Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows: Supplementary Material

A Guide to Notation

(a_n)	A sequence of elements $a_1, a_2, \ldots$
$a(n) = \Theta(b(n))$	$a(n)$ differs from $b(n)$ by at most a constant factor as $n \to \infty$
$u \odot v$	The elementwise product of tensors u and v
$\mathrm{LogSumExp}(a_1, \ldots, a_m)$	$\log\left(\sum_{i=1}^{m} \exp(a_i)\right)$
e^v, where $v \in \mathbb{R}^d$	$(e^{v_1}, \ldots, e^{v_d})$
$\\|v\\|$	The norm of a vector $v \in \mathbb{R}^d$ (our results are agnostic to the specific choice of $\\|\cdot\\|$)
$\\|A\\|_{\mathrm{op}}$	The operator norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ induced by $\\|\cdot\\|$
I_d	The $d \times d$ identity matrix
$\det A$	The determinant of a square matrix A
$\mathrm{D}f(z)$	The Jacobian matrix of a function f evaluated at z
$\mathrm{D}F(z; u)$	The Jacobian matrix of a function $\mathrm{D}F(\cdot; u)$ (i.e. with u fixed) evaluated at z
$\mathrm{Lip}\, f$	The Lipschitz constant of a function f
$\mathrm{BiLip}\, f$	The bi-Lipschitz constant of a function f
$\mathcal{A} \cong \mathcal{B}$	The topological spaces $\mathcal{A}$ and $\mathcal{B}$ are homeomorphic
$\overline{B}$	The topological closure of a set B
$\mathrm{int}(B)$	The interior of a set B
∂B	The boundary of a set B
$\mathrm{supp}\, \mu$	The support of a measure μ
$f\#\mu$	The pushforward of a measure μ by a function f
$\mu_n \xrightarrow{\mathcal{D}} \mu$	Weak convergence of the measures μ_n to μ

B Proofs

B.1 Preliminaries

We require some basic results that we include here for completeness. We will make use of standard definitions and results from topology and real analysis. A complete background to these topics can be found in Dudley (2002).

B.1.1 Supports of Measures

Recall that for a Borel measure μ on a topological space $\mathcal{Z}$, the *support* of μ, denoted $\mathrm{supp}\,\mu$, is the set of all $z \in \mathcal{Z}$ such that $\mu(N_z) > 0$ for every open set N_z containing z.

The following is an immediate consequence:

Proposition B.1. *Suppose μ and ν are Borel measures with μ absolutely continuous with respect to ν. Then*

$$\mathrm{supp}\,\mu \subseteq \mathrm{supp}\,\nu.$$

Proof. Suppose $z \notin \mathrm{supp}\,\nu$. Then there exists an open set N_z containing z such that $\nu(N_z) = 0$. By absolute continuity, we have also that $\mu(N_z) = 0$ and hence $z \notin \mathrm{supp}\,\mu$. $\square$

In general the converse need not hold. For example, the Dirac measure on 0 has support contained within the Lebesgue measure on $\mathbb{R}$ (which has full support), but is not absolutely continuous with respect to it.

The following characterisation is useful:

Proposition B.2. *For any Borel measure μ,*

$$(\operatorname{supp}\mu)^c = \bigcup_{\substack{A \text{ open:} \\ \mu(A)=0}} A, \tag{B.1}$$

and hence $\operatorname{supp}\mu$ is closed.

Proof. This follows directly from the definitions, since $z \notin \operatorname{supp}\mu$ if and only if there exists open N_z with $z \in N_z$ and $\mu(N_z) = 0$, which is just another way of saying that z is contained in the right-hand side of (B.1). It follows that $(\operatorname{supp}\mu)^c$ is open, and hence $\operatorname{supp}\mu$ is closed. $\square$

We mainly care about how the support of a measure is transformed by a pushforward function. The following proposition characterises what occurs in this case.

Proposition B.3. *Suppose $\mathcal{Z}$ and $\mathcal{X}$ are topological spaces. If μ is a Borel measure on $\mathcal{Z}$ such that $\mu((\operatorname{supp}\mu)^c) = 0$, and if $f : \mathcal{Z} \to \mathcal{X}$ is continuous, then*

$$\operatorname{supp} f\#\mu = \overline{f(\operatorname{supp}\mu)}.$$

Proof. Suppose $x \notin \overline{f(\operatorname{supp}\mu)}$. Then x must have an open neighbourhood N_x such that

$$N_x \cap f(\operatorname{supp}\mu) = \emptyset.$$

This implies

$$\begin{aligned} f^{-1}(N_x) \cap \operatorname{supp}\mu &\subseteq f^{-1}(N_x) \cap f^{-1}(f(\operatorname{supp}\mu)) \\ &= f^{-1}(N_x \cap f(\operatorname{supp}\mu)) \\ &= f^{-1}(\emptyset) \\ &= \emptyset. \end{aligned}$$

We then have

$$f\#\mu(N_x) = \mu(f^{-1}(N_x)) = \mu(f^{-1}(N_x) \cap \operatorname{supp}\mu) = 0,$$

where the second equality follows since we assumed $\mu((\operatorname{supp}\mu)^c) = 0$, and hence $x \notin \operatorname{supp} f\#\mu$. Consequently

$$\operatorname{supp} f\#\mu \subseteq \overline{f(\operatorname{supp}\mu)}.$$

In the other direction, suppose $x \in f(\operatorname{supp}\mu)$, so that $x = f(z)$ for some $z \in \operatorname{supp}\mu$. Given an open neighbourhood N_x it then follows from continuity that $f^{-1}(N_x)$ is an open neighbourhood of z, and so

$$f\#\mu(N_x) = \mu(f^{-1}(N_x)) > 0$$

since $z \in \operatorname{supp}\mu$. This entails $\operatorname{supp} f\#\mu \supseteq f(\operatorname{supp}\mu)$, which means

$$\operatorname{supp} f\#\mu = \overline{\operatorname{supp} f\#\mu} \supseteq \overline{f(\operatorname{supp}\mu)}$$

by Proposition B.2. $\square$

Note that in general we need not have $\operatorname{supp} f\#\mu = f(\operatorname{supp}\mu)$. For example, if μ is Gaussian and $f = \arctan$, then

$$f(\operatorname{supp}\mu) = (-1, 1) \neq [-1, 1] = \operatorname{supp} f\#\mu.$$

Likewise, in general we do require the assumption $\mu((\operatorname{supp}\mu)^c) = 0$. This is because there exist examples of nontrivial Borel measures μ such that $\operatorname{supp}\mu = \emptyset$. Taking $f \equiv x_0$ to be any constant $x_0 \in \mathcal{X}$ (in which case f is certainly continuous) then gives

$$\overline{f(\operatorname{supp}\mu)} = \emptyset \neq \{x_0\} = \operatorname{supp} f\#\mu.$$

However, for our purposes, the following proposition shows that this is not a restriction.

Proposition B.4. *Suppose μ is a Borel measure on a separable metric space $\mathcal{Z}$. Then*

$$\mu((\operatorname{supp}\mu)^c) = 0.$$

Proof. Throughout the proof, for each z and $r > 0$, we will denote by $B(z, r)$ an open ball of radius r centered at z. Likewise, for each $z \notin \operatorname{supp}\mu$, let

$$r^\star(z) \coloneqq \sup\{r > 0 \mid \mu(B(z, r)) = 0\}.$$

Observe that $r^\star$ is well-defined (but possibly infinite) since $z \notin \operatorname{supp}\mu$ means there must exist some $r > 0$ such that $\mu(B(z, r)) = 0$.

We first show that $\mu(B(z, r^\star(z))) = 0$ for all $z \notin \operatorname{supp}\mu$. To this end, fix z and choose a sequence $r_m \uparrow r^\star(z)$ with $r_m < r^\star(z)$. We then have

$$B(z, r^\star(z)) = \bigcup_{m=1}^{\infty} B(z, r_m),$$

and so

$$\mu(B(z, r^\star(z))) = \lim_{m\to\infty} \mu(B(z, r_m)) = 0$$

by continuity of measure.

Now, by separability, we can choose a countable sequence $(z_k) \subseteq (\operatorname{supp}\mu)^c$ such that $\overline{\{z_k\}} = \overline{(\operatorname{supp}\mu)^c}$. We show that

$$(\operatorname{supp}\mu)^c = \bigcup_{k=1}^{\infty} B(z_k, r^\star(z_k)),$$

from which the result follows by countable subadditivity. It is clear from (B.1) that the left-hand side is a superset of the right. In the other direction, let $z \in (\operatorname{supp}\mu)^c$. By construction of (z_k), there exists a subsequence $(z_{k'})$ such that $z_{k'} \to z$. For all k' large enough we then have $z_{k'} \in B(z, r^\star(z)/2)$ and hence

$$B(z_{k'}, r^\star(z)/2) \subseteq B(z, r^\star(z))$$

by triangle inequality. It follows that for such k' we have

$$\mu(B(z_{k'}, r^\star(z)/2)) \le \mu(B(z, r^\star(z))) = 0,$$

and so $r^\star(z_{k'}) \ge r^\star(z)/2$ since $r^\star(z_{k'})$ is the supremum. But then we have

$$z \in B(z_{k'}, r^\star(z)/2) \subseteq B(z_{k'}, r^\star(z_{k'})),$$

so that

$$z \in \bigcup_{k=1}^{\infty} B(z_k, r^\star(z_k))$$

and we are done. □

B.2 Lipschitz and Bi-Lipschitz Functions

We assume that $\mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}}$, $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$, and $f : \mathcal{Z} \to \mathcal{X}$. Recall that the *Lipschitz* constant of f, denoted $\operatorname{Lip} f$, is defined as the infimum over $M \in [0, \infty]$ such that

$$\|f(z) - f(z')\| \le M\|z - z'\|$$

for all $z, z' \in \mathcal{Z}$. Likewise the *bi-Lipschitz* constant $\operatorname{BiLip} f$ is defined as the infimum over $M \in [1, \infty]$ such that

$$M^{-1}\|z - z'\| \le \|f(z) - f(z')\| \le M\|z - z'\|$$

for all $z, z' \in \mathcal{Z}$. We prove some basic properties that follow from this definition.

Proposition B.5. $\operatorname{BiLip} f < \infty$ *if and only if f is injective and* $\max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}) < \infty$, *where $f^{-1} : f(\mathcal{Z}) \to \mathcal{Z}$. For all injective f, we then have* $\operatorname{BiLip} f = \max(\operatorname{Lip} f, \operatorname{Lip} f^{-1})$.

Proof. For the first statement, suppose $\mathrm{BiLip}\, f < \infty$. It is immediate that $\mathrm{BiLip}\, f \geq \mathrm{Lip}\, f$. To see that f is injective, note that for $z \neq z'$ we have

$$\|f(z) - f(z')\| \geq (\mathrm{BiLip}\, f)^{-1} \|z - z'\| > 0$$

and so $f(z) \neq f(z')$. On the other hand, for $x, x' \in f(\mathcal{Z})$, we have

$$(\mathrm{BiLip}\, f)^{-1} \|f^{-1}(x) - f^{-1}(x')\| \leq \|f(f^{-1}(x)) - f(f^{-1}(x'))\| = \|x - x'\|,$$

which gives that $\mathrm{BiLip}\, f \geq \mathrm{Lip}\, f^{-1}$. Altogether we have

$$\max(\mathrm{Lip}\, f, \mathrm{Lip}\, f^{-1}) \leq \mathrm{BiLip}\, f < \infty, \tag{B.2}$$

which gives the forward direction.

Next suppose f is injective and that

$$M := \max(\mathrm{Lip}\, f, \mathrm{Lip}\, f^{-1}) < \infty.$$

For $z, z' \in \mathcal{Z}$, we certainly have

$$\|f(z) - f(z')\| \leq M \|z - z'\|.$$

Likewise, since $f(z), f(z') \in f(\mathcal{Z})$,

$$\|z - z'\| = \|f^{-1}(f(z)) - f^{-1}(f(z'))\| \leq M \|f(z) - f(z')\|,$$

so that

$$M^{-1} \|z - z'\| \leq \|f(z) - f(z')\|$$

because injectivity of f means that $M > 0$. From this it follows that

$$\mathrm{BiLip}\, f \leq M < \infty, \tag{B.3}$$

which gives the reverse direction, proving the first statement.

For the second statement, suppose f is injective. Then if $\mathrm{BiLip}\, f < \infty$, (B.2) and (B.3) together give

$$\mathrm{BiLip}\, f = \max(\mathrm{Lip}\, f, \mathrm{Lip}\, f^{-1}).$$

On the other hand, if $\mathrm{BiLip}\, f = \infty$ then $\max(\mathrm{Lip}\, f, \mathrm{Lip}\, f^{-1}) = \infty$ since we would otherwise obtain a contradiction by the first statement of the proposition. This completes the proof. $\square$

It follows directly that if $\mathrm{BiLip}\, f < \infty$, then f is a homeomorphism from $\mathcal{Z}$ to $f(\mathcal{Z})$.[14] Moreover, in this case f maps closed sets to closed sets, as the following result shows:

Proposition B.6. *If* $\mathrm{BiLip}\, f < \infty$ *and* $\mathcal{Z}$ *is closed in* $\mathbb{R}^{d_\mathcal{Z}}$*, then* $f(\mathcal{Z})$ *is closed in* $\mathbb{R}^{d_\mathcal{X}}$*.*

Proof. It is a straightforward consequence of Proposition B.5 that if $(x_n) \subseteq f(\mathcal{Z})$ is Cauchy, then $(f^{-1}(x_n))$ is Cauchy. Consequently $(f^{-1}(x_n))$ converges to some $z_\infty \in \mathcal{Z}$, since $\mathcal{Z}$ is a closed subset of a complete space and therefore complete. But then

$$\begin{aligned}
\|x_n - f(z_\infty)\| &= \|f(f^{-1}(x_n)) - f(z_\infty)\| \\
&\leq M \|f^{-1}(x_n) - z_\infty\| \\
&\to 0
\end{aligned}$$

as $n \to \infty$. Consequently $f(\mathcal{Z})$ is complete, and so $f(\mathcal{Z})$ is closed as desired since the ambient space $\mathbb{R}^{d_\mathcal{X}}$ is complete. $\square$

The Lipschitz constant can be computed from the *operator norm* $\|\cdot\|_{\mathrm{op}}$ of the Jacobian of f. Recall that $\|\cdot\|_{\mathrm{op}}$ is defined as for a matrix $A \in \mathbb{R}^{d_\mathcal{X} \times d_\mathcal{Z}}$ as

$$\|A\|_{\mathrm{op}} := \sup_{\substack{v \in \mathbb{R}^{d_\mathcal{Z}}: \\ \|v\|=1}} \|Av\|$$

where we think of elements of $\mathbb{R}^{d_\mathcal{Z}}$ as column vectors.

[14] Note however that the converse is not true in general: for example, $\exp$ is a homeomorphism from $\mathbb{R}$ to $(0, \infty)$, but $\mathrm{BiLip} \exp = \infty$.

Proposition B.7. *If $\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$, $\mathcal{X} = \mathbb{R}^{d_{\mathcal{X}}}$, and f is everywhere differentiable, then*

$$\operatorname{Lip} f = \sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}.$$

Proof. If $v \in \mathcal{Z}$ with $\|v\| = 1$, then

$$\begin{aligned}
\|[\mathrm{D}f(z)]v\| &= \lim_{t \to 0} \frac{\|f(z+tv) - f(z)\|}{|t|} \\
&\leq \lim_{t \to 0} \frac{(\operatorname{Lip} f)\|(z+tv) - z\|}{|t|} \\
&= \operatorname{Lip} f.
\end{aligned}$$

It follows directly that

$$\|\mathrm{D}f(z)\|_{\mathrm{op}} \leq \operatorname{Lip} f.$$

On the other hand, suppose $\operatorname{Lip} f > M$. Then there exists $z, z' \in \mathcal{Z}$ such that

$$\|f(z) - f(z')\| > M\|z - z'\|.$$

Since f is differentiable, so too is the map $\varphi : [0,1] \to \mathcal{X}$ defined by

$$\varphi(t) \coloneqq f(tz + (1-t)z').$$

By Theorem 5.19 of Rudin (1964), there exists $t_0 \in (0,1)$ such that the derivative φ' satisfies

$$\|\varphi'(t_0)\| \geq \|f(z') - f(z)\| > M\|z - z'\|.$$

But, letting $z_0 \coloneqq t_0 z + (1-t_0)z'$, observe that

$$\begin{aligned}
\varphi'(t_0) &= \lim_{t \to 0} \frac{f(z_0 + t(z - z')) - f(z_0)}{t} \\
&= [\mathrm{D}f(z_0)](z - z'),
\end{aligned}$$

where we think of z, z' as column vectors. As such,

$$\begin{aligned}
\|\mathrm{D}f(z_0)\|_{\mathrm{op}}\|z - z'\| &\geq \|[\mathrm{D}f(z_0)](z - z')\| \\
&= \|\varphi'(t_0)\| \\
&> M\|z - z'\|
\end{aligned}$$

and so

$$\sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}} > M.$$

Since M was arbitrary this means that

$$\operatorname{Lip} f \leq \sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}$$

which gives the result. □

Proposition B.5 and Proposition B.7 then immediately entail the following:

Corollary B.8. *Suppose $\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$ and $\mathcal{X} = \mathbb{R}^{d_{\mathcal{X}}}$. If f is injective, and if f and $f^{-1} : f(\mathcal{Z}) \to \mathcal{Z}$ are everywhere differentiable, then*

$$\operatorname{BiLip} f = \max\left(\sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}, \sup_{x \in f(\mathcal{Z})} \|\mathrm{D}f^{-1}(x)\|_{\mathrm{op}} \right).$$

B.2.1 ARZELÀ-ASCOLI

Our proof of Theorem 2.1 makes use of the Arzelà-Ascoli theorem. This is a standard and foundational result in analysis, but we include a statement here for completeness. To this end, suppose we have a sequence of functions $f_n : \mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{X}} \to \mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$. We say that (f_n) is *pointwise bounded* if, for all $z \in \mathcal{Z}$,

$$\sup_n \|f_n(z)\| < \infty.$$

Likewise, (f_n) is *uniformly equicontinuous* if for every $\epsilon > 0$ there exists $\delta > 0$ such that, for all n,

$$\|f_n(z) - f_n(z')\| < \epsilon$$

whenever $\|z - z'\| < \delta$.

Theorem B.9 (Arzelà-Ascoli)**.** *If a sequence of functions $f_n : \mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}} \to \mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ is pointwise bounded and uniformly equicontinuous, then there exists a subsequence of (f_n) that converges uniformly on every compact subset of $\mathcal{Z}$.*

Proof. The case $d = 1$ is proven for example by Rudin (2006, Theorem 11.28). This can be extended to the case $d > 1$ by a standard argument. In particular, write

$$f_n =: (f_{n,1}, \dots, f_{n,d}),$$

where $f_{n,i} : \mathcal{Z} \to \mathbb{R}$. Then extract a subsequence (f_{n_1}) of (f_n) such that $f_{n,1}$ converges uniformly on every compact subset of $\mathcal{Z}$. Then extract a subsequence of (f_{n_1}) such that the same holds for $f_{n,2}$, and so on. The result is a subsequence $(f_{n'})$ such that each $f_{n',i}$ converges uniformly on compact subsets of $\mathcal{Z}$, from which the same holds for $f_{n'}$ also by the triangle inequality. □

B.3 Pushforward Maps Require Unbounded Bi-Lipschitz Constants

Theorem 2.1. *Suppose P_Z and $P_X^\star$ are probability measures on $\mathbb{R}^{d_\mathcal{Z}}$ and $\mathbb{R}^{d_\mathcal{X}}$ respectively, and that $\operatorname{supp} P_Z \ncong \operatorname{supp} P_X^\star$. Then for any sequence of measurable $f_n : \mathbb{R}^{d_\mathcal{Z}} \to \mathbb{R}^{d_\mathcal{X}}$, we can have $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^\star$ only if*

$$\lim_{n \to \infty} \operatorname{BiLip} f_n = \infty.$$

Proof. We suppose that $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^\star$ and prove the contrapositive. That is, without loss of generality (pass to a subsequence if necessary) we assume

$$M := \sup_n \operatorname{BiLip} f_n < \infty, \tag{B.4}$$

and prove that $\operatorname{supp} P_Z \cong \operatorname{supp} P_X^\star$.

We first show that (f_n) is pointwise bounded. To this end, observe that Prokhorov's theorem (Dudley, 2002, Proposition 9.3.4) means that P_Z is tight and that the sequence $(f_n \# P_Z)$ is uniformly tight. As such, there exists compact $K \subseteq \mathbb{R}^{d_\mathcal{Z}}$ such that $P_Z(K) > 0$, and compact $K' \subseteq \mathbb{R}^{d_\mathcal{X}}$ such that

$$\inf_n f_n \# P_Z(K') > 1 - P_Z(K).$$

For each n, we must then have some $z_n \in K$ such that $f_n(z_n) \in K'$; otherwise $K' \subseteq f_n(K)^c$ and so

$$\begin{aligned} f_n \# P_Z(K') &\leq f_n \# P_Z(f_n(K)^c) \\ &= 1 - f_n \# P_Z(f_n(K)) \\ &= 1 - P_Z(f_n^{-1}(f_n(K))) \\ &= 1 - P_Z(K) \end{aligned}$$

since f_n is injective by Proposition B.5. But for any fixed $z \in \mathbb{R}^{d_\mathcal{Z}}$, this entails

$$\begin{aligned}\sup_n \|f_n(z)\| &\leq \sup_n \|f_n(z_n)\| + \|f_n(z) - f_n(z_n)\| \\ &\leq \sup_{x \in K'} \|x\| + \sup_{z \in K} M\|z - z_n\| \\ &\leq \sup_{x \in K'} \|x\| + 2M \sup_{z \in K} \|z\| \\ &< \infty\end{aligned}$$

since K and K' are compact.

Next, observe that (B.4) easily means (f_n) is uniformly equicontinuous. In particular, for $\epsilon > 0$, choosing $\delta := \epsilon/M$ gives

$$\|f_n(z) - f_n(z')\| \leq M\|z - z'\| < \epsilon$$

for all n whenever $\|z - z'\| < \delta$

Theorem B.9 now entails the existence of a subsequence $(f_{n'})$ that converges uniformly on every compact subset of $\mathbb{R}^{d_\mathcal{Z}}$. In particular, $(f_{n'})$ converges pointwise to a limit that we denote by f_∞. Moreover, f_∞ is bi-Lipschitz. To see this, recall that for all n' and $z, z' \in \mathbb{R}^{d_\mathcal{Z}}$ we have

$$\frac{1}{M}\|z - z'\| \leq \|f_{n'}(z) - f_{n'}(z')\| \leq M\|z - z'\|,$$

by our assumption (B.4). Taking $n' \to \infty$ shows that $\operatorname{BiLip} f_\infty \leq M < \infty$.

We also have that

$$f_{n'}\#P_Z \xrightarrow{\mathcal{D}} f_\infty\#P_Z. \tag{B.5}$$

This follows from the Portmanteau theorem (Dudley, 2002, Theorem 11.3.3). In particular, suppose h is a bounded Lipschitz function, and let $B_r \subseteq \mathbb{R}^{d_\mathcal{Z}}$ denote a ball of radius $r > 0$ at the origin. Then

$$\begin{aligned}\left|\int h(x)\, f_{n'}\#P_Z(\mathrm{d}x) - \int h(x)\, f_\infty\#P_Z(\mathrm{d}x)\right| &= \left|\int h(f_{n'}(z)) - h(f_\infty(z))\, P_Z(\mathrm{d}z)\right| \\ &\leq \int_{B_r} |h(f_{n'}(z)) - h(f_\infty(z))|\, P_Z(\mathrm{d}z) \\ &\quad + \int_{B_r^c} |h(f_{n'}(z))| + |h(f_\infty(z))|\, P_Z(\mathrm{d}z) \\ &\leq P_Z(B_r)(\operatorname{Lip} h) \sup_{z \in B_r} \|f_n(z) - f_\infty(z)\| + 2P_Z(B_r^c) \sup_{z \in \mathbb{R}^{d_\mathcal{Z}}} |h(z)|\,.\end{aligned}$$

Hence

$$\limsup_{n' \to \infty} \left|\int h(x)\, f_{n'}\#P_Z(\mathrm{d}x) - \int h(x) f_\infty\#P_Z(\mathrm{d}x)\right| \leq 2P_Z(B_r^c) \sup_{z \in \mathbb{R}^{d_\mathcal{Z}}} |h(z)|$$

by the uniform convergence of $f_{n'}$ to f_∞ on compact subsets, and since $\operatorname{Lip} h < \infty$. Taking $r \to \infty$, the right-hand side vanishes since h is bounded, and we obtain (B.5).

We are now ready to complete the proof. Since f_∞ is bi-Lipschitz, Proposition B.5 means that f_∞ is a homeomorphism from $\mathbb{R}^{d_\mathcal{Z}}$ to $f_\infty(\mathbb{R}^{d_\mathcal{Z}})$. This certainly gives

$$\operatorname{supp} P_Z \cong f_\infty(\operatorname{supp} P_Z).$$

But now Proposition B.6 means

$$f_\infty(\operatorname{supp} P_Z) = \overline{f_\infty(\operatorname{supp} P_Z)}$$

where the closure is taken in $\mathbb{R}^{d_\mathcal{X}}$. However, from (B.5) we have

$$P_X^\star = f_\infty\#P_Z,$$

which by Proposition B.3 means that

$$\operatorname{supp} P_X^\star = \operatorname{supp} f_\infty \# P_Z = \overline{f_\infty(\operatorname{supp} P_Z)}.$$

Consequently

$$\operatorname{supp} P_X^\star = f_\infty(\operatorname{supp} P_Z) \cong \operatorname{supp} P_Z$$

as desired. □

The following corollary extends the above result to the case where $\operatorname{supp} P_X^\star$ may be homeomorphic to $\operatorname{supp} P_Z$, but $P_X^\star$ is very *close* to a probability measure with non-homeomorphic support to P_Z. Here ρ denotes any metric for the weak topology. In other words, ρ must be a metric on the space of distributions that satisfies $\rho(P_n, P) \to 0$ as $n \to \infty$ if and only if $P_n \xrightarrow{\mathcal{D}} P$. The Lévy-Prokhorov and bounded Lipschitz metrics provide standard examples of such ρ (Villani, 2008, Definition 3.3.10).

Corollary 2.2. *Suppose P_Z and P_X^0 are probability measures on $\mathbb{R}^{d_Z}$ and $\mathbb{R}^{d_X}$ respectively with $\operatorname{supp} P_Z \not\cong \operatorname{supp} P_X^0$. Then there exists nonincreasing $M : [0, \infty) \to [1, \infty]$ with $M(\epsilon) \to \infty$ as $\epsilon \to 0$ such that, for any probability measure $P_X^\star$ on $\mathbb{R}^{d_X}$, we have $\operatorname{BiLip} f \geq M(\epsilon)$ whenever $\rho(P_X^\star, P_X^0) \leq \epsilon$ and $\rho(f \# P_Z, P_X^\star) \leq \epsilon$.*

Proof. Define $M : [0, \infty) \to [1, \infty]$ by

$$M(\epsilon) := \inf \left\{ \operatorname{BiLip} f \mid f : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}, \rho(f \# P_Z, P_X^0) \leq 2\epsilon \right\},$$

with $M(\epsilon) := \infty$ if the infimum is taken over the empty set. Certainly M is nonincreasing. If we have both $\rho(P_X^\star, P_X^0) \leq \epsilon$ and $\rho(f \# P_Z, P_X^\star) \leq \epsilon$, then the triangle inequality gives

$$\rho(f \# P_Z, P_X^0) \leq \rho(f \# P_Z, P_X^\star) + \rho(P_X^\star, P_X^0) \leq 2\epsilon$$

and so $\operatorname{BiLip} f \geq M(\epsilon)$ since the right-hand side is an infimum. It remains only to show that $M(\epsilon) \to \infty$ as $\epsilon \to 0$. For contradiction, suppose there exists $\epsilon_n \to 0$ such that $\sup_n M(\epsilon_n) < \infty$. From the definition of M, this means that for each n there exists $f_n : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$ such that $\rho(f_n \# P_Z, P_X^0) \leq 2\epsilon_n$ and $\operatorname{BiLip} f_n \leq M(\epsilon_n) + 1$. It follows directly that $\rho(f_n \# P_Z, P_X^0) \to 0$ as $n \to \infty$, which in turn means $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^0$ since ρ is a metric for the weak topology. At the same time we have

$$\sup_n \operatorname{BiLip} f_n \leq \sup_n M(\epsilon_n) + 1 < \infty,$$

which contradicts Theorem 2.1, since we assumed $\operatorname{supp} P_Z \not\cong \operatorname{supp} P_X^0$. □

B.4 Variance of the Russian Roulette Estimator

In this section we briefly review the Russian roulette estimator used in Chen et al. (2019), and then discuss some scenarios in which we expect the variance of this estimator to increase unboundedly.

B.4.1 Russian Roulette Estimator

Residual Flows (ResFlows, (Chen et al., 2019)), building off of Invertible Residual Networks (iResNets, (Behrmann et al., 2019)), model the data by repeatedly stacking bijections of the form $f_\ell^{-1}(x) = x + g_\ell(x)$, where $\operatorname{Lip} g_\ell =: \kappa < 1$, as mentioned in (5). The change-of-variable formula for one layer of flow reads as, for $x \in \mathbb{R}^d$,

$$\log p_X(x) = \log p_Z(f_\ell^{-1}(x)) + \operatorname{tr}\left(\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \mathrm{D}g_\ell(x)^j \right). \tag{B.6}$$

To deal with this infinite series, iResNets truncate after a fixed number of terms – this provides a biased estimate of the log-likelihood of a point x under the model. ResFlows rely on an alternative method of estimating (B.6), first using a Russian roulette procedure to rewrite the series as follows:

$$\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \operatorname{tr}\left(\mathrm{D}g_\ell(x)^j\right) = \mathbb{E}_N\left[\sum_{j=1}^{N} \frac{(-1)^{j+1}}{j} \frac{\operatorname{tr}\left(\mathrm{D}g_\ell(x)^j\right)}{p_j} \right] =: S(x),$$

where $N \sim \text{Geom}(p)$ is a geometric random variable, and $p_k := \mathbb{P}(N \geq k)$. Then, taking a single sample $N \sim \text{Geom}(p)$, an unbiased estimator of S is given as S_N, where S_n is defined for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$ as

$$S_n(x) := \sum_{j=1}^{N} \frac{(-1)^{j+1}}{j} \frac{\operatorname{tr}\left(\mathrm{D}g_\ell(x)^j\right)}{p_j} \tag{B.7}$$

for any $x \in \mathbb{R}^d$. We will study the variance of S_N in this section.[15]

First, however, define the quantity $\alpha_j(x)$ for $j \in \mathbb{N}, x \in \mathbb{R}^d$ as

$$\alpha_j(x) := \frac{(-1)^{j+1}}{j} \operatorname{tr}\left(\mathrm{D}g(x)^j\right), \tag{B.8}$$

where we now drop the dependence of g on ℓ. Then, $S(x) = \sum_{j=1}^{\infty} \alpha_j(x)$, and $S_N(x) = \sum_{j=1}^{N} \alpha_j(x)/p_j$.

B.4.2 What might happen when $\kappa \to 1$?

We begin with an informal discussion on the variance of S_N as $\kappa \to 1$. First of all we know that, as $\kappa \to 1$, the mapping f^{-1} gets arbitrarily close to a non-invertible mapping: consider e.g. $g(x) = -\kappa x$, then $f^{-1} = (1-\kappa)\mathrm{Id} \to 0$ as $\kappa \to 1$. This near non-invertibility has implications for the speed of convergence of both $S(x)$ and its gradient,[16] as noted in these two results from Behrmann et al. (2019):

1. **Theorem 3:** $\left|\sum_{j=1}^{n} \alpha_j(x) - \log\det\left(I + \mathrm{D}g(x)\right)\right| \leq -d\left(\log(1-\kappa) + \sum_{j=1}^{n} \frac{\kappa^j}{j}\right)$,
2. **Theorem 4:** $\left\|\nabla_\theta\left(\alpha_j(x) - \log\det\left(I + \mathrm{D}g(x)\right)\right)\right\|_\infty = \mathcal{O}\left(\kappa^n\right)$.

We can see that both bounds become very loose as $\kappa \to 1$, implying we cannot guarantee the fast convergence of either series. It then follows that we cannot invoke the results from Rhee & Glynn (2015) and Beatson & Adams (2019) to argue that the variance of the Russian roulette estimator S_N will be small. Indeed, in the next section, we will look at a specific example where this variance becomes *infinite*.

B.4.3 A Specific Example of Infinite Variance

Now consider the case where $d = 1$. We will show that when $\kappa^2 > 1 - p$, there is a set of x having positive Lebesgue measure such that $S_N(x)$ from (B.7) has infinite variance.

We note that here we have $\operatorname{tr}\left(\mathrm{D}g(x)^j\right) = (g'(x))^j$ for any $j \in \mathbb{N}$. We can thus rewrite α_j from (B.8) as

$$\alpha_j(x) := \frac{(-1)^{j+1}}{j}(g'(x))^j. \tag{B.9}$$

Also recall that $N \sim \text{Geom}(p)$ and $p_j := \mathbb{P}(N \geq j)$ for all $j \in \mathbb{N}$.

Proposition B.10. *For any $x \in \mathbb{R}$ and random variable N satisfying $\operatorname{supp} N = \mathbb{N}$, $S_N(x)$ has finite expectation if $\kappa < 1$.*

Proof. Refer to Lyne et al. (2015, Proposition A.1). □

Proposition B.11. *Under the same conditions as Proposition B.10,*

$$\mathrm{Var} S_N(x) \geq \lim_{n\to\infty} 2\sum_{j=1}^{n} \alpha_j(x) S_{j-1}(x) - \mathbb{E}[S_N(x)]^2.$$

[15] Chen et al. (2019) additionally approximate $\operatorname{tr}\left(\mathrm{D}g_\ell(x)^j\right)$ by the Hutchinson's trace estimator $v^T \mathrm{D}g_\ell(x)^j v$ for $v \sim \mathcal{N}(0, I)$. Since v is independent of N, their estimator has strictly higher variance than (B.7).

[16] With respect to the flow parameters θ

Proof. This proof is taken from Lyne et al. (2015, Proposition A.2); we mostly rewrite the proof but adapt it to our specific setting and notation. Note that we will drop the dependence of S_j and α_j on x throughout the proof.

We know from Proposition B.10 that $\mathbb{E}[S_N(x)]$ is finite. Thus we will simply lower-bound $\mathbb{E}[S_N(x)^2]$.

We will first use induction to show the following holds for any $n \in \mathbb{N}$:

$$\sum_{j=1}^{n} S_j^2(p_j - p_{j+1}) = \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - S_n^2 p_{n+1}. \tag{B.10}$$

The base case is

$$S_1^2(p_1 - p_2) = \frac{\alpha_1^2}{p_1^2} p_1 - S_1^2 p_2 = \alpha_1^2 - S_1^2 p_2$$

since $p_1 = 1$. Now, assume (B.10) holds for some $m \in \mathbb{N}$. Then, for $n = m + 1$,

$$\begin{aligned}\sum_{j=1}^{m+1} S_j^2(p_j - p_{j+1}) &= \sum_{j=1}^{m} S_j^2(p_j - p_{j+1}) + S_{m+1}^2(p_{m+1} - p_{m+2}) \\ &= \alpha_1^2 + \sum_{j=2}^{m} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{m} \alpha_j S_{j-1} - S_m^2 p_{m+1} \\ &\quad + S_{m+1}^2(p_{m+1} - p_{m+2})\end{aligned} \tag{B.11}$$

by the inductive hypothesis. We also have

$$\begin{aligned}p_{m+1}(S_m^2 - S_{m+1}^2) &= p_{m+1}(S_m - S_{m+1})(S_m + S_{m+1}) \\ &= p_{m+1}\frac{\alpha_{m+1}}{p_{m+1}}\left(2S_m + \frac{\alpha_{m+1}}{p_{m+1}}\right) \\ &= \frac{\alpha_{m+1}^2}{p_{m+1}} + 2\alpha_{m+1}S_m.\end{aligned}$$

Substituting this result into (B.11) completes the induction and proves (B.10) for all $n \in \mathbb{N}$.

Now, by Jensen's inequality,

$$S_n^2 = \left(\sum_{j=1}^{n} \frac{p_j \frac{\alpha_j}{p_j}}{p_j}\right)^2 \le \frac{\sum_{j=1}^{n} \frac{\alpha_j^2}{p_j}}{\sum_{j=1}^{n} p_j}.$$

This implies

$$p_{n+1}S_n^2 \le p_n S_n^2 \le \frac{p_n}{\sum_{j=1}^{n} p_j} \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j} \le \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j}$$

since (p_n) is a positive sequence.

This finally implies the following lower bound for any $n \in \mathbb{N}$:

$$\begin{aligned}
\sum_{j=1}^{n} S_j^2 \mathbb{P}(N=j) &= \sum_{j=1}^{n} S_j^2 (p_j - p_{j+1}) \\
&= \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - S_n^2 p_{n+1} \\
&\geq \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j} \\
&= \alpha_1^2 (1 - p_1^{-1}) + 2\sum_{j=2}^{n} \alpha_j S_{j-1} \\
&= 2\sum_{j=2}^{n} \alpha_j S_{j-1},
\end{aligned}$$

where the final line follows because $p_1 = 1$.

Since $\mathbb{E}[S_N^2] = \lim_{n\to\infty} \sum_{j=1}^{n} S_j^2 \mathbb{P}(N=j)$, the proof is complete. □

We are about ready to prove the main result but require one more auxiliary result first.

Proposition B.12. *Suppose* $|b| > 1$. *Then,*

$$\lim_{n\to\infty} \frac{n}{b^n} \sum_{j=1}^{n-1} \frac{b^j}{j} = \frac{1}{b-1}.$$

Proof. We will first show that the limit exists, and then show that it equals $(b-1)^{-1}$. Let

$$c_n = \frac{n}{b^n} \sum_{j=1}^{n-1} \frac{b^j}{j}.$$

We can rewrite this as follows:

$$c_n = \sum_{j=1}^{n-1} \frac{n}{b^{n-j} j} = \sum_{j=1}^{n-1} \frac{n}{b^j (n-j)} = \sum_{j=1}^{n-1} \frac{1}{b^j} + \sum_{j=1}^{n-1} \frac{j}{b^j (n-j)}.$$

Since $b > 1$, the first sum is a convergent geometric series as $n \to \infty$. We can decompose the second sum into its positive and negative terms:

$$\sum_{j=1}^{n-1} \frac{j}{b^j (n-j)} = \sum_{j \geq 1 : b^j > 0}^{n-1} \frac{j}{b^j (n-j)} + \sum_{j \geq 1 : b^j < 0}^{n-1} \frac{j}{b^j (n-j)} \equiv ①_n + ②_n.$$

We can see, for all $n \in \mathbb{N}$,

$$①_n \geq -\sum_{j=1}^{n-1} \frac{j}{|b|^j (n-j)} \qquad \text{and} \qquad ②_n \leq \sum_{j=1}^{n-1} \frac{j}{|b|^j (n-j)}.$$

Furthermore, for all $j \in \{1, \ldots, n-1\}$, we have

$$\frac{j}{n-j} \leq j.$$

Now notice that the series $\sum_{j=1}^{\infty} \frac{j}{|b|^j}$ converges by the ratio test:

$$\lim_{j\to\infty} \left| \frac{\frac{j+1}{|b|^{j+1}}}{\frac{j}{|b|^j}} \right| = \lim_{j\to\infty} \frac{j+1}{j|b|} = \frac{1}{|b|} < 1.$$

This implies the existence of $\lim_{n\to\infty}\sum_{j=1}^{n-1}\frac{j}{|b|^j(n-j)}$. Since the sequence $(①_n)$ (resp. $(②_n)$) is negative, non-increasing, and bounded below (resp. positive, non-decreasing, and bounded above), this implies the existence of $\lim_{n\to\infty} ①_n$ (resp. $\lim_{n\to\infty} ②_n$). Altogether, this implies the existence of

$$\lim_{n\to\infty}\left(\sum_{j=1}^{n-1}\frac{1}{b^j}+\sum_{j=1}^{n-1}\frac{j}{b^j(n-j)}\right)=\lim_{n\to\infty}c_n=:c_\infty.$$

Now we will determine its precise value. Note the following recurrence for all $n\in\mathbb{N}$:

$$c_{n+1}=\frac{n+1}{bn}\left(1+c_n\right).$$

Taking the limit of both sides as $n\to\infty$ gives

$$c_\infty=\frac{1}{b}(1+c_\infty).$$

Solving this gives us $c_\infty=\frac{1}{b-1}$, which completes the proof. $\square$

Proposition B.13. *Suppose $N\sim$ Geom(p), g is continuously differentiable, and $1-p<\kappa^2<1$. Then*

$$\{x\in\mathbb{R}\mid \mathrm{Var}S_N(x)=\infty\}$$

has positive Lebesgue measure.

Proof. From Proposition B.11, for a given $x\in\mathbb{R}$, we can see that showing $\sum_{n=2}^{\infty}\alpha_n(x)S_{n-1}(x)$ diverges is sufficient to prove $\mathrm{Var}S_N(x)$ is infinite.

Consider using the ratio test to assess the convergence of the above series, with terms defined as $a_n(x):=\alpha_n(x)S_{n-1}(x)$. We have the following for any $n\geq 2$:

$$\begin{aligned}\left|\frac{a_{n+1}(x)}{a_n(x)}\right|&=\left|\frac{\alpha_{n+1}(x)S_n(x)}{\alpha_n(x)S_{n-1}(x)}\right|\\&=\frac{\frac{|(g'(x))^{n+1}|}{n+1}}{\frac{|(g'(x))^n|}{n}}\cdot\left|\frac{\sum_{j=1}^{n}\frac{\alpha_j(x)}{p_j}}{\sum_{j=1}^{n-1}\frac{\alpha_j(x)}{p_j}}\right|\\&=\frac{n|g'(x)|}{n+1}\cdot\left|\frac{(-1)^{n+1}\cdot(g'(x))^n}{np_n}\left(\sum_{j=1}^{n-1}\frac{(-1)^{j+1}\cdot(g'(x))^j}{jp_j}\right)^{-1}+1\right|.\end{aligned}$$

Recall $p_j=(1-p)^{j-1}\equiv q^{j-1}$. Then, writing $b=-\frac{g'(x)}{q}$, we have

$$\frac{(-1)^{n+1}\cdot(g'(x))^n}{np_n}\left(\sum_{j=1}^{n-1}\frac{(-1)^{j+1}\cdot(g'(x))^j}{jp_j}\right)^{-1}=\frac{1}{n}b^n\left(\sum_{j=1}^{n-1}\frac{1}{j}b^j\right)^{-1}.$$

Now let us assume that $|g'(x)|^2>q$. We can see that $|g'(x)|^2>q\implies|g'(x)|>q$ since $q\in(0,1)$, which then entails $|b|>1$. Therefore, by Proposition B.12,

$$\lim_{n\to\infty}\frac{n}{b^n}\sum_{j=1}^{n-1}\frac{b^j}{j}=\frac{1}{b-1}.$$

This then implies

$$\begin{aligned}\lim_{n\to\infty}\left|\frac{a_{n+1}(x)}{a_n(x)}\right|&=\lim_{n\to\infty}\frac{n|g'(x)|}{n+1}\left|\frac{1}{\frac{n}{b^n}\sum_{j=1}^{n-1}\frac{b^j}{j}}+1\right|\\&=|g'(x)|\left|\frac{1}{\frac{1}{b-1}}+1\right|=\frac{|g'(x)|^2}{q}>1\end{aligned}$$

since we have assumed that $|g'(x)|^2 > q$. Thus, for all x in the set

$$V_{g,q} \coloneqq \{x \in \mathbb{R} \mid |g'(x)|^2 > q\},$$

the series $\sum_{n=2}^{\infty} \alpha_n(x) S_{n-1}(x)$ diverges by the ratio test. This means that $\mathrm{Var} S_N(x) = \infty$ for all $x \in V_{g,q}$.

Finally, we will prove the set $\{x \in \mathbb{R} \mid \mathrm{Var} S_N(x) = \infty\}$ has positive Lebesgue measure. Recall that $\mathrm{Lip}\, g = \kappa$, which directly implies $\sup_{x\in\mathbb{R}} |g'(x)| = \kappa$ from Proposition B.7 and thus $\sup_{x\in\mathbb{R}} |g'(x)|^2 = \kappa^2$. Then, since $\kappa^2 > q$, there exists $x_0 \in \mathbb{R}$ such that $|g'(x_0)|^2 \in (q, \kappa^2)$. By the continuity of $|g'|$, there is open ball of nonzero radius around x_0, denoted $\mathcal{B}(x_0)$, such that $|g'(x)| > q$ for all $x \in \mathcal{B}(x_0)$. Since $\mathcal{B}(x_0)$ is open and non-empty, it has positive Lebesgue measure. The inclusions

$$\mathcal{B}(x_0) \subseteq V_{g,q} \subseteq \{x \in \mathbb{R} \mid \mathrm{Var} S_N(x) = \infty\}$$

thus conclude the proof.

□

B.4.4 DISCUSSION

Changing p as κ increases An obvious strategy to avoid satisfying the conditions of Proposition B.13 is to set p such that $1 - \kappa^2 > p$. However, lowering p in this way incurs additional computational cost: the average number of iterations per training step is equal to p^{-1}, or is lower-bounded by $(1-\kappa^2)^{-1}$ if $p < 1 - \kappa^2$. Thus, if we send $\kappa \to 1$ to mitigate the bi-Lipschitz constraint (6), we will either incur an infinite computational cost or run the risk of encountering infinite variance.

Higher dimensions Although Proposition B.13 only applies for $d = 1$, it is conceivable that similar results can be derived for $d > 1$, especially when considering the discussion in Section B.4.2. We leave a deeper investigation for future work.

B.5 Density of a CIF

We make precise our heuristic derivation of the density (11) via the following result.

Proposition B.14. *Suppose $\mathcal{Z}, \mathcal{X} \subseteq \mathbb{R}^d$ are open, and that $F(\cdot; u) : \mathcal{Z} \to \mathcal{X}$ is a continuously differentiable bijection with everywhere invertible Jacobian for each $u \in \mathcal{U}$. Under the generative model (8), (X, U) has joint density*

$$p_Z(F^{-1}(x;u))\, p_{U|Z}(u|F^{-1}(x;u))\, |\det \mathrm{D}F^{-1}(x;u)|.$$

Proof. Suppose $h : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is a bounded measurable test function. Then

$$\begin{aligned}
\mathbb{E}[h(X,U)] &= \mathbb{E}[h(F(Z;U),U)] \\
&= \int \left[\int h(F(z;u),u)\, p_Z(z)\, p_{U|Z}(u|z)\, \mathrm{d}z \right] \mathrm{d}u \\
&= \int h(x,u)\, p_Z(F^{-1}(x;u))\, p_{U|Z}(u|F^{-1}(x;u))\, |\det \mathrm{D}F^{-1}(x;u)|\, \mathrm{d}z\, \mathrm{d}u,
\end{aligned}$$

where in the third line we substitute $x \coloneqq F(z;u)$ on the inner integral, which is valid by Theorem 17.2 of Billingsley (2008). Now for $A \subseteq \mathcal{X} \times \mathcal{U}$, let $h \coloneqq \mathbb{I}_A$. It follows that

$$\mathbb{P}((X,U) \in A) = \mathbb{E}[\mathbb{I}_A(X,U)] = \int_A p_Z(F^{-1}(x;u))\, p_{U|Z}(u|F^{-1}(x;u))\, |\det \mathrm{D}F^{-1}(x;u)|\, \mathrm{d}z\, \mathrm{d}u,$$

which gives the result since A was arbitrary. □

B.6 Our Approximate Posterior Does Not Sacrifice Generality

The following result shows that our parameterisation of the approximate posterior $q_{U_{1:L}|X}$ in (15) does not lose generality. In particular, provided each $q_{U_\ell|Z_\ell}$ is sufficiently expressive, we can always recover the exact posterior.

Proposition B.15. *Under the generative model (8), the posterior factors like*

$$p_{U_{1:L}|X}(u_{1:L}|x) = \prod_{\ell=1}^{L} p_{U_\ell|Z_\ell}(u_\ell|z_\ell),$$

where $z_L \coloneqq x$ and $z_\ell \coloneqq F_{\ell+1}^{-1}(z_{\ell+1}; u_{\ell+1})$ for $\ell \in \{1, \ldots, L-1\}$.

Proof. Writing $p_{U_{1:L}|X}$ autoregressively gives

$$p_{U_{1:L}|X}(u_{1:L}|x) = \prod_{\ell=1}^{L} p_{U_\ell|U_{\ell+1:L},X}(u_\ell|u_{\ell+1:L}, x).$$

But now it is clear from the generative model (8) that U_ℓ is conditionally independent of $(U_{\ell+1:L}, X)$ given Z_ℓ, and as such

$$p_{U_\ell|U_{\ell+1:L},X}(u_\ell|u_{\ell+1:L}, x) = p_{U_\ell|Z_\ell}(u_\ell|z_\ell).$$

Substituting this into the above expression then gives the result. □

B.7 Conditions for a CIF to Outperform an Underlying Normalising Flow

For this result, the components of our model are assumed to be parameterised by $\theta \in \Theta$, which we will indicate by by F_θ, $p^\theta_{U|Z}$, and $q^\theta_{U|X}$. We will also use θ to indicate quantities that result from the choice of parameters θ (e.g. P^θ_X for the distribution obtained), and will denote by $\mathcal{L}^\theta$ the corresponding ELBO (14).

Proposition 4.1. *Suppose there exists $\phi \in \Theta$ such that, for some bijection $f : \mathcal{Z} \to \mathcal{X}$, $F_\phi(\cdot; u) = f(\cdot)$ for all $u \in \mathcal{U}$. Likewise, suppose $p^\phi_{U|Z}$ and $q^\phi_{U|X}$ are such that, for some density r on $\mathcal{U}$, $p^\phi_{U|Z}(\cdot|z) = q^\phi_{U|X}(\cdot|x) = r(\cdot)$ for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. If $\mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\theta(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)]$, then*

$$D_{\mathrm{KL}}\big(P^\star_X \,\big\|\, P^\theta_X\big) \leq D_{\mathrm{KL}}(P^\star_X \,\|\, f\#P_Z).$$

Proof. Observe from (11) that

$$p^\phi_{X,U}(x, u) = p_Z(f^{-1}(x))\, |\det \mathrm{D}f^{-1}(x)|\, p_{U|Z}(u|f^{-1}(x)).$$

It then follows from (13) that, under ϕ, the model has density

$$p^\phi_X(x) = p_Z(f^{-1}(x))\, |\det \mathrm{D}f^{-1}(x)| \underbrace{\int p_{U|Z}(u|f^{-1}(x))\ \mathrm{d}u}_{=1}$$

which is exactly the density of the normalising flow $f\#P_Z$. We also obtain the posterior

$$\begin{aligned} p^\phi_{U|X}(u|x) &= \frac{p^\phi_{X,U}(x,u)}{p^\phi_X(x)} \\ &= p_{U|Z}(u|f^{-1}(x)) \\ &= r(u). \end{aligned}$$

Since each $q^\phi_{U|X}(\cdot|x) = r(\cdot)$ also, it follows that $\mathcal{L}^\phi$ is tight, so that $\mathcal{L}^\phi(x) = \log p^\phi_X(x)$ for all $x \in \mathcal{X}$.

Now suppose some $\theta \in \Theta$ has

$$\mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\theta(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)].$$

It follows that

$$\mathbb{E}_{x\sim P^\star_X}[\log p^\theta_X(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)] = \mathbb{E}_{x\sim P^\star_X}[\log p^\phi_X(x)].$$

Subtracting $\mathbb{E}_{x\sim P^\star_X}[\log p^\star_X(x)]$ from both sides and negating gives

$$D_{\mathrm{KL}}\big(P^\star_X \,\big\|\, P^\theta_X\big) \leq D_{\mathrm{KL}}\left(P^\star_X \,\middle\|\, P^\phi_X\right) = D_{\mathrm{KL}}(P^\star_X \,\|\, f\#P_Z).$$

□

B.8 CIFs Can Learn Target Supports Exactly

In this section we give necessary and sufficient conditions for a CIF to learn the support of a target distribution exactly, without needing changes to F. However, our argument applies more generally and does not make specific use of the bijective structure of F. To make this clear, we formulate our result here in terms of a generalisation of the model (7). In particular, we will take P_X as the marginal in X of

$$Z \sim P_Z, \quad U \sim P_{U|Z}(\cdot|Z), \quad X := G(Z, U), \tag{B.12}$$

where $G : \mathcal{Z} \times \mathcal{U} \to \mathcal{X}$. We will assume that

- $\mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}}, \mathcal{U} \subseteq \mathbb{R}^{d_\mathcal{U}}$, and $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ are equipped with the subspace topology;
- P_Z and each $P_{U|Z}(\cdot|z)$ are Borel probability measures on $\mathcal{Z}$ and $\mathcal{U}$ respectively;
- G is continuous with respect to the product topology $\mathcal{Z} \times \mathcal{U}$.

We then have the following formula for $\operatorname{supp} P_X$:

Lemma B.16. *Under the model* (B.12)*,*

$$\operatorname{supp} P_X = \overline{\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z))}.$$

Proof. Denote the joint distribution of (Z, U) by $P_{Z,U}$. Observe from Proposition B.3 that

$$\operatorname{supp} P_X = \overline{G(\operatorname{supp} P_{Z,U})}.$$

Let

$$B := \bigcup_{z \in \operatorname{supp} Z} \{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z).$$

The result follows if we can show that

$$\operatorname{supp} P_{Z,U} = \overline{B},$$

since $\overline{G(\overline{B})} = \overline{G(B)}$ because G is continuous.

We first show that $\operatorname{supp} P_{Z,U} \supseteq \overline{B}$. Suppose $(z, u) \in B$, and let $N_{(z,u)} \subseteq \mathcal{Z} \times \mathcal{U}$ be an open set containing (z, u). Then there exists open N_z and N_u containing z and u respectively such that $N_z \times N_u \subseteq N_{(z,u)}$, since the open rectangles form a base for the product topology. It follows that

$$\begin{aligned} P_{Z,U}(N_{(z,u)}) &\geq P_{Z,U}(N_z \times N_u) \\ &= \int_{N_z} P_{U|Z}(N_u|z') \, P_Z(\mathrm{d}z') \\ &> 0, \end{aligned}$$

since by the definition of B we have $P_Z(N_z) > 0$ and $P_{U|Z}(N_u|z) > 0$ for each $u \in N_z$. From this we have $\operatorname{supp} P_{Z,U} \supseteq B$, and taking the closure of each side gives $\operatorname{supp} P_{Z,U} \supseteq \overline{B}$.

In the other direction, suppose that $(z, u) \notin \overline{B}$. Then there exist open sets N_z and N_u containing z and u respectively such that

$$(N_z \times N_u) \cap \overline{B} = \emptyset.$$

By the definition of B, it follows that if $(z', u') \in N_z \times N_u$ and $z' \in \operatorname{supp} P_Z$, then $u' \notin \operatorname{supp} P_{U|Z}(\cdot|z')$. Otherwise stated, if $z' \in N_z \cap \operatorname{supp} P_Z$, then

$$N_u \cap \operatorname{supp} P_{U|Z}(\cdot|z') = \emptyset.$$

Thus

$$\begin{aligned} P_{Z,U}(N_z \times N_u) &= \int_{N_z} \left[\int_{N_u} P_{U|Z}(\mathrm{d}u'|z') \right] P_Z(\mathrm{d}z') \\ &= \int_{N_z \cap \operatorname{supp} P_Z} \left[\int_{N_u \cap \operatorname{supp} P_{U|Z}(\cdot|z')} P_{U|Z}(\mathrm{d}u'|z') \right] P_Z(\mathrm{d}z') \\ &= 0, \end{aligned}$$

where the second line follows from Proposition B.4. Consequently $(z, u) \notin \operatorname{supp} P_{Z,U}$, which gives $\operatorname{supp} P_{Z,U} \subseteq \overline{B}$. $\square$

We now give necessary and sufficient conditions for the model (B.12) to learn a given target support exactly.

Proposition B.17. *Suppose* $P_X^\star(\partial \operatorname{supp} P_X^\star) = 0$ *and that*

$$\overline{G(\operatorname{supp} P_Z \times \mathcal{U})} \supseteq \operatorname{supp} P_X^\star. \tag{B.13}$$

Then there exists $P_{U|Z}$ *such that* $\operatorname{supp} P_X = \operatorname{supp} P_X^\star$ *if and only if, for all* $z \in \operatorname{supp} P_Z$*, there exists* $u \in \mathcal{U}$ *with*

$$G(z, u) \in \operatorname{supp} P_X^\star.$$

Proof. ($\Rightarrow$) Choose $P_{U|Z}$ such that $\operatorname{supp} P_X = \operatorname{supp} P_X^\star$. Lemma B.16 gives

$$\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z)) \subseteq \operatorname{supp} P_X^\star.$$

Suppose $z \in \operatorname{supp} P_Z$. Then for indeed all $u \in \operatorname{supp} P_{U|Z}(\cdot|z)$ we must have $G(z, u) \in \operatorname{supp} P_X^\star$, which proves this direction since $\operatorname{supp} P_{U|Z}(\cdot|z) \neq 0$ by Proposition B.4.

($\Leftarrow$) For $z \in \operatorname{supp} P_Z$, let

$$A_z \coloneqq \{u \in \mathcal{U} : G(z, u) \in \operatorname{int}(\operatorname{supp} P_X^\star)\}, \tag{B.14}$$

where int denotes the *interior* operator. If $A_z = \emptyset$, define $P_{U|Z}(\cdot|z)$ to be Dirac on some u such that $G(z, u) \in \operatorname{supp} P_X^\star$, which exists by assumption. Otherwise, we let $P_{U|Z}(\cdot|z)$ be a probability measure with support $\overline{A_z}$. To show that such a measure exists, observe that A_z is open since G is continuous. Since $\mathcal{U}$ is separable, we can therefore write

$$A_z = \bigcup_{n=1}^{\infty} B_n$$

for a countable collection of open sets $B_n \subseteq A_z$. We can then define a probability measure μ by

$$\mu(C) \coloneqq \sum_{n=1}^{\infty} 2^{-n}\, \mathbb{I}(C \cap B_n \neq \emptyset)$$

for measurable $C \subseteq \mathcal{U}$. Since $A_z \neq \emptyset$, it is straightforward to see that this is a probability measure with $\mu(A_z) = 1$. Consequently $\operatorname{supp} \mu = \overline{A_z}$ by Proposition B.2, since A_z is open and $\operatorname{supp} \mu$ is the smallest closed set with μ-probability 1.

We show this construction gives $\operatorname{supp} P_X \subseteq \operatorname{supp} P_X^\star$. To this end, we first prove that if $z \in \operatorname{supp} P_Z$ and $u \in \operatorname{supp} P_{U|Z}(\cdot|z)$ then

$$G(z, u) \in \operatorname{supp} P_X^\star.$$

If $A_z = \emptyset$ this is immediate. Otherwise, since $\operatorname{supp} P_{U|Z}(\cdot|z) = \overline{A_z}$, there exists $(u_n) \subseteq A_z$ such that $u_n \to u$. By (B.14), each $G(z, u_n) \in \operatorname{supp} P_X^\star$. By continuity we then have

$$G(z, u_n) \to G(z, u) \in \operatorname{supp} P_X^\star$$

since $\operatorname{supp} P_X^\star$ is closed. It follows that

$$\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z)) \subseteq \operatorname{supp} P_X^\star,$$

which gives $\operatorname{supp} P_X \subseteq \operatorname{supp} P_X^\star$ from Lemma B.16 since $\operatorname{supp} P_X^\star$ is closed.

We now show $\operatorname{supp} P_X \supseteq \operatorname{supp} P_X^\star$. Since $P_X^\star(\partial \operatorname{supp} P_X^\star) = 0$ we have

$$\operatorname{supp} P_X^\star = \overline{\operatorname{int}(\operatorname{supp} P_X^\star)}$$

by Proposition B.2, so that $\operatorname{supp} P_X \supseteq \operatorname{supp} P_X^\star$ if $\operatorname{supp} P_X \supseteq \operatorname{int}(\operatorname{supp} P_X^\star)$. Now suppose $x \in \operatorname{int}(\operatorname{supp} P_X^\star)$. Then there exists $(z_n) \subseteq \operatorname{supp} P_Z$ and $(u_n) \subseteq \mathcal{U}$ such that $G(z_n, u_n) \to x$ by (B.13). But then we must have $G(z_n, u_n) \in \operatorname{int}(\operatorname{supp} P_X^\star)$ for n large enough because x lies in the interior. Consequently, for n large enough,

$$u_n \in A_{z_n} \subseteq \operatorname{supp} P_{U|Z}(\cdot|z_n)$$

and hence $G(z_n, u_n) \in \operatorname{supp} P_X$ by Lemma B.16. This means $x \in \operatorname{supp} P_X$ since $\operatorname{supp} P_X$ is closed. □

The following proposition then gives a straightforward condition under which it is additionally possible to recover the *target* exactly (i.e. not just its support). In our experiments we do not enforce this condition explicitly. However, since we learn the parameters of G here, we can expect our model will approximate this behaviour if doing so produces a better density estimator.

Proposition B.18. *If $G(z, \cdot)$ is surjective for each $z \in \mathcal{Z}$, then there exists $P_{U|Z}$ such that $P_X = P_X^\star$.*

Proof. Fix $z \in \mathcal{Z}$. Surjectivity of $G(z, \cdot)$ means that, for $x \in \mathcal{X}$, there exists $u \in \mathcal{U}$ such that $G(z, u) = x$. Thus we can define $H_z : \mathcal{X} \to \mathcal{U}$ such that

$$G(z, H_z(x)) = x$$

for all $x \in \mathcal{X}$. We then define each

$$P_{U|Z}(\cdot|z) \coloneqq H_z \# P_X^\star.$$

From this it follows that $P_X = P_X^\star$. For, letting $B \subseteq \mathcal{X}$ be measurable,

$$\begin{aligned}
P_X(B) &= \int_{G^{-1}(B)} P_{U|Z}(\mathrm{d}u|z)\, P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(G(z,u))\, H_z \# P_X^\star(\mathrm{d}u) \right] P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(G(z,H_z(x)))\, P_X^\star(\mathrm{d}x) \right] P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(x)\, P_X^\star(\mathrm{d}x) \right] P_Z(\mathrm{d}z) \\
&= P_X^\star(B),
\end{aligned}$$

which gives the result. □

C Experimental Details

Our choices (10) and (17) required parameterising s, t, μ^p, Σ^p, μ^q, and Σ^q. Since these terms are naturally paired, at each layer of our model we set

$$\begin{aligned}
[s(u), t(u)] &\coloneqq \mathtt{NN}_F(u), \\
[\mu^p(z), \varsigma^p(z)] &\coloneqq \mathtt{NN}_p(z), \\
\Sigma^p(z) &\coloneqq \operatorname{diag}(e^{\varsigma^p(z)}), \\
[\mu^q(x), \varsigma^q(x)] &\coloneqq \mathtt{NN}_q(x), \\
\Sigma^q(x) &\coloneqq \operatorname{diag}(e^{\varsigma^q(x)}),
\end{aligned}$$

where $\mathtt{NN}$ denotes a separate neural network and $\varsigma^p(z), \varsigma^q(x) \in \mathbb{R}^d$.

In all experiments we trained our models to maximise either the log-likelihood (for the baseline flows) or the ELBO (for the CIFs) using the ADAM optimiser (Kingma & Ba, 2015) with default hyperparameters and no weight decay. The ELBO was estimated using a single sample per datapoint (i.e. a single call to Algorithm 1). We used a held-out validation set and trained each model until its validation score stopped improving, except for the NSF tabular data experiments where we train for a fixed number of epochs as specified in Durkan et al. (2019). After training, we used validation performance to select the best parameters found during training for use at test time (again except for the NSF experiments, where we just test with the final model). Both validation and test scores were computed using the exact log-likelihood for the baseline and the importance sampling estimate (16) for the CIFs, with $m = 5$ samples for validation and $m = 100$ for testing.

C.1 Tabular Data Experiments

Following Papamakarios et al. (2017), we experimented with the POWER, GAS, HEPMASS, and MINIBOONE datasets from the UCI repository (Bache & Lichman, 2013), as well as a dataset of 8×8 image patches extracted from the BSDS300 dataset (Martin et al., 2001). We preprocessed these datasets identically to Papamakarios et al. (2017), and used the same train/validation/test splits. For all CIF-ResFlow models, we used a batch size of 1000 and a learning rate of 10^{-3}. For the MAF experiments, we used a batch size of 1000 and a learning rate of 10^{-3}, except for BSDS300 where we used a learning rate of 10^{-4} to control the instability of the baseline. For the NSF experiments, we used batch sizes and learning rates as dictated by Durkan et al. (2019, Table 5), along with their cosine learning rate annealing scheme.

Also, for all CIF models, each U_ℓ had the same dimension $d_\mathcal{U}$, which we took to be roughly a quarter of the dimensionality of the data (except in Section C.1.4 for which $d_\mathcal{U} = d_\mathcal{X}$). In particular, we set $d_\mathcal{U} \coloneqq 2$ for POWER and GAS, $d_\mathcal{U} \coloneqq 5$ for HEPMASS, $d_\mathcal{U} \coloneqq 10$ for MINIBOONE, and $d_\mathcal{U} \coloneqq 15$ for BSDS300.

C.1.1 Residual Flows

The residual blocks in all ResFlow models used multilayer perceptrons (MLPs) with 4 hidden layers of 128 hidden units (denoted 4×128), LipSwish nonlinearities (Chen et al., 2019, (10)) before each linear layer, and a residual connection from the input to the output. We did not use any kind of normalisation (e.g. ActNorm or BatchNorm) for these experiments. For all models we set $\kappa = 0.9$ in (5) to match the value for the 2-D experiments in the codebase of Chen et al. (2019). Other design choices followed Chen et al. (2019). In particular:

- We always exactly computed several terms at the beginning of the series expansion of the log Jacobian, and then used Russian Roulette sampling (Kahn, 1955) to estimate the sum of the remaining terms. In particular, at training time we computed 2 exact terms, while at test time we computed 20 exact terms;
- We used a geometric distribution with parameter 0.5 for the number of terms to compute in our Russian Roulette estimators;
- We used the Skilling-Hutchinson trace estimator (Skilling, 1989; Hutchinson, 1990) to estimate the trace in the log Jacobian term;
- At both training and test time, we used a single Monte Carlo sample of (n, v) to estimate (6) of Chen et al. (2019);

However, note that for these experiments, for the sake of simplicity, we did not use the memory-saving techniques in (8) and (9) of Chen et al. (2019), nor the adaptive power iteration scheme described in their Appendix E.

For $\mathtt{NN}_F$, $\mathtt{NN}_p$, and $\mathtt{NN}_q$ we used 2×10 MLPs with $\tanh$ nonlinearities These networks were much smaller than 4×128, and hence the CIF-ResFlows had only roughly 1.5-4.5% more parameters (depending on the dimension of the dataset) than the otherwise identical 10-layer ResFlows, and roughly 10% of the parameters of the 100-layer ResFlows.

The 100-layer ResFlows were significantly slower to train than the 10-layer models, and for POWER, GAS, and BSDS300 we were forced to stop these before their validation loss had converged. However, to ensure a fair comparison, we allocated more total computing power to these models than to the 10-layer models, which were terminated properly. In particular, we trained each 100-layer ResFlow on POWER and GAS for a total of 10 days on a single NVIDIA GeForce GTX 1080 Ti, and on BSDS300 for a total of 7 days. In contrast, the 10-layer ResFlows converged after around 1 day on POWER, 4.5 days on GAS, and around 3 days on BSDS300. Likewise, the 10-layer CIF-ResFlows converged after around 1 day on POWER, 6 days on GAS, and 2 days on BSDS300.

Table C.4: MAF and CIF-MAF parameter configurations for POWER and GAS.

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	NN_p SIZE	NN_q SIZE	NN_F SIZE
MAF	5, 10, 20	2×100, 2×200, 2×400	-	-	-
CIF-MAF	5, 10	2×128	$2 \times 100, 2 \times 200$	$2 \times 100, 2 \times 200$	2×128

Table C.5: MAF and CIF-MAF parameter configurations for HEPMASS and MINIBOONE

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	NN_p SIZE	NN_q SIZE	NN_F SIZE
MAF	5, 10, 20	2×128, 2×512, 2×1024	-	-	-
CIF-MAF	5, 10	2×128	$2 \times 128, 2 \times 512$	$2 \times 128, 2 \times 512$	2×128

C.1.2 MASKED AUTOREGRESSIVE FLOWS

The experiment comparing MAF baselines to CIF-MAFs was inspired by the experimental setup in Papamakarios et al. (2017). For each dataset, we specified a set of hyperparameters over which to search for both the baselines and the CIFs; these hyperparameters are provided in Table C.4, Table C.5, and Table C.6. Then, we trained each model until no validation improvement had been observed for 50 epochs. We then evaluated the model with the best validation score among all candidate models on the test dataset to obtain a log-likelihood score. We performed this procedure with three separate random seeds, and report the average and standard error across the runs in Table 1.

We searched over all combinations of parameters listed in Table C.4, Table C.5, and Table C.6. For example, on HEPMASS or MINIBOONE, our set of candidate MAF models included: for $L = 5$, an autoregressive network of size of either 2×128, 2×512, or 2×1024; for $L = 10$, an autoregressive network size of either 2×128, 2×512, or 2×1024; and for $L = 20$, again an autoregressive network size of either 2×128, 2×512, or 2×1024; this gave us a total of 9 candidate MAF models for each seed. The set of candidate CIF-MAF models can similarly be determined via the table and gave us a total of 8 candidate models for each seed. We maintained this split of 9 candidates for MAF and 8 candidates for CIF-MAF across datasets to fairly compare against the baseline by allowing them more configurations. We also considered deeper and wider MAF models to compensate for the additional parameters introduced by NN_F, NN_p, and NN_q in the CIF-MAFs. Finally, we allowed the baseline MAF models to use batch normalization between MADE layers as recommended by Papamakarios et al. (2017), but we do not use them within CIF-MAFs as the structure of our F generalises this transformation.

We should note that our evaluation of models is slightly different from Papamakarios et al. (2017). For the model which scores best on the validation set, Papamakarios et al. (2017) report the average and standard deviation of log-likelihood across the points in the test dataset. However, our error bars emerge as the error in average test-set log-likelihood across *multiple* runs of the same experiment; this style of evaluation is often employed in other works as well (e.g. FFJORD (Grathwohl et al., 2019), NAF (Huang et al., 2018), and SOS (Jaini et al., 2019) as noted in Durkan et al. (2019, Table 1)).

C.1.3 NEURAL SPLINE FLOWS

The experiment comparing NSF baselines to CIF-NSFs mirrors the experimental setup in Durkan et al. (2019). Specifically, we constructed baseline NSFs that exactly copied the settings in Durkan et al. (2019, Table 5). We also built CIF-NSFs using these baseline settings, although for the CIF-NSF-1 model we lowered the number of hidden channels in the autoregressive networks so that the total number of trainable parameters matched that of the baseline. Our parameter settings are provided in Table C.7; note that parameter settings are homogeneous across datasets, besides MINIBOONE for which we reduced

Table C.6: MAF and CIF-MAF parameter configurations for BSDS300

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	NN_p SIZE	NN_q SIZE	NN_F SIZE
MAF	5, 10, 20	2×512, 2×1024, 2×2048	-	-	-
CIF-MAF	5, 10	2×512	$2 \times 128, 2 \times 512$	$2 \times 128, 2 \times 512$	2×128

Table C.7: CIF-NSF configurations for all tabular datasets. The number of hidden features in the autoregressive network is referred to as n_h.

	$\mathtt{NN}_p$ SIZE	$\mathtt{NN}_q$ SIZE	$\mathtt{NN}_F$ SIZE	n_h VS. BASELINE
CIF-NSF-1 (MINIBOONE)	3×50	2×10	3×25	FEWER
CIF-NSF-1 (NON-MINIBOONE)	3×200	2×10	3×100	FEWER
CIF-NSF-2	3×200	2×10	3×100	SAME

Table C.8: Mean $\pm$ standard error of average test set log-likelihood (higher is better). Best performing runs are shown in bold. CIF-Id-1 had $s \equiv 0$ and $t = \mathrm{Id}$. CIF-Id-2 had $s \equiv 0$ and $t = \mathtt{NN}_F$. CIF-Id-3 had $(s, t) = \mathtt{NN}_F$.

	POWER	GAS	HEPMASS	MINIBOONE
CIF-ID-1 ($\mathtt{NN}_q = 10 \times 2$)	0.43 ± 0.01	10.92 ± 0.10	-17.06 ± 0.05	-11.26 ± 0.03
CIF-ID-1 ($\mathtt{NN}_q = 100 \times 4$)	0.42 ± 0.01	10.86 ± 0.16	-17.44 ± 0.09	-10.91 ± 0.04
CIF-ID-2 ($\mathtt{NN}_q = 10 \times 2$)	0.45 ± 0.01	10.43 ± 0.08	-17.63 ± 0.10	-11.13 ± 0.08
CIF-ID-2 ($\mathtt{NN}_q = 100 \times 4$)	0.47 ± 0.01	10.89 ± 0.18	-17.51 ± 0.09	-10.75 ± 0.07
CIF-ID-3 ($\mathtt{NN}_q = 10 \times 2$)	$\mathbf{0.50 \pm 0.01}$	$\mathbf{11.32 \pm 0.14}$	-17.08 ± 0.02	-10.45 ± 0.04
CIF-ID-3 ($\mathtt{NN}_q = 100 \times 4$)	$\mathbf{0.50 \pm 0.01}$	$\mathbf{11.58 \pm 0.12}$	$\mathbf{-16.68 \pm 0.07}$	$\mathbf{-10.01 \pm 0.04}$

the size $\mathtt{NN}_p$ and $\mathtt{NN}_F$ by a factor of 4 as per Durkan et al. (2019)[17]. We trained both NSFs and CIF-NSFs for a number of training epochs corresponding to the number of training steps divided by the number of batches in the training set, i.e.

$$n_e = \lceil n_s / (n_t / n_b) \rceil ,$$

where n_e is the number of epochs, n_s is the number of training steps, n_b is the batch size, and n_t is the number of training data points. Note that n_s and n_b are from Durkan et al. (2019, Table 5), and n_t is fixed by the pre-processing steps from Papamakarios et al. (2017). We then evaluated the test-set performance of each model after the pre-specified number of epochs, averaging across three seeds, and put the results in Table 1. We again average randomness across seeds, rather than across points in the test set, as discussed in the previous section.

We quickly note here that we selected our parameters after trying a few settings on various UCI datasets. There were other settings which performed better for individual datasets that are not included here, as we would like the proposed configurations to be as homogeneous as possible. It appeared as though the NSF models were already fairly good at modelling the data, which allowed us to make $\mathtt{NN}_q$ much smaller while still achieving good inference.

We also should note that we wrapped our code around the NSF bijection code from `https://github.com/bayesiains/nsf`. We also disable weight decay in all of these experiments without observing any problems with convergence.

C.1.4 ABLATING f

We ran ablation experiments to gain some insight into the relative importance of f in (9). In particular, we considered a 10 layer model ($L = 10$) where at each layer U_ℓ had the same dimension as the data and $f = \mathrm{Id}$ was the identity. We refer to this model as CIF-Id.

We considered three parameterisations of CIF-Id. The first had $s \equiv 0$ and $t = \mathrm{Id}$, which from our choice (10) of $p_{U|Z}$ corresponds to stacking the following generative process:

$$\begin{aligned} Z &\sim P_Z \\ \epsilon &\sim \mathrm{Normal}(0, I_d) \\ X &\coloneqq Z - \mu^p(Z) - e^{\varsigma^p(Z)} \odot \epsilon. \end{aligned} \tag{C.1}$$

[17] Indeed, there was no choice of n_h which would allow us to achieve the same number of parameters as the baseline for the models noted in row 2 of Table C.7.

Observe this generalise ResFlows, since (5) can be realised by sending $\varsigma^p \to -\infty$ and having $\mu^p < 1$. Accordingly, we took $\mathtt{NN}_p$ to be a 4×128 MLP to match the size of the residual blocks used in our tabular ResFlow experiments.

The second CIF-Id parameterisation had $s \equiv 0$ and $t = \mathtt{NN}_F$, which amounts to replacing (C.1) with

$$X := Z - t\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right).$$

To align with the first CIF-Id, we took $\mathtt{NN}_F$ and $\mathtt{NN}_p$ to be 2×128 MLPs, and zeroed out the s output of $\mathtt{NN}_F$ to obtain $s \equiv 0$. The third parameterisation had $(s, t) = \mathtt{NN}_F$, which replaces (C.1) with

$$X := \exp\left(-s\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right)\right) \odot Z - t\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right).$$

Again, we took $\mathtt{NN}_F$ and $\mathtt{NN}_p$ to be 2×128 MLPs in this case.

We ran all configurations with two different choices of $\mathtt{NN}_q$: a 2×10 MLP as in our tabular ResFlow experiments, as well as a 4×100 MLP. The results are given in Table C.8.[18] Observe that these models performed comparably or better than the 100-layer ResFlows, but worse than the CIF-ResFlows and CIF-MAFs in Table 1. As discussed in Section 4.1.1, we conjecture this occurs because a CIF-Id requires greater complexity from $p_{U|Z}$ to make up for its simple choice of f, which in turn makes inference harder and hence the ELBO (14) looser, resulting in a poorer model that is learned overall. Likewise, note that the best performance in all cases was obtained when $(s, t) = \mathtt{NN}_F$. This provides some justification for the generality of our choice of (9), as opposed to simpler alternatives that omit s or t.

C.2 Image Experiments

In all our image experiments we applied the same uniform dequantisation scheme as Theis et al. (2016), after which we applied the logit transform of Dinh et al. (2017) with $\alpha = 10^{-5}$ for Fashion-MNIST and $\alpha = 0.05$ for CIFAR10.

C.2.1 RESFLOW

For our baseline ResFlow experiments we used the same architecture as Chen et al. (2019). In particular, our convolutional residual blocks (denoted Conv-ResBlock) had the form

$$\text{LipSwish} \to 3 \times 3 \text{ Conv} \to \text{LipSwish} \to 1 \times 1 \text{ Conv} \to \text{LipSwish} \to 3 \times 3 \text{ Conv},$$

while our fully connected residual blocks (denoted FC-ResBlock) had the form

$$\text{LipSwish} \to \text{Linear} \to \text{LipSwish} \to \text{Linear},$$

with a residual connection from the input to the output in both cases. The overall architecture of the flow in all cases was:

$$\text{Image} \to \text{LogitTransform}(\alpha) \to k\times \text{ Conv-ResBlock} \to [\text{Squeeze} \to k\times \text{ Conv-ResBlock}] \times 2 \to 4\times \text{ FC-ResBlock},$$

where the Squeeze operation was as defined by Dinh et al. (2017). Like Chen et al. (2019), we used ActNorm layers (Kingma & Dhariwal, 2018) before and after each residual block.

Due to computational constraints, the models we considered were smaller than those used by Chen et al. (2019). In particular, our smaller ResFlow models used 128 hidden channels in their Conv-ResBlocks, 64 hidden channels in the linear layers of their FC-ResBlocks, and had $k = 4$. Our larger ResFlow models used 256 hidden channels in their Conv-ResBlocks, 128 hidden channels in the linear layers of their FC-ResBlocks, and had $k = 6$. In contrast, Chen et al. (2019) used 512 hidden channels in their Conv-ResBlocks, 128 hidden channels in their FC-ResBlocks, and had $k = 16$.

As described for our tabular experiments, we used the same estimation scheme as Chen et al. (2019). Additionally:

- We took $\kappa = 0.98$;
- We used the Neumann gradient series expression for the log Jacobian (Chen et al., 2019, (8)) and computed gradients in the forward pass (Chen et al., 2019, (9)) to reduce memory overhead;

[18] Due to computational constraints we did not run these experiments on BSDS300.

- We used an adaptive rather than a fixed number of power iterations for spectral normalisation (Gouk et al., 2018), with a tolerance of 0.001;

For the CIF-ResFlows, we augmented the smaller baseline ResFlow by treating each composition of ActNorm $\to$ ResBlock, as well as the final ActNorm, as an instance of f in (9). Each $\mathtt{NN}_F$, $\mathtt{NN}_p$, and $\mathtt{NN}_q$ was a ResNet (He et al., 2016a;b) consisting of 2 residual blocks with 32 hidden channels (denoted 2×32). We gave each U_ℓ the same shape as a single channel of Z_ℓ, and upsampled to the dimension of Z_ℓ by adding channels at the output of each $\mathtt{NN}_F$. Note that we did not experiment with using the larger baseline ResFlow model as the basis for a CIF.

For all models we used a learning rate of 10^{-3} and a batch size of 64.

Figure C.3 through to Figure C.8 show samples synthesised from the ResFlow and CIF-ResFlow density models trained on MNIST and CIFAR-10.

C.2.2 REALNVP

For our RealNVP-based image experiments, we took the baseline to be a RealNVP with the same architecture used by Dinh et al. (2017) for their CIFAR-10 experiments. In particular, we used 10 affine coupling layers with the corresponding alternating channelwise and checkerboard masks. Each coupling layer used a ResNet (He et al., 2016a;b) consisting of 8 residual blocks of 64 channels (denoted 8×64). We replicated the multi-scale architecture of Dinh et al. (2017), squeezing the channel dimension after the first 3 coupling layers, and splitting off half the dimensions after the first 6. This model had 5.94M parameters for Fashion-MNIST and 6.01M parameters for CIFAR-10.

For the CIF-RealNVP, we considered each affine coupling layer to be an instance of f in (9). When choosing the size of our networks, we sought to maintain roughly the same depth over which gradients were propagated as in the baseline. To this end, our coupling networks were 4×64 ResNets, each $\mathtt{NN}_p$ and $\mathtt{NN}_q$ were 2×64 ResNets, and each $\mathtt{NN}_F$ was a 2×8 ResNet. We gave each U_ℓ the same shape as a single channel of Z_ℓ, and upsampled to the dimension of Z_ℓ by adding channels at the output of $\mathtt{NN}_F$. Our model had 5.99M parameters for Fashion-MNIST and 6.07M parameters for CIFAR-10.

For completeness, we also trained a RealNVP model with coupler networks of size 4×64 to match our CIF-RealNVP configuration. This model had 2.99M parameters for Fashion-MNIST and 3.05M for CIFAR-10.

In all cases for these experiments we used a learning rate of 10^{-4} and a batch size of 100.

Figure C.9 through to Figure C.14 show samples synthesised from the RealNVP and CIF-RealNVP density models trained on Fashion-MNIST and CIFAR-10.

C.3 2-D Experiments

To gain intuition about our model, we ran experiments on some simple 2-D datasets. For the datasets in Figure 1, we used a 10-layer ResFlow, a 100-layer ResFlow, and 10-layer CIF-ResFlow. For the CIF-ResFlows we took $d_{\mathcal{U}} = 1$. Other architectural and training details were the same as for the tabular experiments described in Section 5.1 and Section C.1.1. The resulting average test set log-likelihoods for the top dataset were:

- -1.501 for the 10-layer ResFlow
- -1.419 for the 100-layer ResFlow
- -1.409 for the 10-layer CIF-ResFlow

The final average test set log likelihoods for the bottom dataset were:

- -2.357 for the 10-layer ResFlow
- -2.287 for the 100-layer ResFlow
- -2.275 for the 10-layer CIF-ResFlow

Note that in both cases the CIF-ResFlow slightly outperformed the 100-layer ResFlow.

We additionally ran several experiments comparing a baseline MAF against a CIF-MAF on the 2-D datasets shown in Figure C.15. The baseline MAFs had 20 autoregressive layers, while the CIF-MAFs had 5. The network used at each layer had 4 hidden layers of 50 hidden units (denoted 4×50). For the CIF-MAF, we took $d_{\mathcal{U}} = 1$, and used 2×10 MLPs for $\mathtt{NN}_F$ and 4×50 MLPs for $\mathtt{NN}_p$ and $\mathtt{NN}_q$. In total the baseline MAF had 160160 parameters, while our model had 119910 parameters.

The results of these experiments are shown in Figure C.15. Observe that CIF-MAF consistently produces a more faithful representation of the target distribution than the baseline, and in all cases achieved higher average test set log probability. A failure mode of our approach is exhibited in the spiral dataset, where our model still lacks the power to fully capture the topology of the target. However, we did not find it difficult to improve on this: by increasing the size of $\mathtt{NN}_p$ to 8×50 (and keeping all other parameters fixed), we were able to obtain the result shown in Figure C.16. This model had a total of 221910 parameters. We also tried a larger MAF model with autoregressive networks of size 8×50, (obtaining 364160 parameters total). This model diverged after approximately 160 epochs. The result after 150 epochs is shown in Figure C.16.

Figure C.3: Synthetic MNIST samples generated by the small baseline ResFlow model

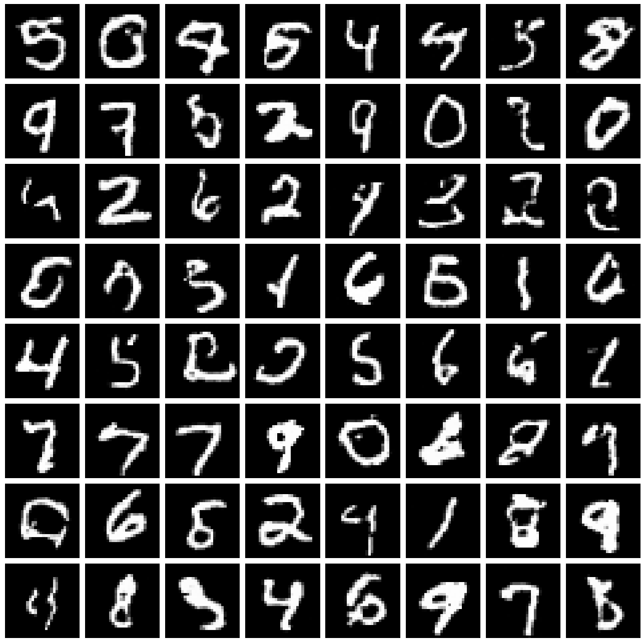

Figure C.4: Synthetic MNIST samples generated by the large baseline ResFlow model

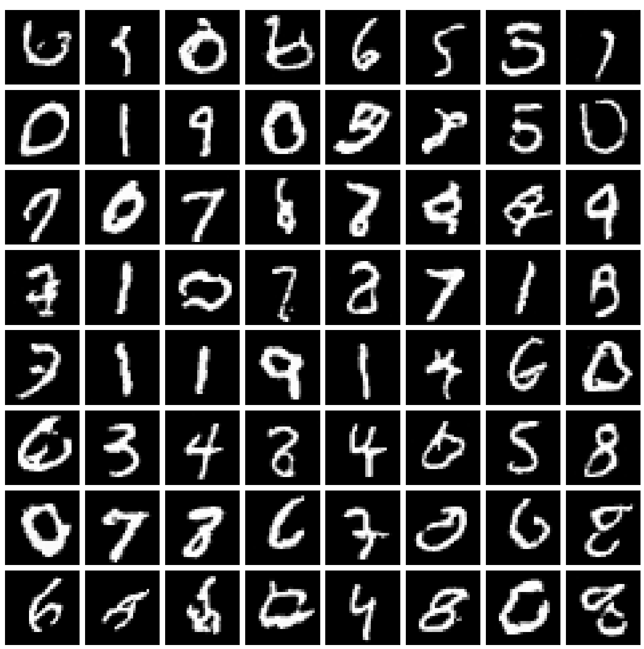

Figure C.5: Synthetic MNIST samples generated by the CIF-ResFlow model

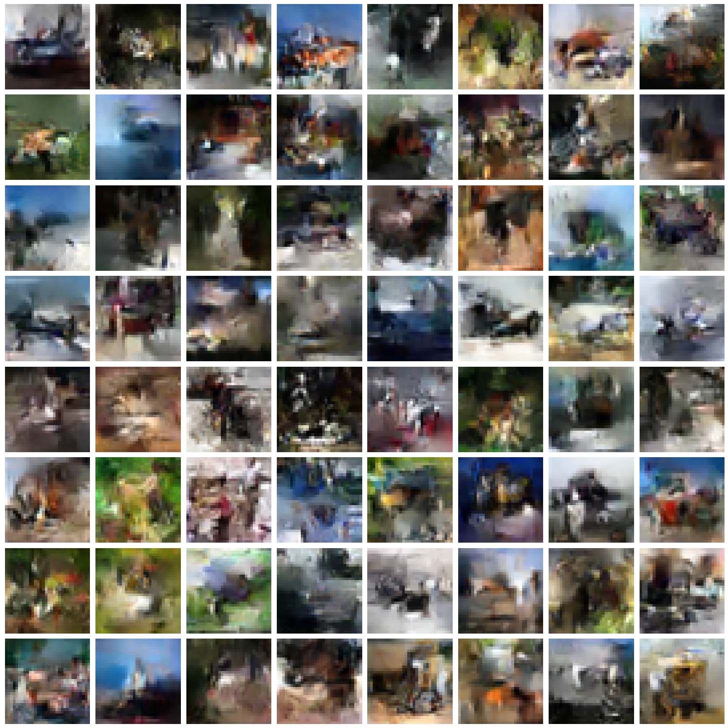

Figure C.6: Synthetic CIFAR-10 samples generated by the small baseline ResFlow model

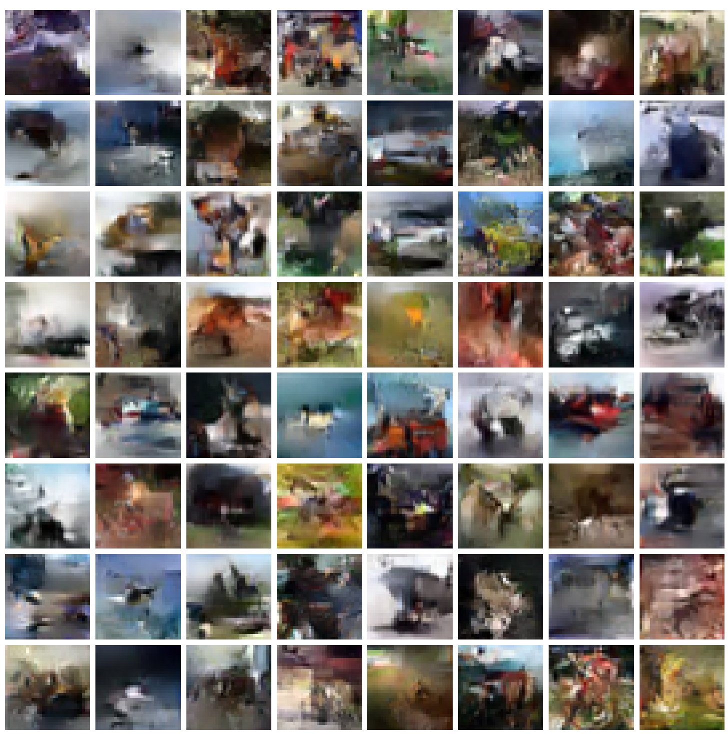

Figure C.7: Synthetic CIFAR-10 samples generated by the large baseline ResFlow model

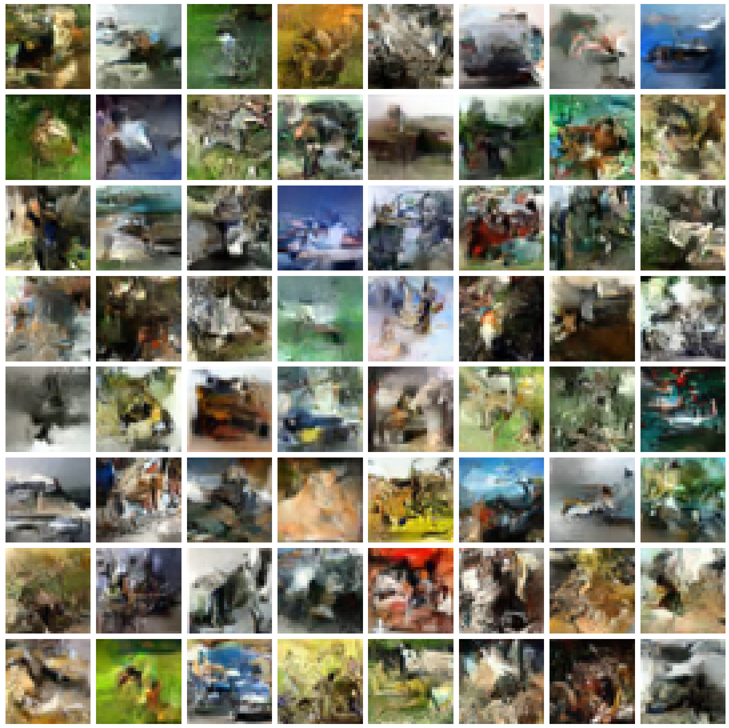

Figure C.8: Synthetic CIFAR-10 samples generated by the CIF-ResFlow model

Figure C.9: Synthetic Fashion-MNIST samples generated by RealNVP with coupling networks of size 4×64

Figure C.10: Synthetic Fashion-MNIST samples generated by RealNVP with coupling networks of size 8×64

Figure C.11: Synthetic Fashion-MNIST samples generated by CIF-RealNVP

Figure C.12: Synthetic CIFAR-10 samples generated by RealNVP with coupling networks of size 4×64

Figure C.13: Synthetic CIFAR-10 samples generated by RealNVP with coupling networks of size 8×64

Figure C.14: Synthetic CIFAR-10 samples generated by CIF-RealNVP

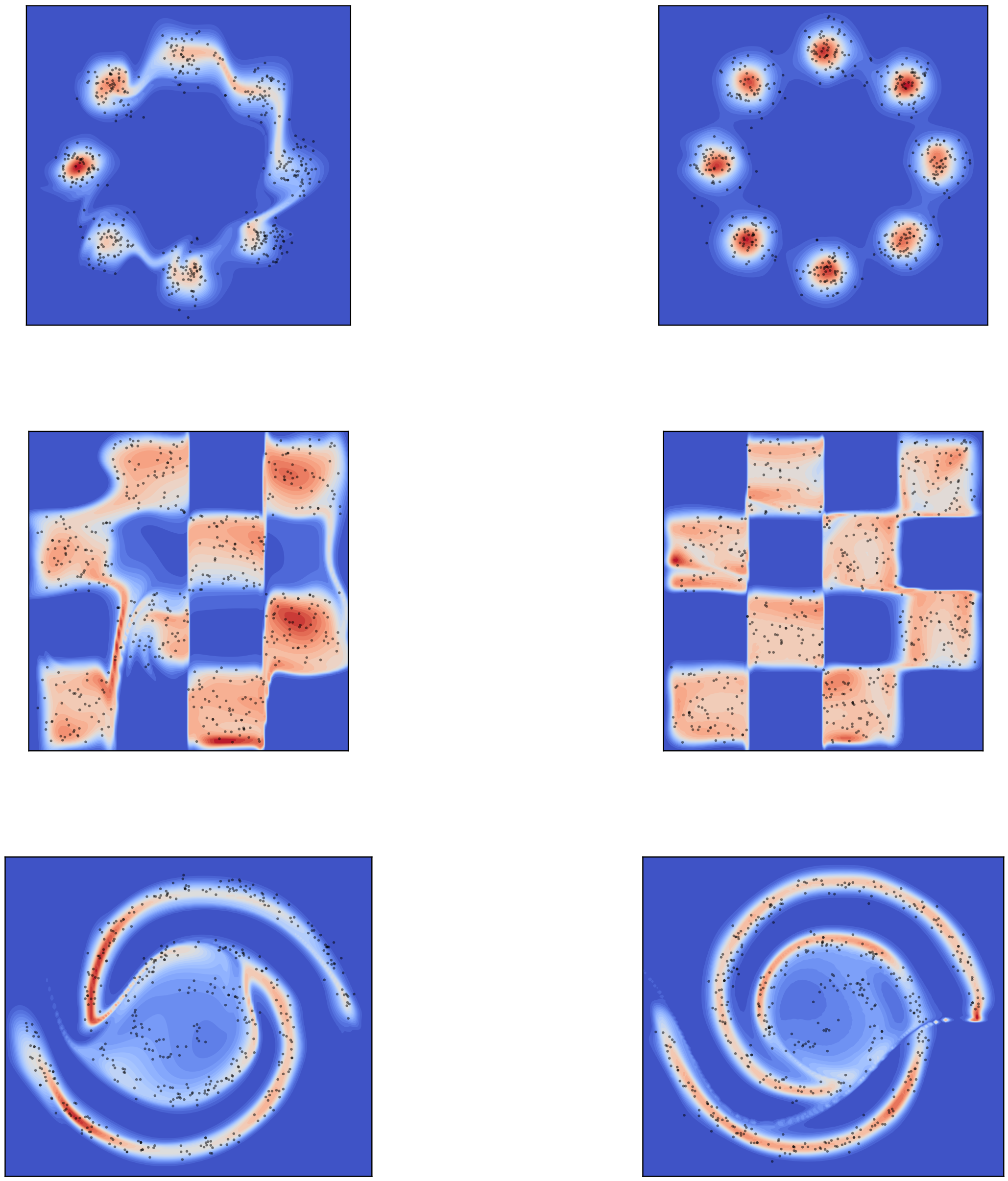

Figure C.15: Density models learned by a standard 20 layer MAF (left) and by a 5 layer CIF-MAF (right) for a variety of 2-D target distributions. Samples from the target are shown in black.

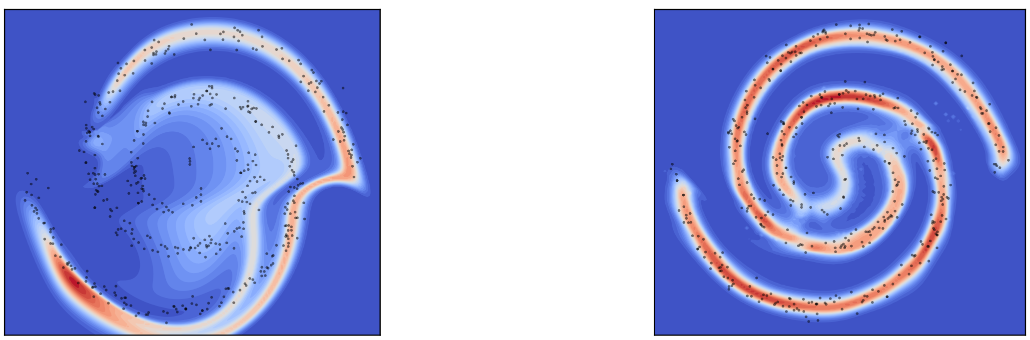

Figure C.16: Density models learned by a larger 20 layer MAF (left) and a larger 5 layer CIF-MAF (right) for the spirals dataset.