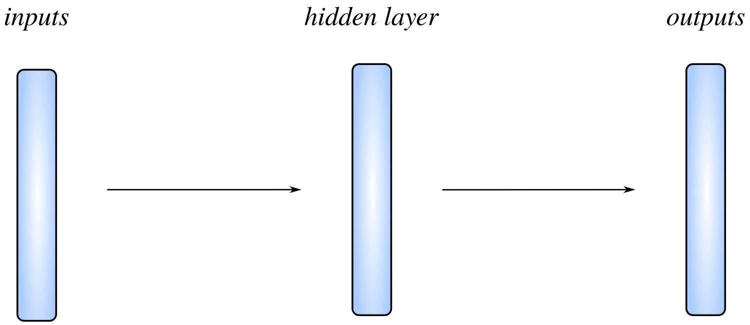
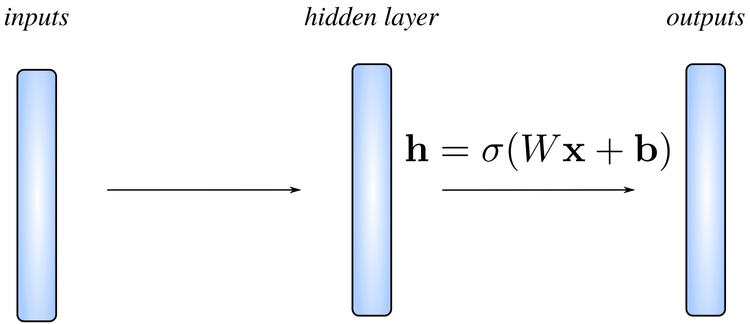
Augmented neural ODEs

Olof Mogren, PhD, RISE

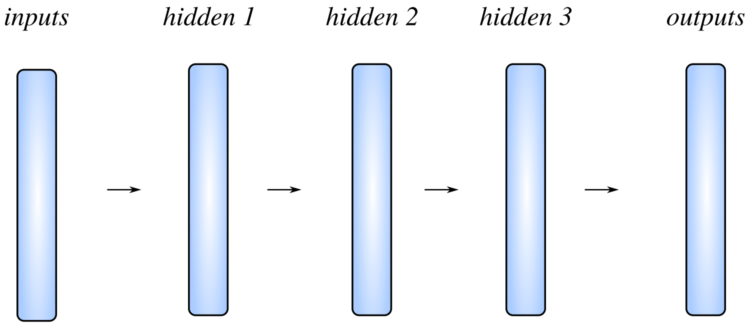
Neural networks



Neural networks



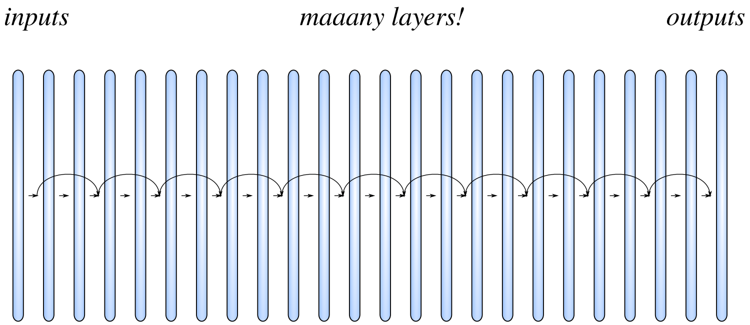
Deep neural networks



Deep nets

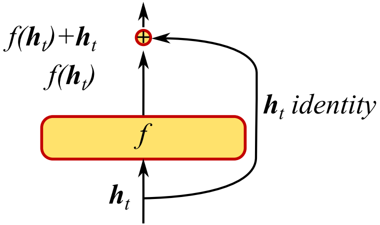
- More layers ->
  - Decreased length of step taken in each layer

Residual neural networks



Residual connections

- $\mathbf{h}_{t+1} = \mathbf{h}_t + f(\mathbf{h}_t, \theta_t)$
- A layer learns the difference between $\mathbf{h}_t$ and $\mathbf{h}_{t+1}$
- Deeper net $\rightarrow$ smaller differences
- What happens in the limit?



Ordinary differential equation for Resnets

- In the limit, state **h** updates:

$$\frac{\partial \mathbf{h}(t)}{\partial t} = f(\mathbf{h}(t), t, \theta)$$

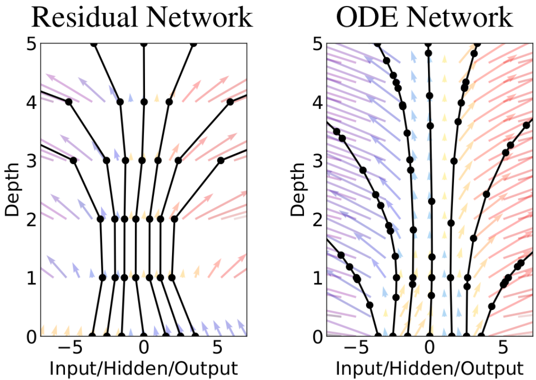
Continuous depth neural networks

- $\mathbf{h}(t)$ - state
  - $\mathbf{h}(0)$ (input vector)
  - $\mathbf{h}(T)$ (output vector, also $\phi(\mathbf{x})$, where $\mathbf{x} = \mathbf{h}(0)$)
  - $\mathbf{h}(t)$, $t \in (0, T)$ (internal state)
- State is transformed continuously from $\mathbf{h}(0)$ to $\mathbf{h}(T)$
- Parameterize the gradient of the state with a neural net f:

$$\frac{\partial \mathbf{h}(t)}{\partial t} = f(\mathbf{h}(t), t, \theta)$$

Chen, Ruvanova, Bettencourt, Duvenaud, 2018 (NeurIPS Best paper award)

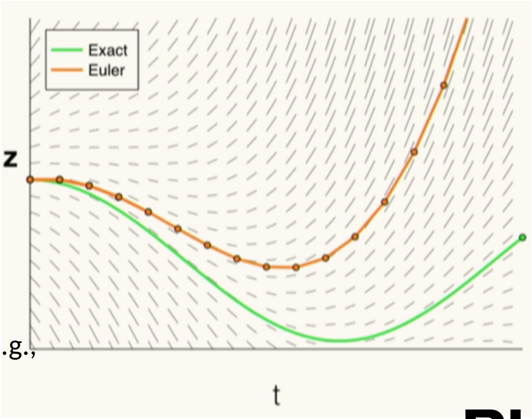
Continuous depth neural networks (2)



Chen, Ruvanova, Bettencourt, Duvenaud, 2018

Ordinary differential equation (ODE) solvers

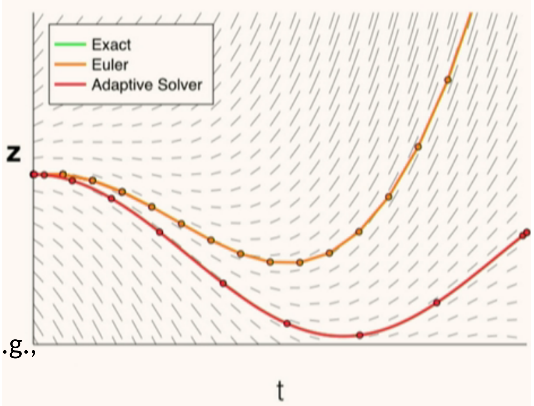
- Vector-valued **h** changes in time
- Time-derivative: $\frac{\partial \mathbf{h}}{\partial t} = f(\mathbf{h}(t), t)$
- Initial-value problem: given $\mathbf{h}_{t_0}$, find
  - $\mathbf{h}_{t_1} = \mathbf{h}_{t_0} + \int_{t_0}^{t_1} f(\mathbf{h}_t, t, \theta) dt$
- Oldest and simplest: Euler's method
- Takes a small step h in gradient's direction
  - $\mathbf{h}(t+h) = \mathbf{h} + hf(\mathbf{h}, t)$
- Modern solvers: 120 years of improvements e.g., (Hairer, et.al., 1987)
  - Approximation error guarantees
  - Adaptive evaluation strategy



Chen, et.al., 2018

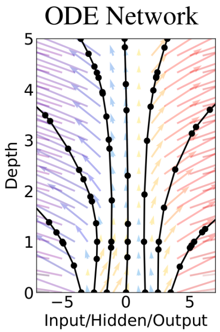
Ordinary differential equation (ODE) solvers

- Vector-valued **h** changes in time
- Time-derivative: $\frac{\partial \mathbf{h}}{\partial t} = f(\mathbf{h}(t), t)$
- Initial-value problem: given $\mathbf{h}_{t_0}$, find
  - $\mathbf{h}_{t_1} = \mathbf{h}_{t_0} + \int_{t_0}^{t_1} f(\mathbf{h}_t, t, \theta) dt$
- Oldest and simplest: Euler's method
- Takes a small step h in gradient's direction
  - $\mathbf{h}(t + h) = \mathbf{h} + hf(\mathbf{h}, t)$
- Modern solvers: 120 years of improvements e.g., (Hairer, et.al., 1987)
  - Approximation error guarantees
  - Adaptive evaluation strategy

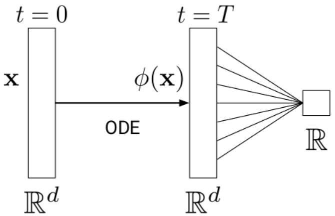


Chen, et.al., 2018

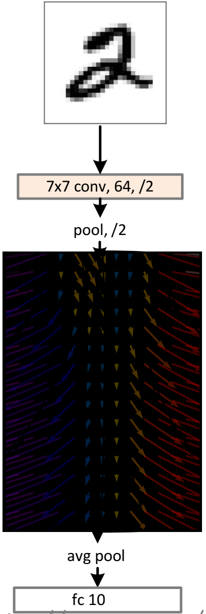
ODENet: The steps of the ODE solver defines the neural network.



ODENet/Neural ODE (NODE) architecture



Dupont, Doucet, Teh, 2019



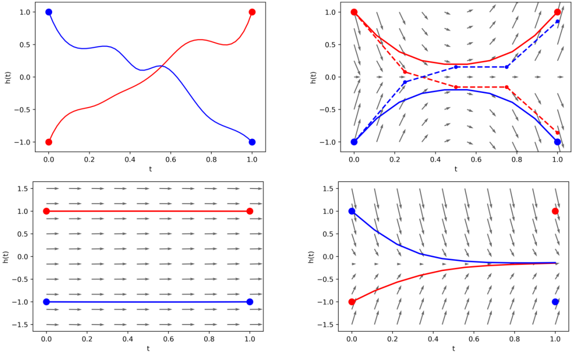
Drop-in replacement for Resnets

- Same performance with fewer parameters.

	Test Error	# Params
1-Layer MLP	1.60%	0.24 M
ResNet	0.41%	0.60 M
ODE-Net	0.42%	0.22 M

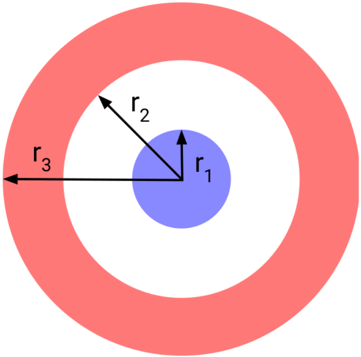
Chen, Rubanova, Bettencourt, Duvenaud, 2018

1d function impossible with NODE



Dupont, Doucet, Teh, 2019

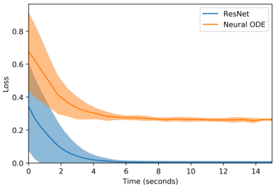
g(x): n-d function impossible with NODE



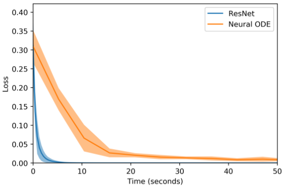
$$g(x) = -1 \text{ if } ||x|| \leq r_1, g(x) = 1 \text{ if } r_2 \leq ||x|| \leq r_3$$

Dupont, Doucet, Teh, 2019

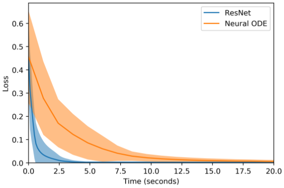
Training loss NODE vs ResNet



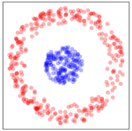
(a) $g(\mathbf{x})$ in $d = 1$

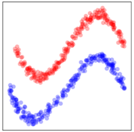


(b) $g(\mathbf{x})$ in $d = 2$



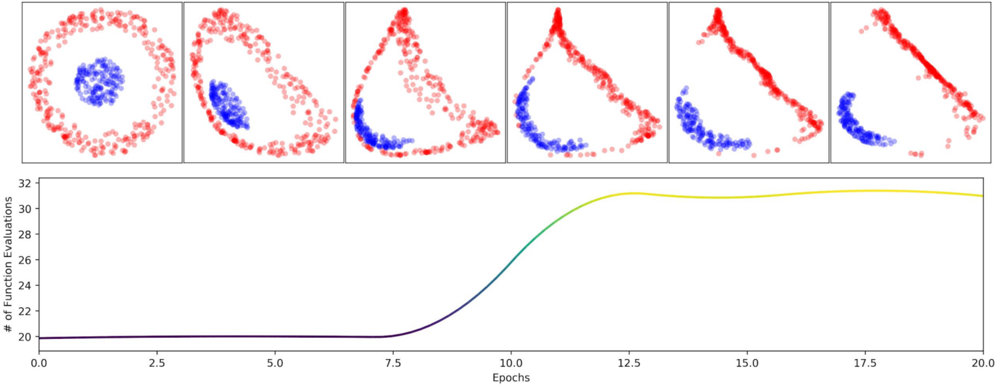
(c) Separable function in $d = 2$





Dupont, Doucet, Teh, 2019

NFE of g(x)



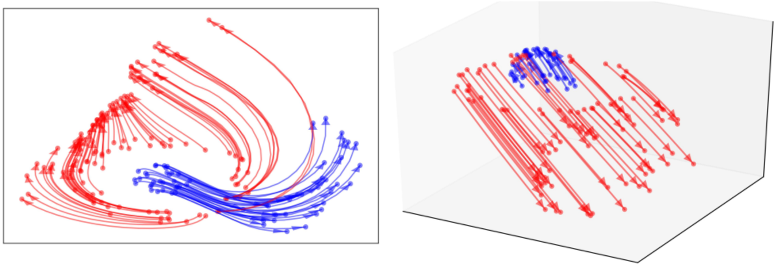
Dupont, Doucet, Teh, 2019

Augmented neural ODEs (ANODEs)

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{h}(t)\\\mathbf{a}(t)\end{bmatrix} = \mathbf{f}(\begin{bmatrix}\mathbf{h}(t)\\\mathbf{a}(t)\end{bmatrix}, t), \qquad \begin{bmatrix}\mathbf{h}(0)\\\mathbf{a}(0)\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\\mathbf{0}\end{bmatrix}$$

Dupont, Doucet, Teh, 2019

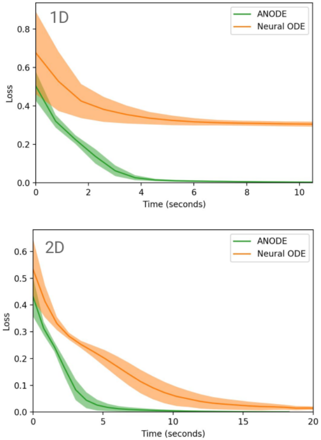
ANODE learns simpler flows

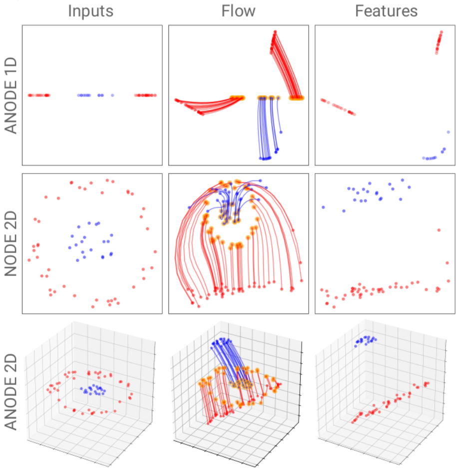


Neural ODE Augmented Neural ODE

Dupont, Doucet, Teh, 2019

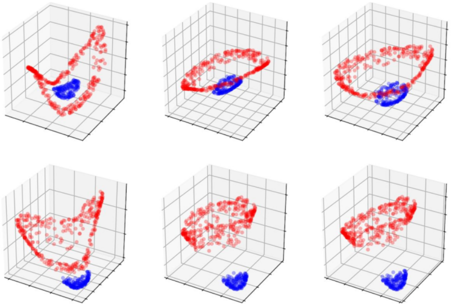
NODE vs ANODE

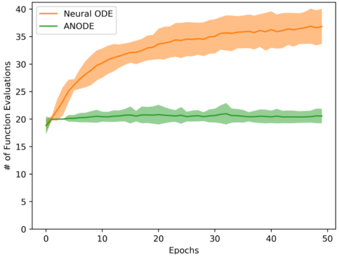




$g(x)$ not perfectly solved by NODE.
Dupont, Doucet, Teh, 2019

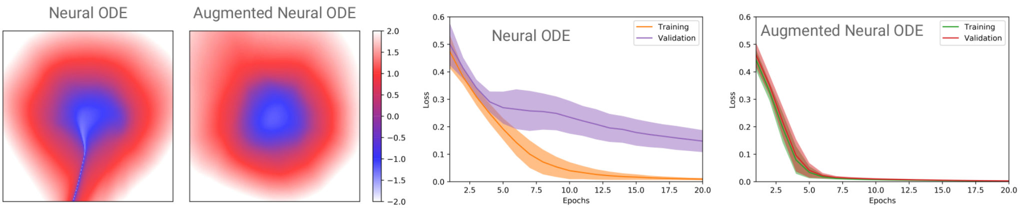
Feature evolution, ANODE





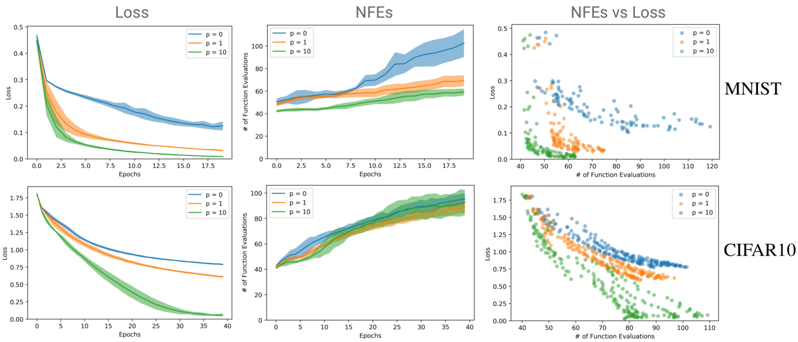
ANODE learns simpler flows, requires less NFEs.
Dupont, Doucet, Teh, 2019

Smoothness and generalization



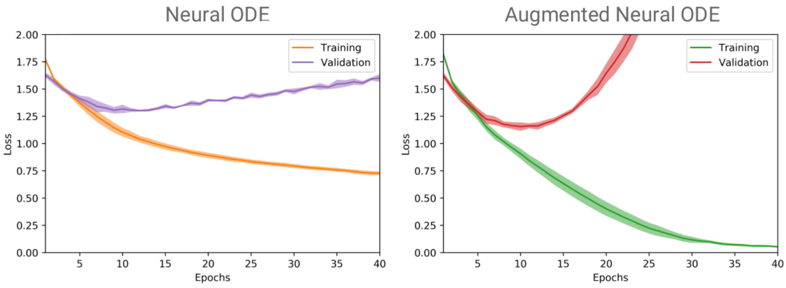
ANODE: smaller generalization gap.
Dupont, Doucet, Teh, 2019

MNIST, CIFAR10



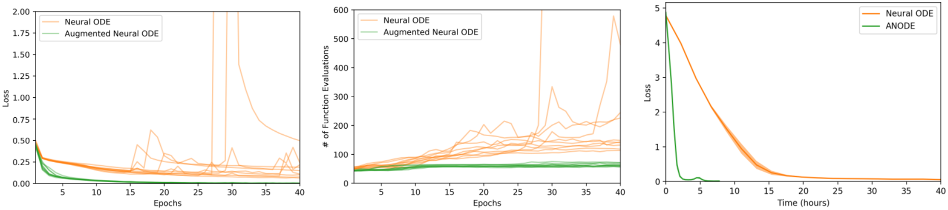
p - size of augmented dimension.
Dupont, Doucet, Teh, 2019

Generalization



CIFAR10
Dupont, Doucet, Teh, 2019

Stability



CIFAR10
Dupont, Doucet, Teh, 2019

Reading list

- Behrman, et.al., Invertible residual networks, ICML 2019, arxiv:1811.00995
- Garriga-Allonso, et.al., Deep convolutional networks as shallow gaussian processes, ICLR 2019, arxiv:1808.05587
- Ruthotto, Haber, Deep neural networks motivated by partial differential equations, 2018, arxiv:1804.04272
- Grathwohl, et.al., FFJORD: Free-form continuous dynamics for scalable reversible generative models, ICLR 2019

- Reversibility; train generative model using maximum likelihood

Appendix

How to train the ODENet

- Adjoint sensitivity method (Pontryagin et al., 1962)
- Continuous time limit of standard back-propagation
- Solve another ODE in reverse direction
- Error guarantees
- Dynamic step sizes

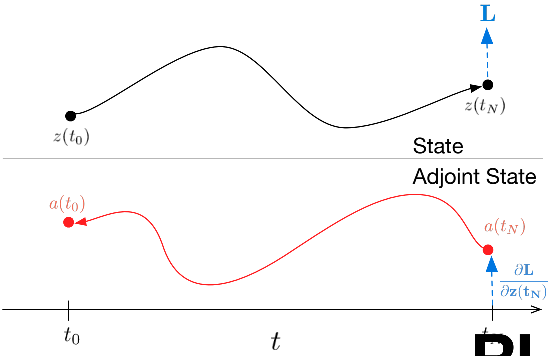
$$\mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{h}(t)}$$

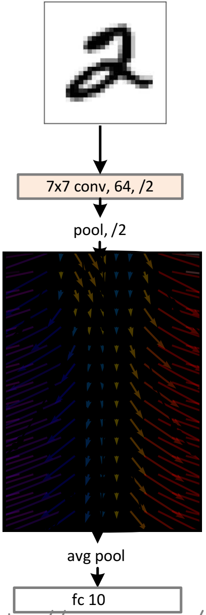
$$\frac{\partial \mathbf{a}(t)}{\partial t} = \mathbf{a}(t) \frac{\partial f(\mathbf{h}(t), t, \theta)}{\partial \mathbf{h}(t)}$$

$$\frac{\partial L}{\partial \theta} = \int_{t_1}^{t_o} \mathbf{a}(t) \frac{\partial f(\mathbf{h}(t), t, \theta)}{\partial \theta}$$

O(1) Memory Gradients

- No need to store activations, just run dynamics backwards from output.
- Reversible ResNets (Gomez et al., 2018) must partition dimensions.





Drop-in replacement for Resnets

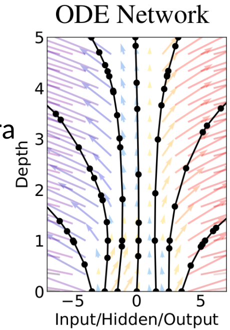
- Same performance with fewer parameters.

	Test Error	# Params
1-Layer MLP	1.60%	0.24 M
ResNet	0.41%	0.60 M
ODE-Net	0.42%	0.22 M

Chen, Rubanova, Bettencourt, Duvenaud

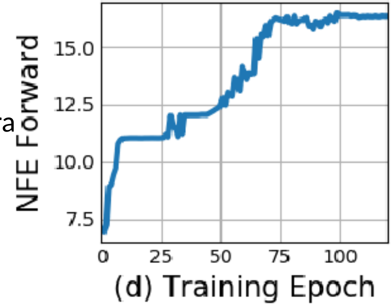
How deep are ODE-nets?

- ‘Depth’ is left to ODE solver.
- Dynamics become more demanding during tra
- 2-4x the depth of resnet architectures



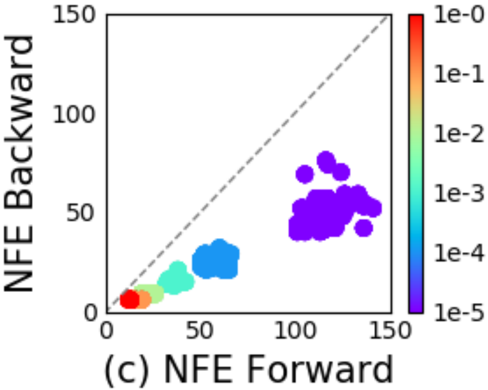
How deep are ODE-nets?

- ‘Depth’ is left to ODE solver.
- Dynamics become more demanding during tra
- 2-4x the depth of resnet architectures



Reverse vs Forward Cost

- Empirically, reverse pass roughly half as expensive as forward pass
- Again, adapts to instance difficulty
- Num evaluations comparable to number of layers in modern nets

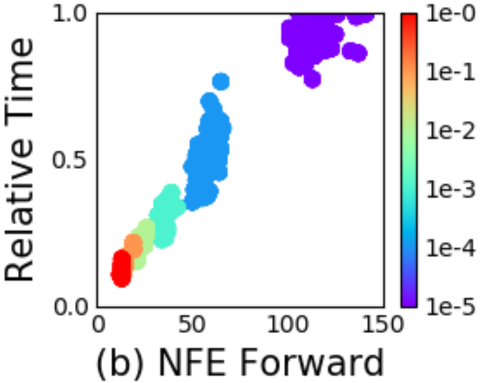


Chen, Rubanova, Bettencourt, Duvenaud

Speed-Accuracy Tradeoff

output = ODESolve(f, z0, t0, t1, theta, tolerance)

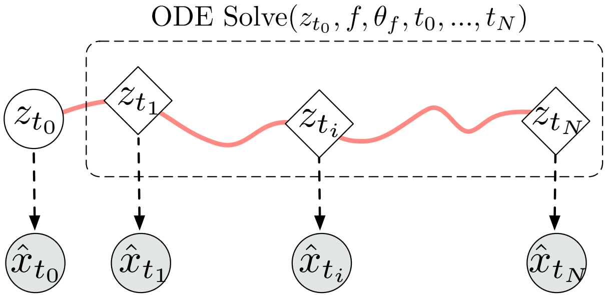
- Time cost is dominated by evaluation of dynamics
- Roughly linear with number of forward evaluations



tolerance

Chen, Rubanova, Bettencourt, Duvenaud

Continuous-time models



- Well-defined state at all times
- Dynamics separate from inference
- Irregularly-timed observations.

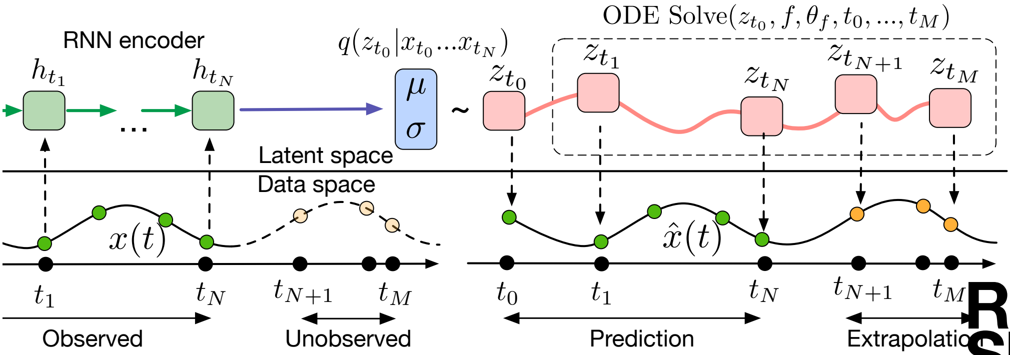
$$\mathbf{z}_{t_0} \sim p(\mathbf{z}_{t_0})$$
$$\mathbf{z}_{t_1}, \mathbf{z}_{t_2}, \ldots, \mathbf{z}_{t_N} = \text{ODESolve}(\mathbf{z}_{t_0}, f, \theta_f, t_0, \ldots, t_N)$$
$$\text{each} \quad \mathbf{x}_{t_i} \sim p(\mathbf{x}|\mathbf{z}_{t_i}, \theta_{\mathbf{x}})$$

Chen, Rubanova, Bettencourt, Duvenaud

RI. SE

Continuous-time RNNs

- Can do VAE-style inference with an RNN encoder
- Actually, more like a Deep Kalman Filter

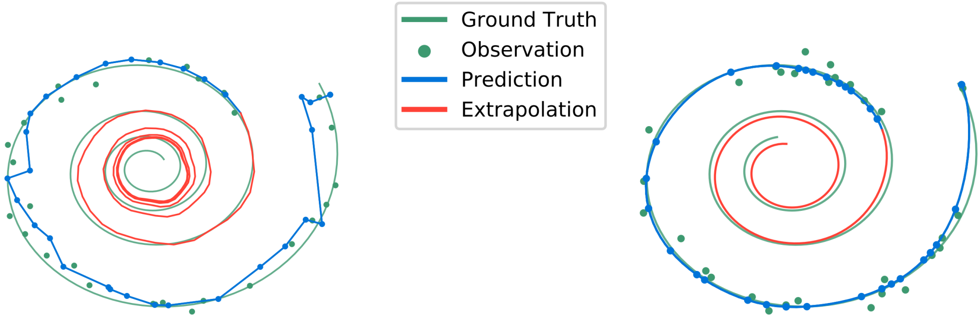


Chen, Rubanova, Bettencourt, Duvenaud

Continuous-time models

Recurrent Neural Net

Latent ODE



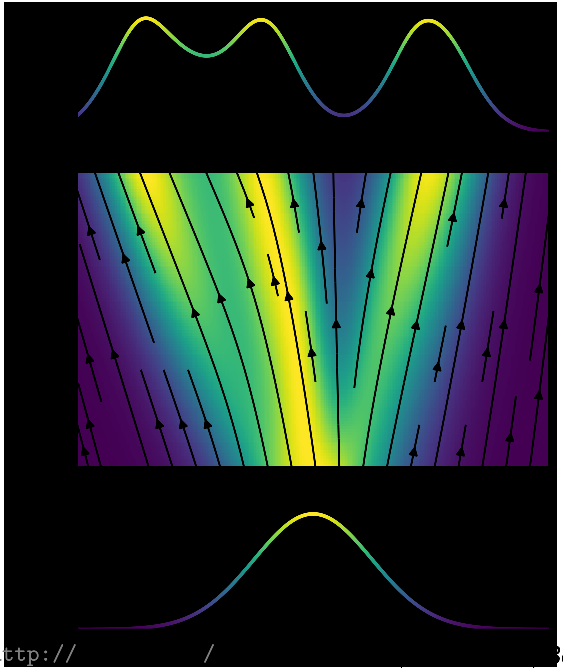
Chen, Rubanova, Bettencourt, Duvenaud

Normalizing flows

Tabak & Vanden-Eijnden 2010

- The transformation of a probability density through a sequence of invertible mappings
- Change of variables rule
- Produces a valid probability distribution
- Requires computing the determinant: $O(M^3)$

$$q(\mathbf{h}') = q(\mathbf{h}) \left| det \frac{\partial f^{-1}}{\partial \mathbf{h}'} \right| = q(\mathbf{h}) \left| det \frac{\partial f}{\partial \mathbf{h}} \right|^{-1}$$



Instantaneous Change of Variables

$$\frac{d\mathbf{z}}{dt} = f(\mathbf{z}(t), t)$$

$$\Downarrow$$

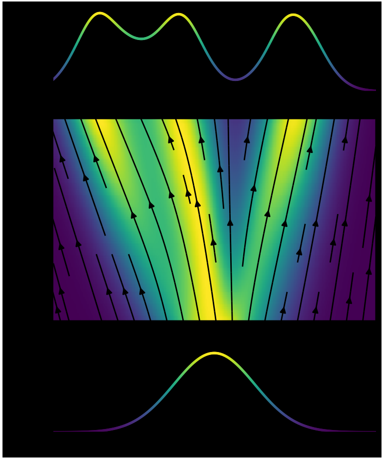
$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\mathbf{tr}\left(\frac{df}{d\mathbf{z}(t)}\right)$$

- Worst-case cost O(D^2).
- Only need continuously differentiable f

http:// / Bettencourt, Duvenaud

Continuous Normalizing Flows

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \mathrm{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)\, dt$$



- Reversible dynamics, so can train from data by maximum likelihood
- No discriminator or recognition network, train by SGD
- No need to partition dimensions

Chen, Rubanova, Bettencourt, Duvenaud

Trading Depth for Width

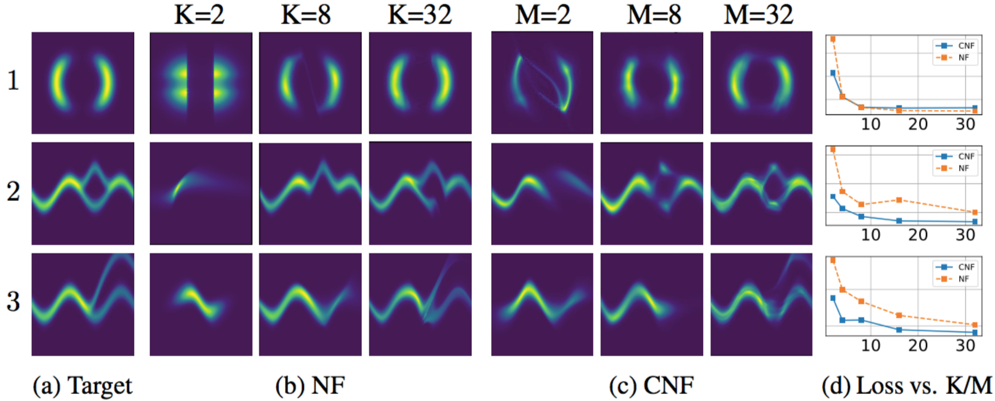


Figure 5: Comparison of NF and CNFs on learning generative models (noise → data) trained to minimize the reverse KL.

Chen, Rubanova, Bettencourt, Duvenaud

Concluding remarks

- Memory efficiency (constant)
- The ODE solver takes a tolerance parameter, trade-off accuracy vs running time
- Time-series with irregular observation times
- Continuous normalizing flows
- Computation time not not guaranteed
- 2-4 times slower than Resnets