AUTM Flow: Atomic Unrestricted Time Machine for Monotonic Normalizing Flows (Supplementary material)

Difeng Cai[1] **Yuliang Ji**[1] **Huan He**[2] **Qiang Ye**[3] **Yuanzhe Xi**[1]

[1]Department of Mathematics, Emory University, Atlanta, GA, USA
[2]Department of Computer Science, Emory University, Atlanta, GA, USA
[3]Department of Mathematics, University of Kentucky, Lexington, KY, USA

A PROOFS

Proof of Theorem 1. Write $v = v(t, x)$ as a function of t and x. In fact,

$$v(t, x) = x + \int_0^t g(v(t, x), t)dt.$$

Define $u(t) = \frac{\partial v}{\partial x}$. It follows from the formula of v that

$$u(t) = \frac{\partial v}{\partial x} = 1 + \int_0^t \frac{\partial g}{\partial v}\frac{\partial v}{\partial x}dt. \tag{1}$$

Then we see that

$$\frac{du}{dt} = \frac{\partial g}{\partial v}\frac{\partial v}{\partial x} = \frac{\partial g}{\partial v}u.$$

This implies that

$$u(t) = C\exp\left(\int_0^t \frac{\partial g}{\partial v}dt\right).$$

Recall from (1) that $u(0) = 1$, so $C = 1$. Since $q(x) = v(1, x)$, we now conclude that

$$q'(x) = \frac{\partial v}{\partial x}(1, x) = u(1) = \exp\left(\int_0^1 \frac{\partial g}{\partial v}dt\right).$$

□

Proof of Theorem 2. Obviously, monotonicity implies invertibility, so it suffices to show that q is strictly increasing. There are several ways to prove this. The simplest way is to use Theorem 1 to see that $q'(x) > 0$, so q must be increasing. Below we present a different proof without using the analytic expression of $q'(x)$.

Let $v_x(t)$ denote the function in (4) with $v(0) = x$, where $g(v, t)$ is continuous in t and uniformly Lipschitz continuous in v. We need to show that for any $x < x'$, there holds $q(x) < q(x')$. We prove this by contradiction. Assume that there exist $x < x'$ such that $q(x) \geq q(x')$. There are two cases to consider: $q(x) = q(x')$ and $q(x) > q(x')$.

Case 1: $q(x) = q(x') = C$ for some constant C.
In this case, $v_x(1) = v_{x'}(1) = C$. Define $w_a(t) = v_a(1-t)$ for any $a \in \mathbb{R}$ and $t \in [0, 1]$. Then it is easy to see that $w_x(t)$, $w_{x'}(t)$ are both solutions to the ODE

$$\frac{dw}{dt} = -g(w(t), 1-t), \quad w(0) = C, \quad t \in [0, 1]. \tag{2}$$

Note that $w_x(t)$ and $w_{x'}(t)$ are two different solutions of (2) because $w_x(1) = x < x' = w_{x'}(1)$. This contradicts the uniqueness of solution to the ODE (which is well-posed since g is Lipschitz) and we conclude that the assumption $q(x) = q(x')$ can *not* hold.

Case 2: $q(x) > q(x')$.
In this case, we have $v_x(1) > v_{x'}(1)$ and $v_x(0) < v_{x'}(0)$. Applying intermediate value theorem to $v_x(t) - v_{x'}(t)$ yields that there exists $\tau \in (0, 1)$ such that

$$v_x(\tau) = v_{x'}(\tau) = C$$

for some constant C. Similar to Case 1, if we define $w_a(t) = v_a(\tau - t)$ for $t \in [0, \tau]$, then we can deduce that the ODE

$$\frac{dw}{dt} = -g(w(t), 1-t), \quad w(0) = C \tag{3}$$

has two different solutions $w_x(t)$ and $w_{x'}(t)$ as $w_x(\tau) = x < x' = w_{x'}(\tau)$, which contradicts the well-posedness of (3).

We now conclude that the inequality $q(x) \geq q(x')$ can *not* hold. Consequently, $q(x)$ must be strictly increasing and the proof is complete. □

Proof of Theorem 3. Since the set of increasing Lipschitz continuous functions is dense in $\mathcal{M}$, it suffices to consider Lipschitz functions in $\mathcal{M}$.

We need to show that, given an arbitrary increasing Lipschitz continuous function $\phi(x)$, there exists a family of AUTM bijections $\{q_s(x)\}_{s>0} \subset \mathcal{Q}$ that converge compactly to $\phi(x)$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

as $s \to 0$, i.e. $q_s|_K \to \phi|_K$ uniformly on any compact set $K \subset \mathbb{R}$ as $s \to 0$. We construct q_s as follows.

For $s > 0$, we define

$$g_s(v, t) = \phi(v(te^{-\frac{1}{s}})) - v(te^{-\frac{1}{s}}).$$

Then we define

$$q_s(x) = x + \int_0^1 g_s(v_s, t)dt,$$

where

$$v_s(t) := x + \int_0^t g_s(v_s, z)dz.$$

We prove that for any x, $q_s(x)$ converges to $\phi(x)$ as $s \to 0$. In fact, we will show that the convergence rate is $O(e^{-\frac{1}{s}})$.

First we prove that $\max_{t\in[0,1]} |v_s(t)|$ is uniformly bounded for $s > 0$. To show this, we investigate the differential equation that $v_s(t)$ satisfies. For notational convenience, we drop the subscript s in v_s in the proof below. The dependence on s will be stated explicitly when needed. Note that

$$\frac{dv}{dt} = g_s(v, t) = \phi(v(te^{-\frac{1}{s}})) - v(te^{-\frac{1}{s}}), \quad v(0) = x.$$

Equivalently, by a change of variable $\tau = te^{-\frac{1}{s}}$, we have

$$\frac{dv}{d\tau} = e^{\frac{1}{s}}[\phi(v(\tau)) - v(\tau)], \quad v(0) = x. \tag{4}$$

In the following, we first analyze the property of the solution v to the initial value problem in (4), and then we prove the uniform boundedness.

Results on initial value problem (4). We prove in the following that the solution $v(\tau)$ of (4) must fall into one of the three cases below:

$$\begin{aligned} &(\text{I})v'(\tau) = 0 \text{ for all } \tau; \\ &(\text{I})v'(\tau) > 0 \text{ for all } \tau; \\ &(\text{I})v'(\tau) < 0 \text{ for all } \tau. \end{aligned}$$

Case (I): We first show that if $v'(a) = 0$ for some $a \geq 0$, then $v'(\tau) = 0$ everywhere. In fact, $v'(a) = e^{\frac{1}{s}}[\phi(v(a)) - v(a)] = 0$ implies $\phi(v(a)) - v(a) = 0$. Note that $v(\tau) = v(a)$ is then an equilibrium solution. Since ϕ is Lipschitz, from the uniqueness theorem of the initial value problem, $v(\tau) = v(a)$ is the only solution and thus $v'(\tau) = 0$ for all τ.

Case (II): If $v'(0) > 0$, then it is easy to see that $v'(\tau) > 0$ for all τ. In fact, if $v'(a) = 0$ for some $a > 0$, then we know from the result above that $v'(0) = 0$, a contradiction; if $v'(a) < 0$ for some $a > 0$, then because $v'(\tau)$ is continuous, intermediate value theorem implies that there must be a point $b \in (0, a)$ such that $v'(b) = 0$, which then implies $v'(0) = 0$ according to the result above, a contradiction.

Case (III): If $v'(0) < 0$, a similar argument shows that $v'(\tau) < 0$ for all τ.

Thus we conclude that there can only be three cases for the solution $v(\tau)$, as shown in (I), (II), (III).

Proof of uniform boundedness of $|v_s|$. Next we show that every solution v of (4) is uniformly bounded in s.

If v falls into Case (I), it is easy to see that $v(\tau) = v(0) = x$ independent of s, thus uniformly bounded.

If v falls into Case (II), then $\phi(v) - v > 0$, and the equation in (4) can be equivalently written as

$$\frac{dv}{\phi(v) - v} = e^{\frac{1}{s}}d\tau.$$

Let $G(x)$ denote the anti-derivative of $\frac{1}{\phi(x)-x}$. Then it follows that

$$G(v) = \int_0^\tau e^{\frac{1}{s}}d\tau + A = \tau e^{\frac{1}{s}} + A = t + A,$$

where A is a constant independent of s. In fact, setting $t = 0$ (or equivalently, $\tau = 0$) yields that

$$G(v(0)) = G(x) = 0 + A = A.$$

Thus $A = G(x)$. Since $G'(v) = \frac{1}{\phi(v)-v} = \frac{1}{v'} \cdot e^{\frac{1}{s}} > 0$, we know from inverse function theorem that G^{-1} exists and is continuous and strictly increasing. Therefore,

$$v = G^{-1}(G(v)) = G^{-1}(t + A) = G^{-1}(t + G(x)) \tag{5}$$

is uniformly bounded in s and t since G is independent of s, t, and $t + G(x) \in [G(x), 1 + G(x)]$ with $t \in [0, 1]$. Therefore, in Case (II), $|v_s(t)|$ is uniformly bounded in s and t.

If v falls into Case (III), the uniform boundedness of $|v_s(t)|$ can be derived analogously as in Case (II).

Now we conclude that $\max_{t\in[0,1]} |v_s(t)|$ is uniformly bounded with respect to s.

Proof of convergence $q_s(x) \to \phi(x)$. Next we show that the uniform boundedness of $|v_s|$ implies the convergence $q_s(x) \to \phi(x)$ as $s \to 0$. Since ϕ is Lipschitz, we see that $|g_s(v_s, t)|$ is also uniformly bounded in s and t. Thus $M := \sup_{s>0, t\in[0,1]} |g_s(v_s, t)| < \infty$. Note that

$$\phi(x) = x + \int_0^1 \phi(x) - x dt.$$

Let L denote the Lipschitz constant of ϕ. Then we deduce from the definition of q_s and v_s that

$$\begin{aligned}|q_s(x)-\phi(x)| &= \left|\int_0^1 \phi(v_s(te^{-\frac{1}{s}}))-\phi(x)+x-v_s(te^{-\frac{1}{s}})dt\right| \\ &\le (L+1)\int_0^1 |v_s(te^{-\frac{1}{s}})-x|dt \\ &= (L+1)\int_0^1 \left|\int_0^{te^{-\frac{1}{s}}} g_s(v_s,z)dz\right| dt \\ &\le (L+1)\int_0^1 Mte^{-\frac{1}{s}}dt \\ &= (L+1)\frac{M}{2}e^{-\frac{1}{s}} \to 0, \quad s\to 0.\end{aligned}$$

This proves the pointwise convergence $q_s(x) \to \phi(x)$ as $s \to 0$. Since K is compact and $|v_s|$ is continuous in x (see (5) for example), it can be deduced from the argument above that $|v_s|$ is in fact uniformly bounded in $s > 0, t \in [0,1]$ and $x \in K$. Thus the convergence proof still holds with constant M chosen as an upper bound of $|g_s|$ over $s > 0, t \in [0,1]$ and $x \in K$. The rate of convergence is still $O(e^{-\frac{1}{s}})$ as $s \to 0$.

The proof of Theorem 3 is now complete.

□

Proof of Theorem 4. This is an immediate result of Theorem 3. Since compact convergence implies pointwise convergence, it suffices to show the compact convergence. According to the proof of Theorem 3, for each F_k (where each entry is a monotone continuous function), we can construct a family of triangular AUTM transformations $T_{s,k}$ (parametrized by $s > 0$) that converge compactly to F_k with a rate of $O(e^{-\frac{1}{s}})$ as $s \to 0$. Then it follows immediately that $T_s := T_{s,1} \circ T_{s,2} \circ \cdots \circ T_{s,p}$ converges compactly to $F = F_1 \circ F_2 \circ \cdots \circ F_p$ with rate $O(e^{-\frac{1}{s}})$ as $s \to 0$, which completes the proof. □

Proof of Theorem 5. The proof follows essentially the same argument as the proof of Theorem 3 in which $\kappa_s(t) = 1$. More precisely, for a general positive kernel $\kappa_s(t)$ satisfying (10), the positivity of $\kappa_s(t)$ is used in obtaining results for the initial value problem; the bounded L^1 norm of $\kappa_s(t)$ is used in proving the uniform boundedness of $|v_s|$; the asymptotic property as $s \to 0$ is used in proving the convergence $q_s \to \phi$. Therefore, similar to the proof of Theorem 3, we conclude that $q_s|_K \to \phi|_K$ uniformly as $s \to 0$. We remark that the class of kernels in (10) includes $\kappa_s(t) = 1$, the normalized Gaussian kernel $\kappa_s(t) = C_s e^{-\frac{t^2}{s}}$ with $C_s = \left(\int_0^1 e^{-\frac{z^2}{s}}dz\right)^{-1}$, and it can be computed that the convergence rate for the latter is $O(s^{-\frac{1}{2}}e^{-\frac{1}{s}})$. □

B EXPERIMENT DETAILS

B.1 HYPERPARAMETERS FOR DENSITY ESTIMATION DATASETS

We list the hyperparameters in Table 1. Hyperparameters are obtained after extensive grid search. For the number of layers, we tried 5,10,20. For the hidden layer dimensions, we tried $10d, 20d, 40d$, where d is the dimension of the vector in the dataset. We trained our model by using Adam. We stop the training process when there is no improvement on validation set in several epochs.

B.2 HYPERPARAMETERS FOR CIFAR10 AND IMAGENET32

We list the hyperparameters in Table 2. In this experiment, most hyperparameters come from [Kaiming et al., 2016]. We use 14 AUTM coupling layers with 8 residual blocks for each layer in our model. Like [Kingma and Dhariwal, 2018], before each coupling layers, there is an actnorm layer and a conv 1×1 layer. Each residual block has three convolution layers with 128 channels. Our method is trained for 100 epochs with batch size 64. We trained our model by using Adamax with Polyak.

C RECONSTRUCTION ON IMAGE DATASET

We examine the reserve step of our AUTM layer by showing the reconstruction of images. The used model is the same as the model in Section 6.2 and use CIFAR10 and ImageNet32 dataset in this experiment. We compute the inverse of our layer by using iterative method with the reverse of integral as the initial guess. As Figure 1 shows, the average L_1 reconstruction error converges in 15 steps. Also, Figure 1 shows that the reconstructed images look the same as original images.

We show the result of the reconstruction process of our method in Figure 1.

D CODE

Our code is available at `https://anonymous.4open.science/r/AUTM-2B1B`. We use some code from BNAF[De Cao et al., 2020].

References

Nicola De Cao, Wilker Aziz, and Ivan Titov. Block neural autoregressive flow. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115

Table 1: Hyperparameters for Power, GAS, Hepmass, Miniboone, BSDS300 datasets, d is the dimension of the vector in the dataset

	POWER	GAS	Hepmass	Miniboone	BSDS300
layers	10	10	10	5	10
hidden layer dimensions	40d	40d	40d	10d	40d
epochs	450	1000	500	1000	1000
batch size	256	256	256	256	128
optimizer	adam	adam	adam	adam	adam
learning rate	0.01	0.01	0.01	0.01	0.01
lr decay rate	0.5	0.5	0.5	0.5	0.5

Table 2: Hyperparameters for CIFAR-10 and ImageNet32 datasets

	CIFAR10	ImageNet32
layers	14	14
residential blocks	8	8
hidden channels	128	128
epochs	2500	50
batch size	64	64
optimizer	adamax	adamax
learning rate	0.01	0.01
lr decay	0.5	0.5
lr decay epoch	[30,60,90]	[30,60,90]

of *Proceedings of Machine Learning Research*, pages 1263–1273, 2020.

He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, pages 10215–10224, 2018.

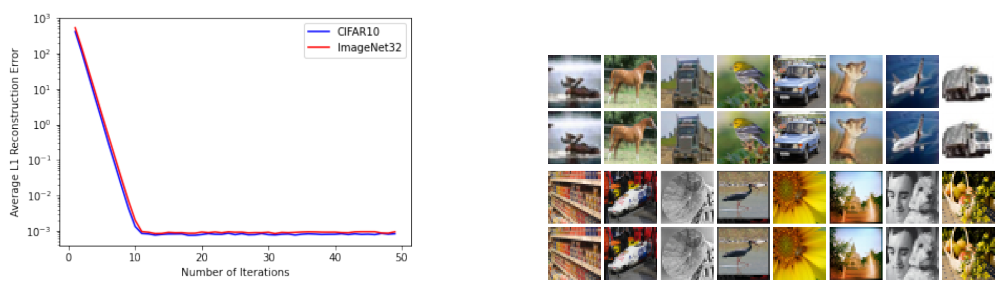

Figure 1: **Left**: The average value of the L_1 reconstruction error for 64 images. **Right**: The reconstruction of selected images in CIFAR10 and ImageNet32 dataset. The 1st, 3rd rows are the original images, and the 2nd, 4th rows are the reconstructions.