# FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models

Will Grathwohl*, Ricky T. Q. Chen*, Jesse Bettencourt, Ilya Sutskever, David Duvenaud

*Equal Contribution University of Toronto, Vector Institute, OpenAI

## Main Idea

- Construct generative model for data which replaces discrete invertible transformations with system of continuous-time dynamics using Instantaneous Change of Variables.
- Replace expensive Jacobian trace with stochastic estimate to compute unbiased log-density in *linear-time* while allowing *unrestricted* network architectures.
- Flexibility allows us to achieve better performance than previous reversible generative models.

## Normalizing Flows

Generate data $x$ by:

$$z \sim p(z) \qquad x = F_\theta(z)$$

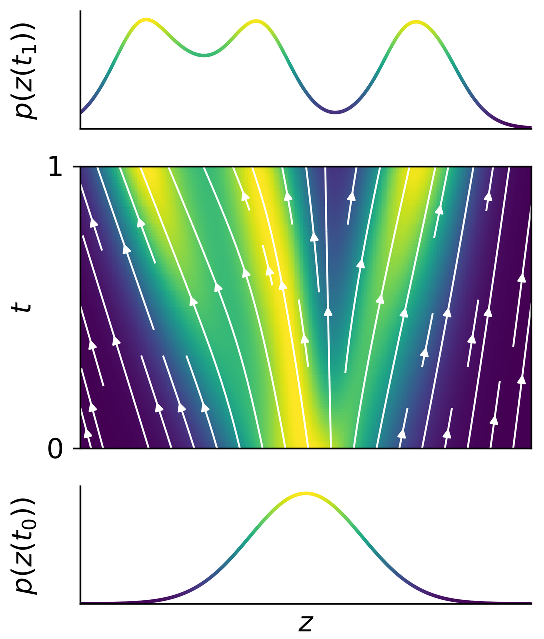
With invertible $F_\theta$ then:

$$\log p(x) = \log p_z(F_\theta^{-1}(x)) - \log \left| \frac{\partial F_\theta}{\partial x} \right| \quad (1)$$

Restricted, simple transformations allow $\log \left|\frac{\partial F_\theta}{\partial x}\right|$ to be computed efficiently.
Compose multiple simple transformations for an expressive transformation satisfying invertibility and efficient log-determinants:

$$F_\theta(z) = F_\theta^{t_1} \circ \cdots \circ F_\theta^{t_0}(z)$$

## Continuous Normalizing Flows




Model the generative process with continuous dynamics:

$$z_0 \sim p(z_0)$$
$$\frac{\partial z_t}{\partial t} = f_\theta(z_t, t)$$
$$x = z_1 = z_0 + \int_{t_0}^{t_1} f_\theta(z_t, t)\,dt$$

To obtain the density we solve the initial value problem (IVP):

$$\log p(x) = \log p(z_0) - \int_{t_0}^{t_1} \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial z_t}\right) dt \quad (2)$$

## Unbiased Log-Density Estimation

$\mathrm{Tr}(\partial f / \partial z_t)$ cannot be computed efficiently for unrestricted $f$.
We utilize two techniques to estimate it efficiently:

**Vector-Jacobian Products**: Explicitly computing the Jacobian $\frac{\partial f}{\partial z_t}$ cannot be done efficiently, but reverse-mode automatic differentiation cheaply computes $e^T \frac{\partial f}{\partial z_t}$ can be for any vector $e$.

**Stochastic Trace Estimators**: For any matrix A and a distribution $p(e)$ over vectors where $\mathbb{E}[e] = 0$, $\mathrm{Cov}[e] = I$, then:

$$\mathrm{Tr}(A) = \mathbb{E}_{p(e)}[e^T A e]$$

The Monte-Carlo estimator derived from this expectation is known as Hutchinson's estimator.
Combining these two we can build an efficient unbiased estimator

$$\mathrm{Tr}\left(\frac{\partial f}{\partial z_t}\right) = \mathbb{E}_{p(e)}\left[\underbrace{\left(e^T \frac{\partial f}{\partial z_t}\right)}_{\text{VJP}} e\right]$$

which can be combined with Equation 2 to give

$$\log p(x) = \log p(z_0) - \mathbb{E}_{p(e)}\left[\int_{t_0}^{t_1} e^T \frac{\partial f}{\partial z_t} e\right] \quad (3)$$

## Training with Adjoint Backpropagation

Given an objective of the form

$$L(z_1) = L\left(\underbrace{\int_0^1 f(z_t, t, \theta)\,dt}_{\text{Solution to IVP}}\right)$$

we can obtain $\frac{\partial L}{\partial \theta}$ for gradient-based optimization by solving another IVP.
We define a new quantity, the adjoint, $a_t$, which has dynamics $\frac{\partial a_t}{\partial t}$

$$a_t = -\frac{\partial L}{\partial z_t} \qquad \frac{\partial a_t}{\partial t} = -a_t^T \frac{\partial f(z_t, t, \theta)}{\partial z_t}$$

then solving backwards in time gives the desired gradients of the loss with respect to the parameters

$$\frac{\partial L}{\partial \theta} = \int_{t_1}^{t_0} a_t^T \frac{\partial f(z_t, t, \theta)}{\partial \theta} dt$$

This allows us to use a black-box ODE solver to compute $z_1$ and also $\partial L / \partial \theta$.

## Density Estimation: Qualitative

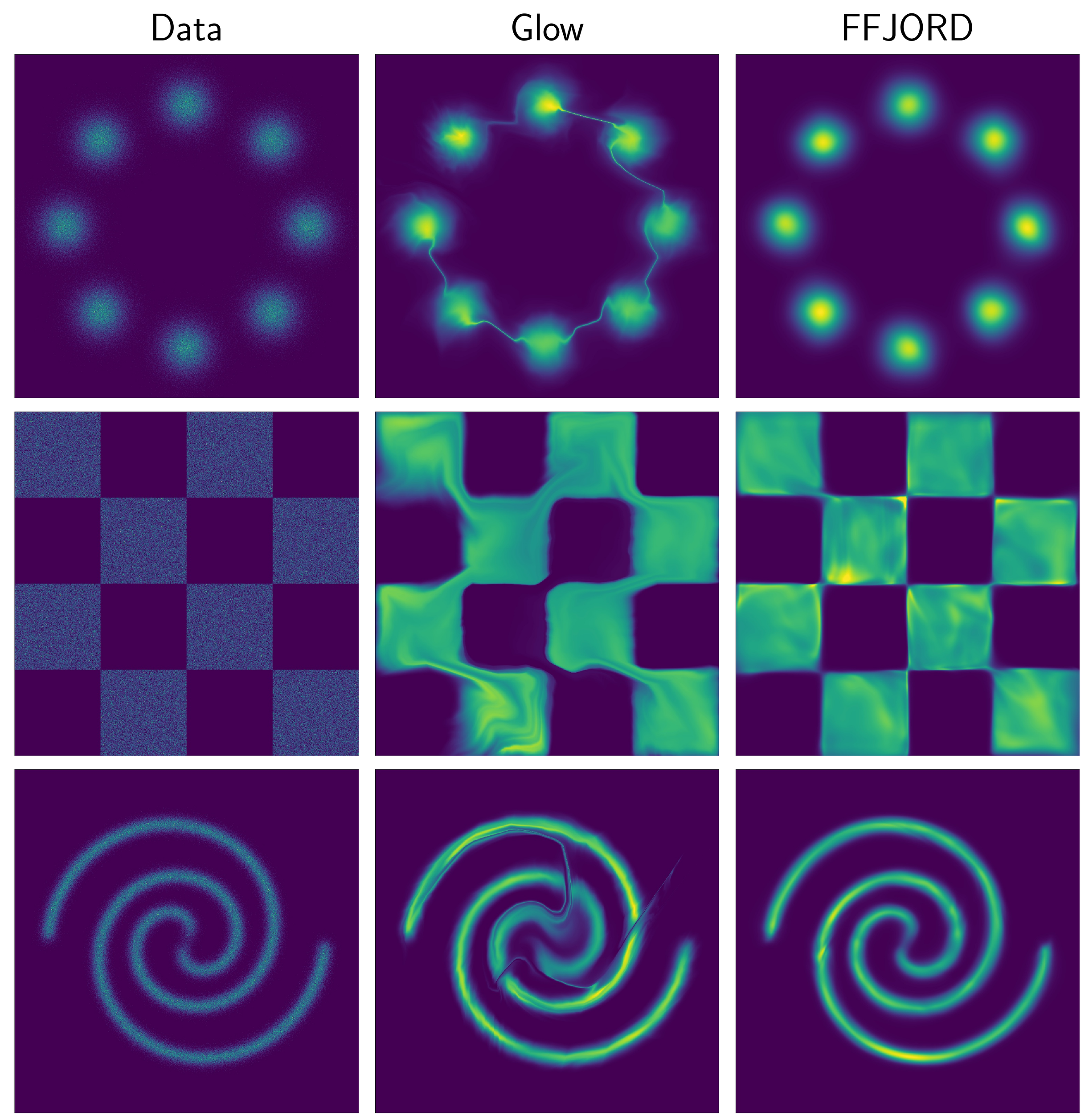



Table: Glow and FFJORD trained on 2D densities.


Table: Samples from FFJORD trained on MNIST and CIFAR10.

## Density Estimation: Quantitative

| | POWER | GAS | HEPMASS | MINIB | BSDS | MNIST | CIFAR10 |
|---|---|---|---|---|---|---|---|
| Real NVP | -.17 | -8.33 | 18.71 | 13.55 | -153.28 | 1.06 | 3.49 |
| Glow | -.17 | -8.15 | 18.92 | 11.35 | -155.07 | 1.05 | **3.35** |
| FFJORD | **-.46** | **-8.59** | **15.26** | **10.43** | **-157.67** | **0.99*** | 3.40 |
| MADE | 3.08 | -3.56 | 20.98 | 15.59 | -148.85 | 2.04 | 5.67 |
| MAF | -.24 | -10.08 | 17.70 | 11.75 | -155.69 | 1.89 | 4.31 |
| TAN | -.48 | -11.19 | 15.12 | 11.01 | -157.03 | - | - |
| DDSF | -.62 | -11.96 | 15.09 | 8.86 | -157.73 | - | - |

Table: Density estimation experiments. Negative log-likelihood on test set.

## Variational Inference

| | MNIST | Omniglot | Frey Faces | Caltech |
|---|---|---|---|---|
| No Flow | 86.55 ± .06 | 104.28 ± .39 | 4.53 ± .02 | 110.80 ± .46 |
| Planar | 86.06 ± .31 | 102.65 ± .42 | 4.40 ± .06 | 109.66 ± .42 |
| IAF | 84.20 ± .17 | 102.41 ± .04 | 4.47 ± .05 | 111.58 ± .38 |
| Sylvester | 83.32 ± .06 | 99.00 ± .04 | 4.45 ± .04 | 104.62 ± .29 |
| FFJORD | **82.82** ± **.01** | **98.33** ± **.09** | **4.39** ± **.01** | **104.03** ± **.43** |

Table: Variational inference experiments. Negative ELBO on test set.

## Advantages / Disadvantages

Advantages:

- Guaranteed inverse by reversing order of integration, regardless of model parameterization
- Efficient, unbiased log-probability estimation without restricting the Jacobian of the transformation
- Does not require dimension splitting or ordering choices
- Reversible generative models can now be defined with standard neural network architectures

Disadvantages:

- Must rely on adaptive numerical ODE solvers for stable training
- Computation time determined by solver, not user
- Currently 4-5x slower than other reversible generative models (Glow, Real-NVP)

## Conclusion

- We have presented a new class of reversible generative models.
- Our model utilizes continuous dynamics to side-step many issues in previous discrete-time reversible models.
- Our model achieves state-of-the-art results on a number of challenging density estimation and variational inference benchmarks.
- Our approach demonstrates the utility of using continuous-time dynamics and should motive further development of Neural ODEs.

## References

[1]Rezende, Mohamed. "Variational Inference Normalizing Flows." (2015)
[2]Dinh, Sohl-Dickstein, Bengio. "Density estimation Real NVP." (2016)
[3]Chen, Rubanova, Bettencourt, Duvenaud. "Neural ODEs." (2018)
[4]Kingma, Dhariwal. "Glow: Generative Flow with Invertible 11 Convolutions." (2018)