O((2n+1)^{-1}) = O(n^{-1})$ *accurate on compact sub-intervals of* $(0,1)$*. This is precisely the accuracy one would expect from the degree* $2n+1$ *Bernstein approximation on any compact subinterval of* $(0,1)$*.*

In our NF, at each step, the estimation is done using a univariate polynomial, and hence, the overall convergence rate is, in fact, the minimal univariate convergence rate of $O(n^{-1})$ (equivalently, the error upper bound is the maximum of univariate upper bounds), and in general, cannot be improved further regardless of how regular the density transformation is. However, our experiments show that our model on average has a significantly smaller error than the given theoretical upper-bound.

4 Examples of Bernstein-type approximations

In this section, we illustrate how to use Bernstein-type polynomials to approximate diffeomorphisms between densities. We restrict our attention to densities on R. Suppose F and G are the distribution functions of the two probability densities P_z and P_x on R. Then the *increasing rearrangement* $f = G^{-1} \circ F$ is the unique increasing transformation that pushes forward P_z to P_x, and this generalizes to higher dimensions [8, Chapter 1]. Now, we can explicitly write down their degree$-n$ Bernstein-type approximations, $B_n(f)$ along with convergence rates.

Example 2. Uniform$[0,1]$ *to a continuous and non-zero density* P *on* $[0,1]$ *: Note that* $G(x) = \int_0^x P(s)\,ds$*,* $x \in [0,1]$ *is strictly increasing and hence, invertible on* $[0,1]$*. So,* $f(x) = G^{-1}(x)$*, and* G^{-1} *is once continuously differentiable. Then*

$$B_n(f)(x) = \sum_{k=0}^{n} G^{-1}\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}. \tag{18}$$

and $\|B_n(f) - f\|_\infty = O(n^{-1/2})$.

Example 3. Kumaraswamy(α,β) *to* Uniform$[0,1]$ *: Here,* $\alpha,\beta > 0$ *and for* $x \in [0,1]$*,* $F(x) = 1-(1-x^\alpha)^\beta$ [7] *and* $G(x) = x$*. Therefore,* $f(x) = F(x)$*. Then*

$$B_n(f)(x) = \sum_{k=0}^{n} F\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}. \tag{19}$$

When $\alpha,\beta \geq 1$*,* $\|B_n(f) - f\| = O(n^{-1})$.

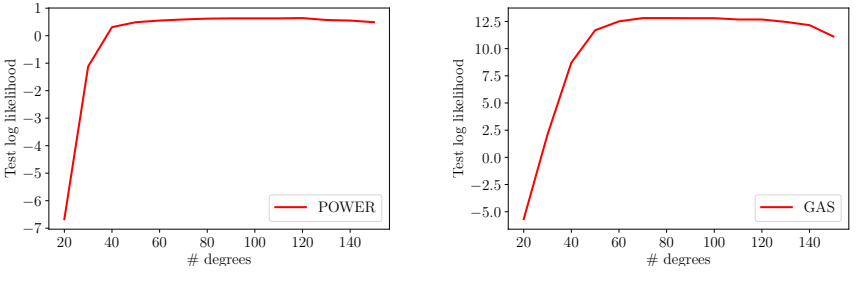
5 Hyper-parameters and training details

For optimization, we used the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_1 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, where parameters refer to the usual notation. An initial learning rate of 0.01 was used for updating the weights with a decay factor of 10% per 50 iterations. We initialized all the trainable weights randomly from a standard normal distribution and used maximum likelihood as the objective function for training. We observed that a single layer model with 100 degree polynomials performed well for the real-world data.

In contrast, for 2D toy distributions and and images we used higher number of layers (8) with 15 degree polynomials in each layer. For all the experiments, we use a Kumaraswamy distribution with parameters $a = 2$ and $b = 5$ as the base density. Using a standard normal distribution after converting it to a density on $[0, 1]$ using a nonlinear transformation, e.g., $\frac{1+\tanh(z)}{2}$, also yielded similar results.

6 Training stability for higher degree polynomials

Typically, polynomial-based models such as SOS yield training instability as their target ranges are not compact. This is because higher degree approximations could increase the range of outputs without bound, and in turn cause gradients to explode while training. As a solution, they opt to use a higher number of layers with lower degree polynomials. In contrast, our model can entertain higher degree approximations without any instability which allows more design choices. Figure 1 demonstrates this behavior experimentally.



(a) Power (b) Gas

Figure 1: Test log-likelihood against the number of degrees used for the Bernstein approximation in a single layer model on POWER and GAS datasets. Slight dip in the performance for degrees 100+ but shows no training instability.

According to Figure 1, the model hits a peak in performance at a certain degree and shows a slight drop in performance at higher degrees. Nevertheless, the model does not exhibit unstable behavior at higher degrees as opposed to SOS-flows – an indication of the superior training stability of our model. This further illustrates that our model provides the option to design shallow models by increasing the number of degrees in the polynomials instead of deeper models with a higher number of layers.

7 Ablation study

We compare the performance of different variants of our model against a simple task in order to better understand the design choices. For this, we use a standard normal as the base distribution, and a mixture of five Gaussians with means $= (-5, -2, 0, 2, 5)$, variances $= (1.5, 2, 1, 2, 1)$, and weights 0.2 each, as the target. Figure 2 depicts the results.

Clearly, we were able to increase the expressiveness of the transformation by increasing the degree of the polynomials, as well as the number of layers. However, it is also visible that using an unnecessarily higher degree over-parametrizes the model, and hence, deteriorate the output. As discussed in the main article and in Section 6, we are able to use polynomials with degree as high as 100 in this experiment and others with no cost to the training stability because the training is done for a compactly supported target.

We also examine how the initial base distribution affects the performance. We use a mixture of seven Gaussians with means $= (-7, -5, -2, 0, 2, 5, 7)$, variances $= (1, 1, 2, 2, 2, 1, 1)$, and weights $= (0.8, 0.2, 0.2, 0.6, 0.2, 0.2, 0.8)$, as the target. We used a model with a 100-degree polynomial and a single layer for this experiment. Figure 3 illustrates the results. Although all priors capture the multimodes, when Uniform$[0, 1]$ is used the model was not able to predict that the density is almost zero for large negative values.

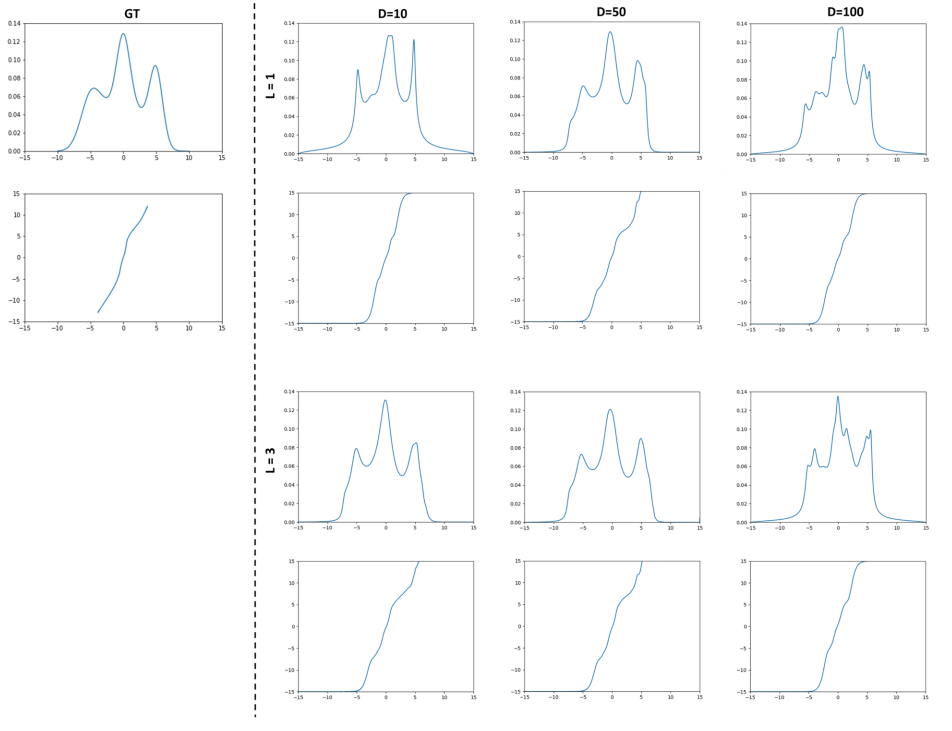


Figure 2: Ablation study with different varients of our model. **D** and **L** denotes the degree of the used polynomials and the number of layers, respectively. Corresponding transformation functions are also shown below the predicted densities.

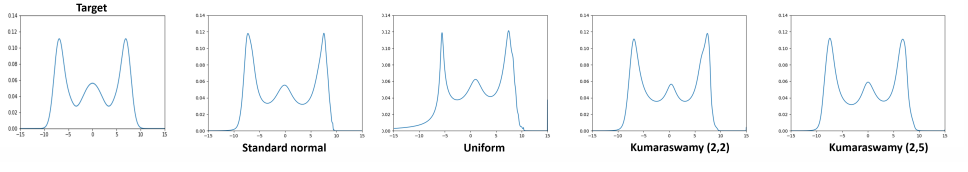


Figure 3: Approximation of the target density starting from various initial densities (the initial distributions are noted below the densities).