April 2nd, 2020

Mathematics of Neural ODEs

Vikram Voleti

PhD student - Mila, University of Montreal — **Prof. Christopher Pal**

Visiting Researcher - University of Guelph — **Prof. Graham Taylor**

voletiv.github.io

@ (virtual) University of Guelph

Contents

1. **Ordinary Differential Equations (ODEs)**
    - **Initial Value Problems**
    - **Numerical Integration methods**
    - **Fundamental theorem of ODEs**
2. Neural ODEs
3. Recent research

Ordinary Differential Equations (ODEs)

1st order Ordinary Differential Equation:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta)$$

x is a variable we are interested in,

t is (typically) time,

f is a function of x and t, it is the differential,

θ parameterizes f (optionally).

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Many physical processes follow this template!

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

Example:

$$\frac{dx}{dt} = 2t; \; x(0) = 2; \; x(1) = \ ?$$

$$\begin{aligned} \Rightarrow x(1) &= x(0) + \int_0^1 2t \, dt \\ &= x(0) + \left(t^2\big|_{t=1} - t^2\big|_{t=0}\right) \\ &= 2 + 1^2 - 0^2 \\ &= 3 \end{aligned}$$

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \boxed{\int_{t_0}^{t_1} f(x(t), t, \theta)\, dt}$$

What if this cannot be analytically integrated?

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \boxed{\int_{t_0}^{t_1} f(x(t), t, \theta)\, dt}$$

What if this cannot be analytically integrated?

Example:

$$\frac{dx}{dt} = 2xt \ ; \ x(0) = 3$$
$$\Rightarrow \int \frac{1}{2x}\, dx = \int t\, dt$$
$$\Rightarrow \frac{1}{2}\log x = \frac{1}{2}t^2 + c_0$$
$$\Rightarrow x(t) = ce^{t^2}$$
$$x(0) = 3 \Rightarrow c = 2$$
$$\therefore x(t) = 2e^{t^2}$$
$$\Rightarrow x(1) = 5.436$$

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

Approximations to $\int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$

i.e. **Numerical Integration :**

- Euler method
- Runge-Kutta methods
- ...

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \;?$$

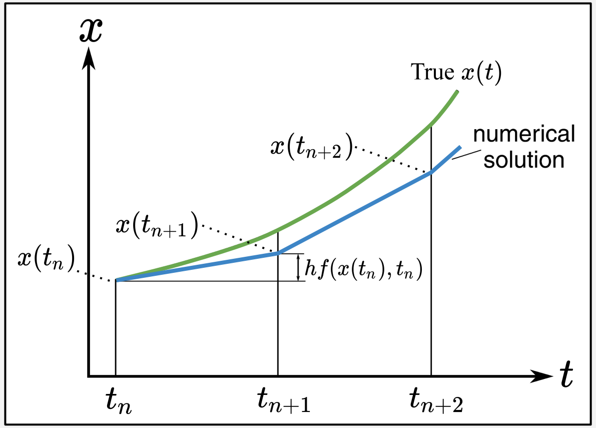
Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

1st-order Runge-Kutta / Euler's method:

$t_{n+1} = t_n + h$ → Step size h

$x(t_{n+1}) = x(t_n) + hf(x(t_n), t_n)$ → Update using derivative f



https://guide.freecodecamp.org/mathematics/differential-equations/eulers-method/

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \;?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

1st-order Runge-Kutta / Euler's method:

$$t_{n+1} = t_n + h$$
$$x(t_{n+1}) = x(t_n) + hf(x(t_n), t_n)$$

Example:

$$\frac{dx}{dt} = f(x,t) = 2xt \;;\; x(0) = 3;\; x(1) = \;?$$

(Solution: $x(t) = 2e^{t^2}$; $x(1) = 5.436$)

$\boxed{h = 0.25}$

$$\begin{aligned}
x(0.25) &= x(0) + 0.25 * f(x(0), 0) \\
&= 3 + 0.25 * (2 * 3 * 0) \\
&= 3 \\
x(0.5) &= x(0.25) + 0.25 * f(x(0.25), 0.25) \\
&= 3 + 0.25 * (2 * 3 * 0.25) \\
&= 3.375 \\
x(0.75) &= x(0.5) + 0.25 * f(x(0.5), 0.5) \\
&= 3.375 + 0.25 * (2 * 3.375 * 0.5) \\
&= 4.21875 \\
x(1) &= x(0.75) + 0.25 * f(x(0.75), 0.75) \\
&= 4.21875 + 0.25 * (2 * 4.21875 * 0.75) \\
&= 5.8008
\end{aligned}$$

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

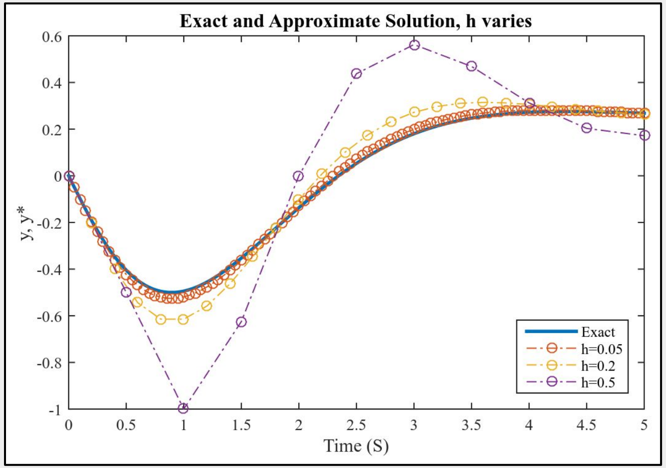
Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

1st-order Runge-Kutta / Euler's method:

$$t_{n+1} = t_n + h$$
$$x(t_{n+1}) = x(t_n) + hf(x(t_n), t_n)$$

Step size matters!



https://lpsa.swarthmore.edu/NumInt/NumIntFirst.html

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

1st-order Runge-Kutta / Euler's method:

$$t_{n+1} = t_n + h$$

$$s_1 = f(x(t_n),\ t_n)$$

$$x(t_{n+1}) = x(t_n) + hs_1$$

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

2nd-order Runge-Kutta method:

$$t_{n+1} = t_n + h$$

$$s_1 = f(x(t_n),\ t_n)$$

$$s_2 = f(x(t_n + \tfrac{h}{2}),\ t_n + \tfrac{h}{2}) = f(x(t_n) + \tfrac{h}{2} s_1,\ t_n + \tfrac{h}{2})$$

$$x(t_{n+1}) = x(t_n) + h s_2$$

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \; ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

★ **4th-order Runge-Kutta method:**

$$t_{n+1} = t_n + h$$

$$s_1 = f(x(t_n),\; t_n)$$

$$s_2 = f(x(t_n) + \tfrac{h}{2} s_1,\; t_n + \tfrac{h}{2})$$

$$s_3 = f(x(t_n) + \tfrac{h}{2} s_2,\; t_n + \tfrac{h}{2})$$

$$s_4 = f(x(t_n) + h s_3,\; t_n + h)$$

$$x(t_{n+1}) = x(t_n) + \tfrac{h}{6}(s_1 + 2s_2 + 2s_3 + s_4)$$

Default ODE solver used in MATLAB:
https://blogs.mathworks.com/loren/2015/09/23/ode-solver-selection-in-matlab/

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

Many other ODE solvers to choose from!

Mila

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \; ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

Many other ODE solvers to choose from!

Considerations to choose an ODE solver:

- Stiff v/s Non-stiff ODE
- # of calculations per iteration
- Implicit v/s Explicit solver
- Single-step size v/s Multi-step size (adaptive)

https://blogs.mathworks.com/loren/2015/09/23/ode-solver-selection-in-matlab/

https://math.temple.edu/~queisser/assets/files/Talk3.pdf

Ordinary Differential Equations (ODEs)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = x(t_0) + \int_{t_0}^{t_1} f(x(t), t, \theta)\, dt$$

$$\boxed{x(t_1) = \text{ODESolve}(\ f(x(t), t, \theta),\ x(t_0),\ t_0,\ t_1\ )}$$

Final time

Initial time

Initial value

Differential

Any ODE solver of our choice!

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \;?$$

Solution:

$$x(t_1) = \text{ODESolve}(\; f(x(t), t, \theta),\; x(t_0),\; t_0,\; t_1 \;)$$

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \;\; x(t_0) \text{ is given}; \;\; x(t_1) = \;?$$

Solution:

$$x(t_1) = \text{ODESolve}(\; f(x(t), t, \theta),\; x(t_0),\; t_0,\; t_1 \;)$$

Fundamental Theorem of ODEs

Suppose f is continuously differentiable.

Then, the solution to the initial value problem is **unique**!

http://faculty.bard.edu/belk/math213/InitialValueProblems.pdf

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given;} \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = \text{ODESolve}(\ f(x(t), t, \theta),\ x(t_0),\ t_0,\ t_1\ )$$

Fundamental Theorem of ODEs

Suppose f is continuously differentiable.

1. The solution curves for this differential equation completely fill the plane, and
2. Solution curves for different solutions **do not intersect**.

http://faculty.bard.edu/belk/math213/InitialValueProblems.pdf

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = \text{ODESolve}(\ f(x(t), t, \theta),\ x(t_0),\ t_0,\ t_1\ )$$

Fundamental Theorem of ODEs

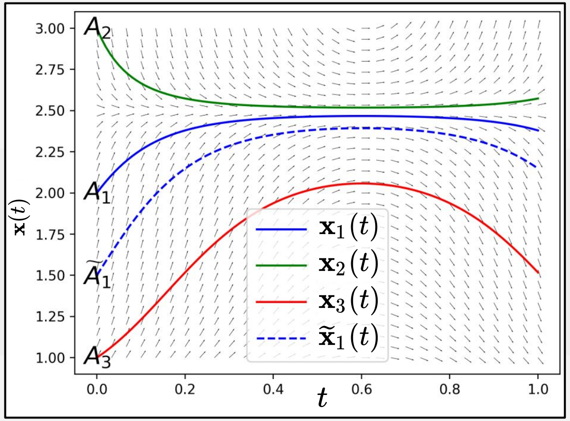
Suppose f is continuously differentiable.

1. The ... differential equation ... tely fil

2. Solution curves for different solutions **do not intersect.**

Geometrically, $x(t)$ is a flow!

http://faculty.bard.edu/belk/math213/InitialValueProblems.pdf



https://openreview.net/pdf?id=B1e9Y2NYvS

1. Ordinary Differential Equations (ODEs)
    - Initial Value Problems
    - Numerical Integration methods
    - Fundamental theorem of ODEs
2. **Neural ODEs** (Chen et al., 2018)
    - **Adjoint method**
    - **Applications**
3. Recent research

Neural ODEs (Chen et al., 2018)

Initial value problem:

$$\frac{dx(t)}{dt} = f(x(t), t, \theta); \quad x(t_0) \text{ is given}; \quad x(t_1) = \ ?$$

Solution:

$$x(t_1) = \text{ODESolve}(\ f(x(t), t, \theta),\ x(t_0),\ t_0,\ t_1\ )$$

f **is a neural network!**

Paradigm shift: whereas earlier *f* was pre-defined/hand-designed according to the domain, here we would like to estimate an *f* that suits our objective.

https://arxiv.org/pdf/1806.07366.pdf

Neural ODEs (Chen et al., 2018)

ODEs

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

Euler discretization

Vector notation

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f(\mathbf{x}_n, t_n, \theta)$$

https://arxiv.org/pdf/1806.07366.pdf

Residual networks

$$\mathbf{x}_{l+1} = \text{ResBlock}(\mathbf{x}_l, \theta)$$

$$\mathbf{x}_{l+1} = \mathbf{x}_l + g(\mathbf{x}_l, \theta)$$

Skip connection

https://arxiv.org/pdf/1512.03385.pdf

Neural ODEs (Chen et al., 2018)

ODEs

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

Euler discretization

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f(\mathbf{x}_n, t_n, \theta)$$

Forward propagation:

$$\mathbf{x}(t_1) = \mathrm{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$$

https://arxiv.org/pdf/1806.07366.pdf

Residual networks

$$\mathbf{x}_{l+1} = \mathrm{ResBlock}(\mathbf{x}_l, \theta)$$

$$\mathbf{x}_{l+1} = \mathbf{x}_l + g(\mathbf{x}_l, \theta)$$

Skip connection

$$\mathbf{y}_{pred} = \mathrm{ResNet}(\mathbf{x})$$

Stacked ResBlocks

https://arxiv.org/pdf/1512.03385.pdf

Neural ODEs (Chen et al., 2018)

ODEs

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

Euler discretization

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f(\mathbf{x}_n, t_n, \theta)$$

Forward propagation:

$$\mathbf{x}(t_1) = \mathrm{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$$

$$L(\mathbf{x}(t_1)) \rightarrow \frac{\partial L}{\partial \theta}$$

How to compute this?

Update θ to reduce L

https://arxiv.org/pdf/1806.07366.pdf

Residual networks

$$\mathbf{x}_{l+1} = \mathrm{ResBlock}(\mathbf{x}_l, \theta)$$

$$\mathbf{x}_{l+1} = \mathbf{x}_l + g(\mathbf{x}_l, \theta)$$

Skip connection

$$\mathbf{y}_{pred} = \mathrm{ResNet}(\mathbf{x})$$

Stacked ResBlocks

$$L(\mathbf{y}_{pred}) \rightarrow \frac{\partial L}{\partial \theta}$$

Update θ to reduce L

https://arxiv.org/pdf/1512.03385.pdf

Neural ODEs (Chen et al., 2018)

ODEs

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

Euler discretization

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f(\mathbf{x}_n, t_n, \theta)$$

Forward propagation:

$$\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$$

$$L(\mathbf{x}(t_1)) \rightarrow \boxed{\frac{\partial L}{\partial \theta}}$$

Update θ to reduce L

Back-propagate through the ODE Solver!

https://arxiv.org/pdf/1806.07366.pdf

ODEs

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

Euler discretization

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f(\mathbf{x}_n, t_n, \theta)$$

Forward propagation:

$$\mathbf{x}(t_1) = \mathrm{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$$

$$L(\mathbf{x}(t_1)) \rightarrow \boxed{\frac{\partial L}{\partial \theta}}$$

Update θ to reduce L

Back-propagate through the ODE Solver!

High memory cost -

need to save all activations of all iterations of ODESolve.

Can we do better?

Yes.

https://arxiv.org/pdf/1806.07366.pdf

$$L(\mathrm{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

$$\boxed{\frac{\partial L}{\partial \theta}} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta}\ dt$$

https://arxiv.org/pdf/1806.07366.pdf

$$\boxed{L(\mathrm{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

$$\frac{\partial L}{\partial \theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta}\, dt$$

We need a(t)

https://arxiv.org/pdf/1806.07366.pdf

$$\boxed{L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$

https://arxiv.org/pdf/1806.07366.pdf

$$L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Can be computed using autodiff

https://arxiv.org/pdf/1806.07366.pdf

Mila

$$\boxed{L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \quad \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Can be computed using autodiff

We can use a(t1) as initial value,
and integrate backwards from t1 to t to get a(t).

We'll use t0 as a proxy for t

https://arxiv.org/pdf/1806.07366.pdf

Neural ODEs (Chen et al., 2018)

$$\boxed{L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Initial value

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

Backward integration from t1 to t0

https://arxiv.org/pdf/1806.07366.pdf

$$L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \dfrac{\partial L}{\partial \mathbf{x}(t_1)}$

Initial value

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(\underbrace{-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}}_{\substack{\textit{Vector-Jacobian Product} \\ \textit{(can be efficiently evaluated by autodiff)}}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

https://arxiv.org/pdf/1806.07366.pdf

$$L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \dfrac{\partial L}{\partial \mathbf{x}(t_1)}$

But we need x(t)

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}\left(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0\right)$$

(and we don't want to have saved x(t) in memory from forward-prop)

https://arxiv.org/pdf/1806.07366.pdf

Mila

$$\boxed{L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^{\top} \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \ \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Initial value

$$x(t_0) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta)\ ,\ \mathbf{x}(t_1),\ t_1,\ t_0)$$

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(-\mathbf{a}(t)^{\top} \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

https://arxiv.org/pdf/1806.07366.pdf

Neural ODEs (Chen et al., 2018)

$$L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \ \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Back-propagation:

$$x(t_0) = \text{ODESolve}(\quad f(\mathbf{x}(t), t, \theta) \quad,\ \mathbf{x}(t_1),\ t_1,\ t_0)$$

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

$$\therefore \frac{\partial L}{\partial \theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta}\, dt$$

https://arxiv.org/pdf/1806.07366.pdf

$$L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \dfrac{\partial L}{\partial \mathbf{x}(t_1)}$

Back-propagation:

$$x(t_0) = \text{ODESolve}(\ \ f(\mathbf{x}(t), t, \theta)\ \ ,\ \mathbf{x}(t_1),\ t_1,\ t_0)$$

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

$$\therefore \frac{\partial L}{\partial \theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta}\, dt = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta},\ \ \mathbf{0}_{|\theta|}\ \ ,\ t_1,\ t_0)$$

Initial value is 0

https://arxiv.org/pdf/1806.07366.pdf

Neural ODEs (Chen et al., 2018)

$$\boxed{L(\text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )) \rightarrow \frac{\partial L}{\partial \theta}}$$

Adjoint method (Pontryagin et al., 1962)

$$\text{adjoint } \mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{x}}\ ;\ \frac{d\mathbf{a}}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}}$$

Forward propagation: $\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ ) \Rightarrow \mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$

Back-propagation:

$$x(t_0) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta)\ ,\ \mathbf{x}(t_1),\ t_1,\ t_0)$$

$$\Rightarrow \mathbf{a}(t_0) = \frac{\partial L}{\partial \mathbf{x}(t_0)} = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}},\ \frac{\partial L}{\partial \mathbf{x}(t_1)},\ t_1,\ t_0)$$

$$\therefore \frac{\partial L}{\partial \theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta}\, dt = \text{ODESolve}(-\mathbf{a}(t)^\top \frac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta},\ \mathbf{0}_{|\theta|},\ t_1,\ t_0)$$

Combine the 3 ODE Solves into 1!

https://arxiv.org/pdf/1806.07366.pdf

Neural ODEs (Chen et al., 2018)

Forward propagation:

$$\mathbf{x}(t_1) = \text{ODESolve}(\ f(\mathbf{x}(t), t, \theta),\ \mathbf{x}(t_0),\ t_0,\ t_1\ )$$

Compute $L(\mathbf{x}(t_1))$.

$$\mathbf{a}(t_1) = \frac{\partial L}{\partial \mathbf{x}(t_1)}$$

Back-propagation:

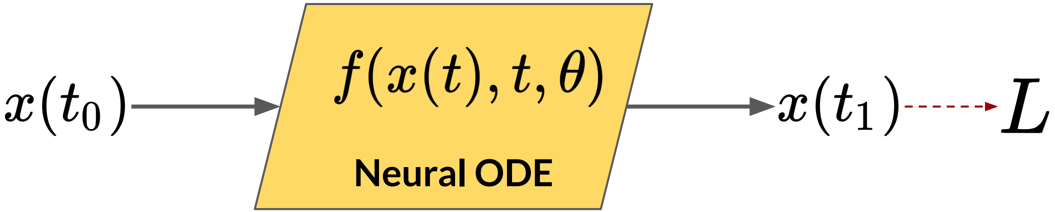
$$\begin{bmatrix} \mathbf{x}(t_0) \\ \dfrac{\partial L}{\partial \mathbf{x}(t_0)} \\ \dfrac{\partial L}{\partial \theta} \end{bmatrix} = \text{ODESolve}\left(\begin{bmatrix} f(\mathbf{x}(t), t, \theta) \\ -\mathbf{a}(t)^\top \dfrac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \mathbf{x}} \\ -\mathbf{a}(t)^\top \dfrac{\partial f(\mathbf{x}(t), t, \theta)}{\partial \theta} \end{bmatrix}, \begin{bmatrix} \mathbf{x}(t_1) \\ \dfrac{\partial L}{\partial \mathbf{x}(t_1)} \\ \mathbf{0}_{|\theta|} \end{bmatrix}, t_1, t_0 \right)$$

Update θ to reduce L

https://arxiv.org/pdf/1806.07366.pdf

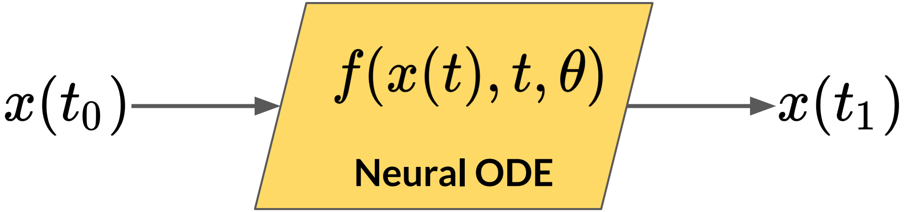
Neural ODEs (Chen et al., 2018)

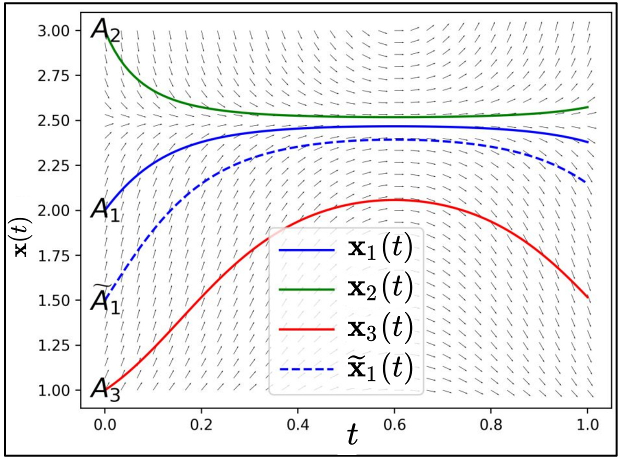
https://arxiv.org/pdf/1806.07366.pdf



Neural ODEs (Chen et al., 2018)

https://arxiv.org/pdf/1806.07366.pdf





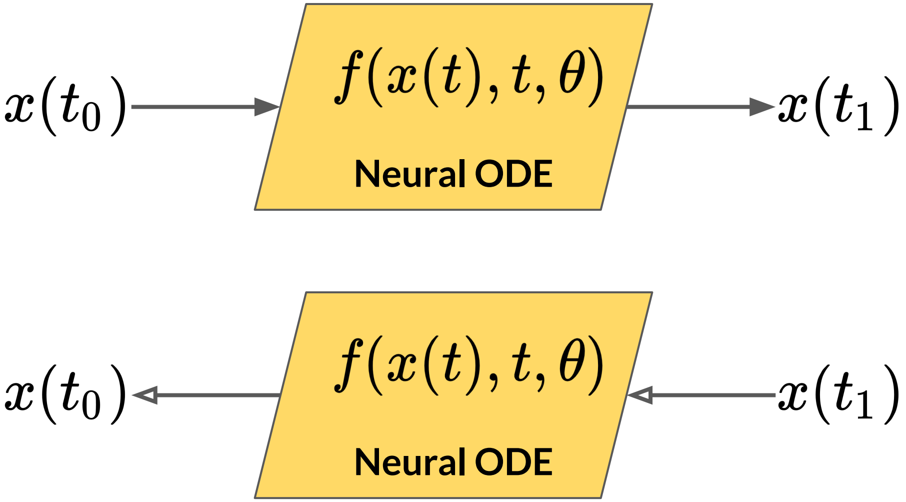
https://openreview.net/pdf?id=B1e9Y2NYvS

Neural ODEs describe a homeomorphism (flow).

- They preserve dimensionality.
- They form non-intersecting trajectories.

Neural ODEs (Chen et al., 2018)

https://arxiv.org/pdf/1806.07366.pdf



Neural ODEs are **reversible** models!
Just integrate forward/backward in time.

https://arxiv.org/pdf/1806.07366.pdf

Applications

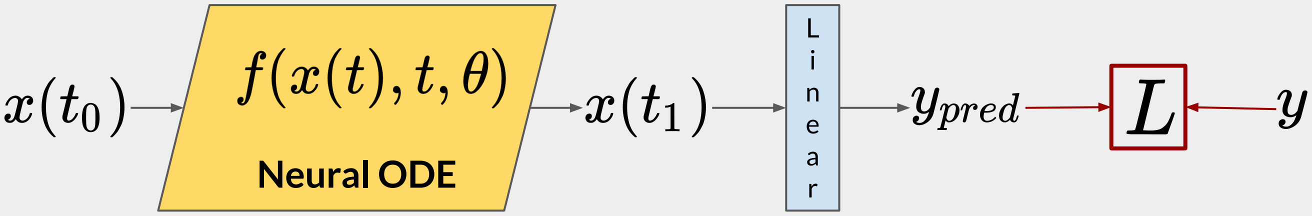
Supervised Learning | **Continuous Normalizing Flows** | **Generative Latent Models**

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | Generative Latent Models

ODE-Net:



~ Replacement for ResNets

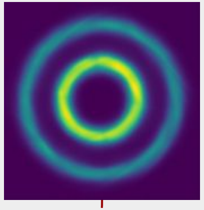
Neural ODEs (Chen et al., 2018)

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | **Continuous Normalizing Flows** | Generative Latent Models

Target distribution

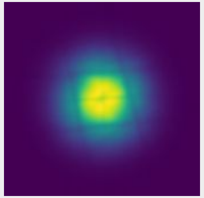


(such as real image manifold)

$\mathbf{x}(t_0) \longleftrightarrow f(\mathbf{x}(t), t, \theta) \longleftrightarrow \mathbf{x}(t_1)$

Neural ODE

Noise distribution



Sample from target distribution (such as an image)

Sample from noise distribution (such as Gaussian)

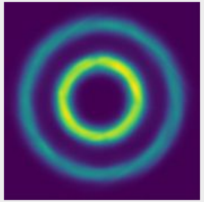
Neural ODEs (Chen et al., 2018)

https://arxiv.org/pdf/1806.07366.pdf

Applications

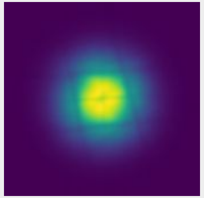
Supervised Learning | **Continuous Normalizing Flows** | Generative Latent Models

Target distribution



$\mathbf{x}(t_0) \longleftrightarrow$ $f(\mathbf{x}(t), t, \theta)$ Neural ODE $\longleftrightarrow \mathbf{x}(t_1)$

Noise distribution



Likelihood estimation

using Change of Variables formula

$$\mathbf{x}_1 = g(\mathbf{x}_0) \Rightarrow \log p(\mathbf{x}_1) = \log p(\mathbf{x}_0) - \log |\det \frac{\partial g}{\partial \mathbf{x}_0}|$$

Train *f* to maximize the likelihood of the samples from target distribution *log p(x_1)*,
by computing *x(t_0)* using the Neural ODE with *x(t_1)* as the initial value, and the Change of Variables formula.

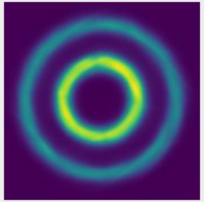
Neural ODEs (Chen et al., 2018)

https://arxiv.org/pdf/1806.07366.pdf

Applications

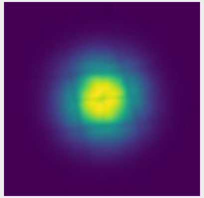
Supervised Learning | **Continuous Normalizing Flows** | Generative Latent Models

Target distribution



$\mathbf{x}(t_0) \leftrightarrow$ $f(\mathbf{x}(t), t, \theta)$ Neural ODE $\leftrightarrow \mathbf{x}(t_1)$

Noise distribution



$\longrightarrow$

Likelihood estimation
using Change of Variables formula

$\longleftarrow$

Generate samples

Sample from the noise distribution, transform it into a sample from the target distribution using the trained Neural ODE.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

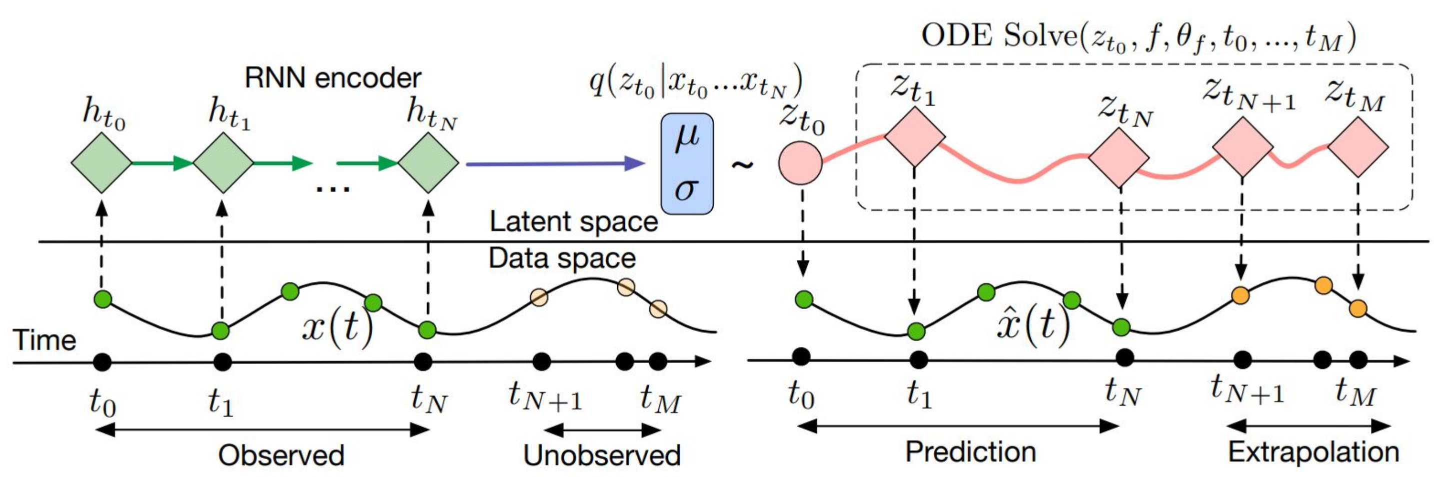


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

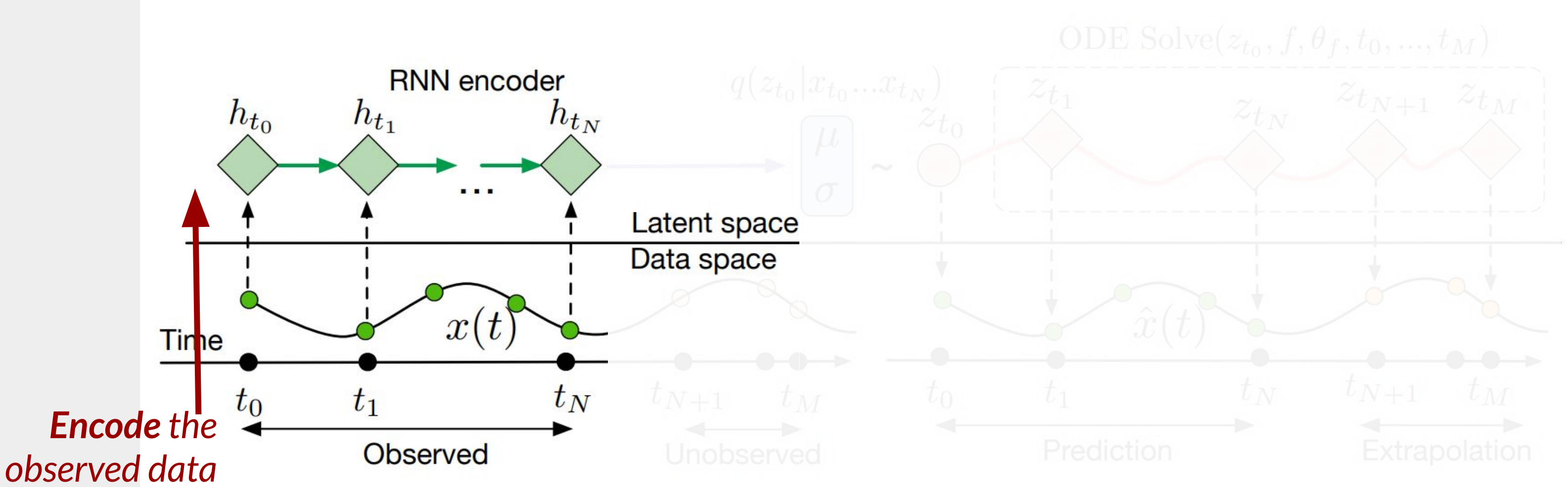


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

***Encode** into a latent distribution (such as Gaussian)*

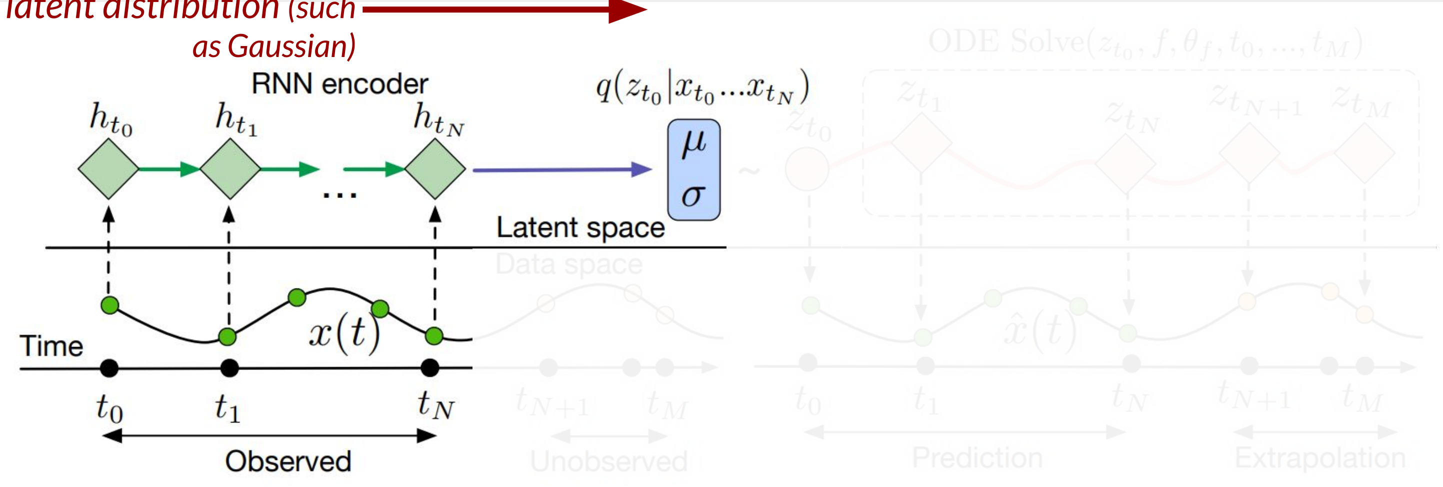


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

Sample *from the latent distribution*

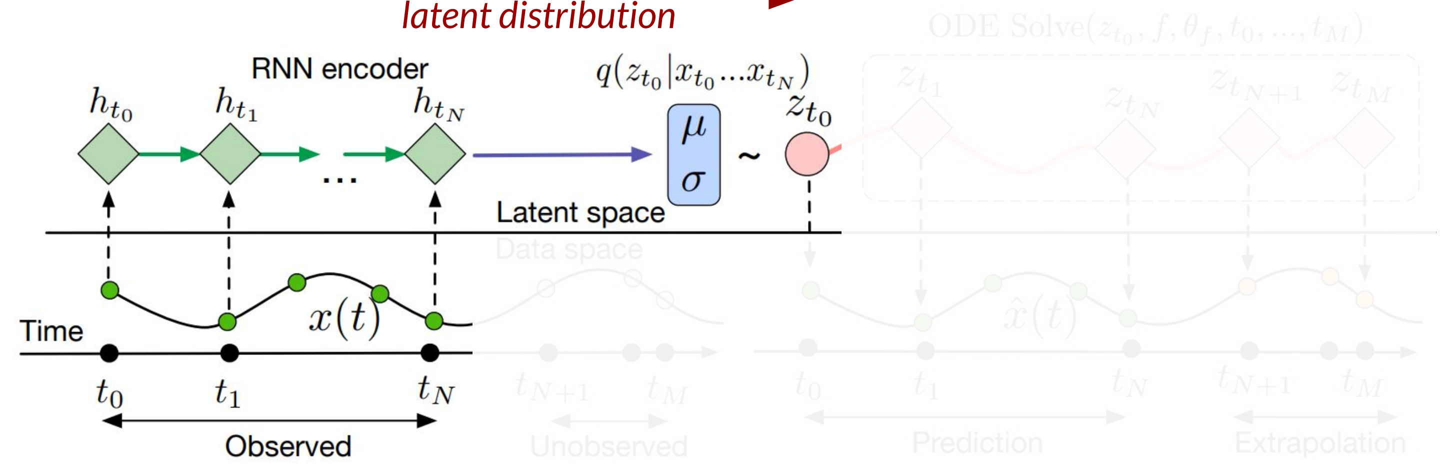


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

Use the sample as ***initial value*** *to a Neural ODE*

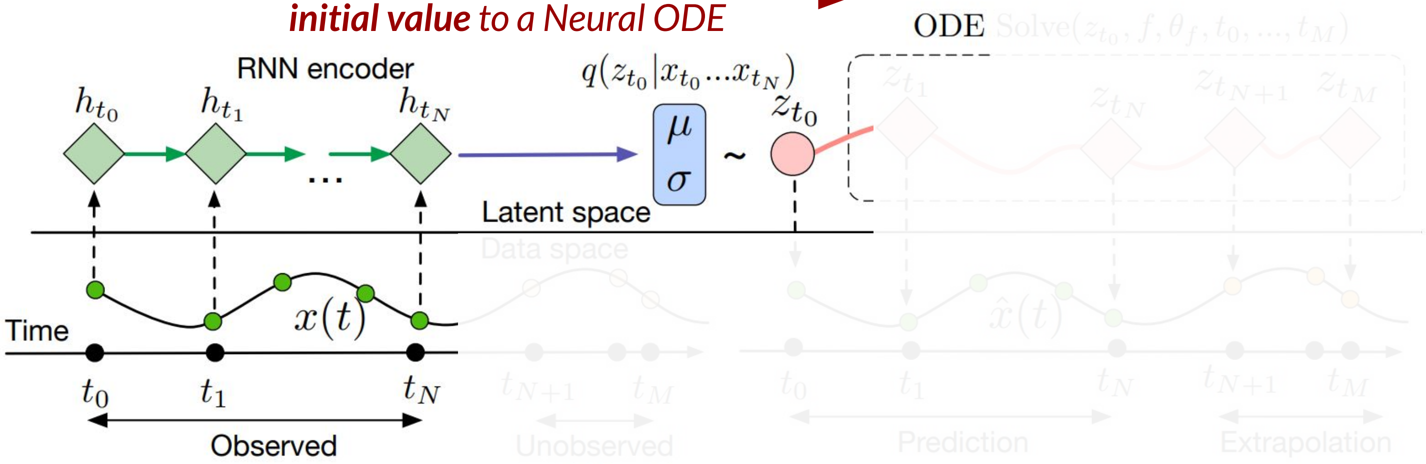


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

Solve *the Neural ODE for latent points in later time steps*

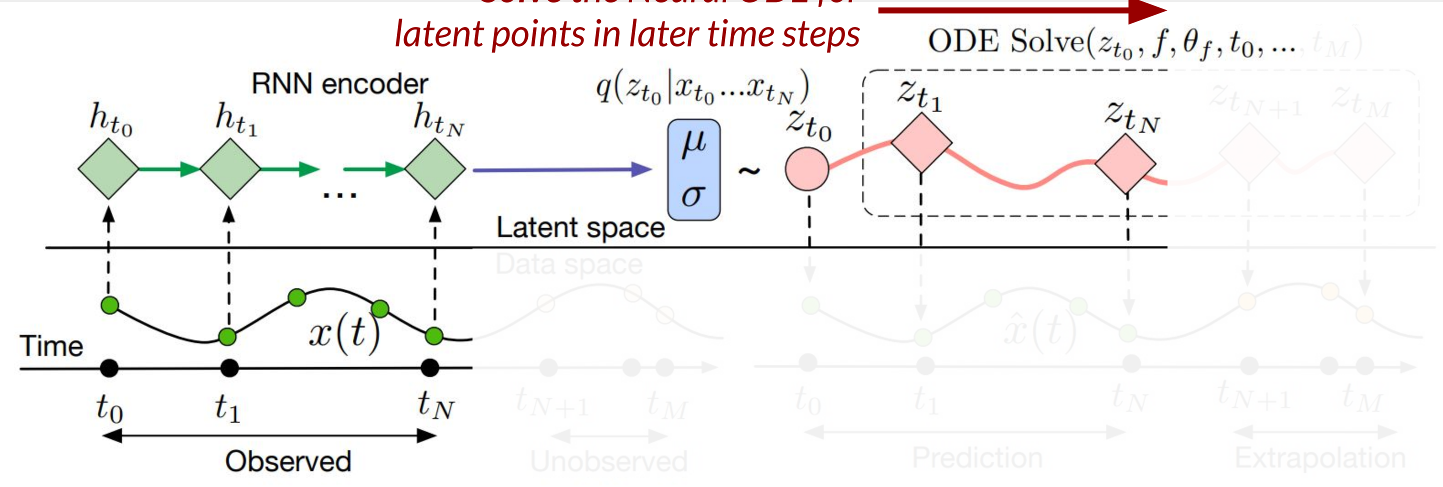


Figure 6: Computation graph of the latent ODE model.

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

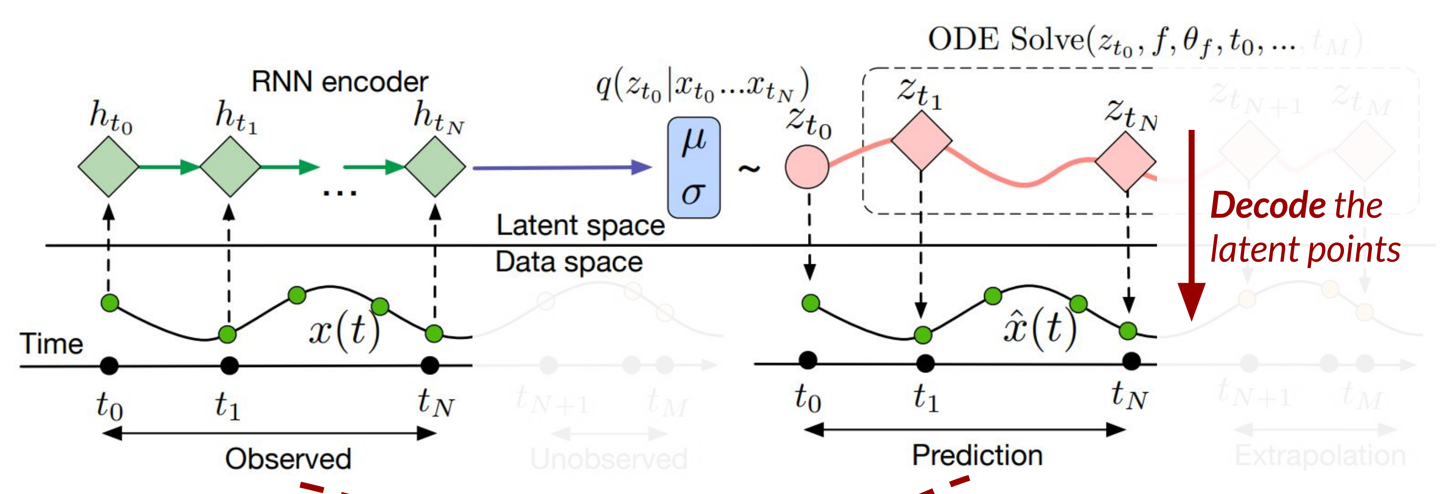


Figure 6: Computation graph of the latent ODE model.

Decode the latent points

*Compute **loss***

https://arxiv.org/pdf/1806.07366.pdf

Applications

Supervised Learning | Continuous Normalizing Flows | **Generative Latent Models**

*Solve the Neural ODE for latent points in **future** time steps*

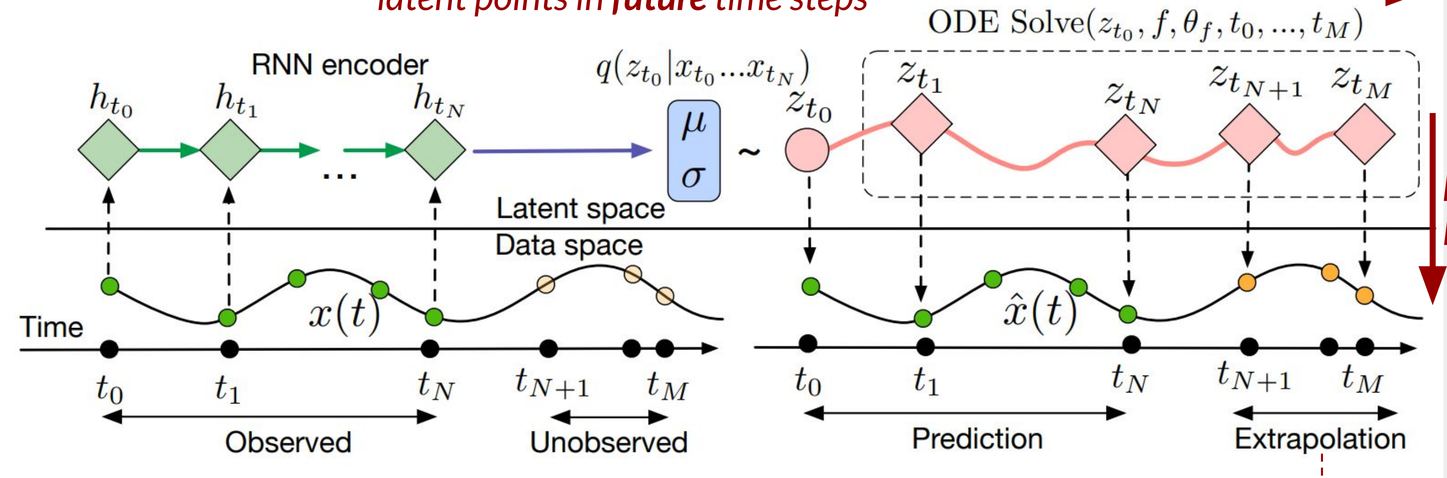


Figure 6: Computation graph of the latent ODE model.

***Decode** the latent points*

Future generation!

1. Ordinary Differential Equations (ODEs)
   - Initial Value Problems
   - Numerical Integration methods
   - Fundamental theorem of ODEs
2. Neural ODEs (Chen et al., 2018)
   - Adjoint method
   - Applications
3. **Recent research**

FFJORD: Free-form Continuous Dynamics For Scalable Reversible Generative Models (Grathwohl et al., ICLR 2019)

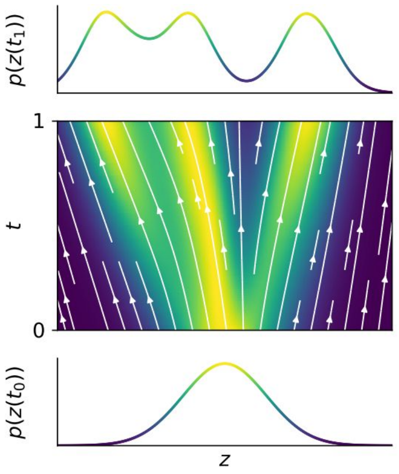
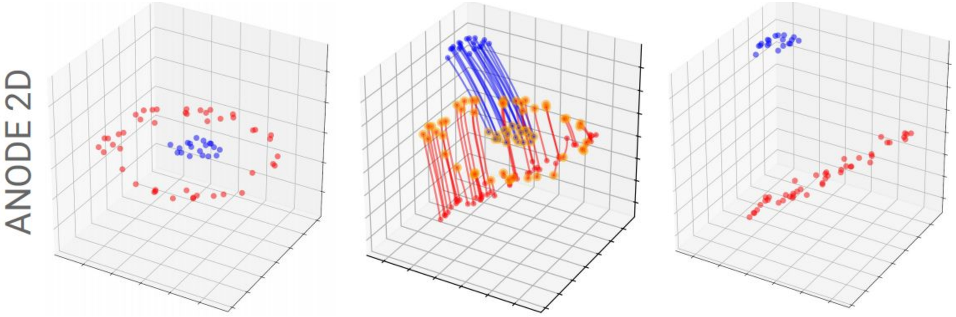


Figure 1: FFJORD transforms a simple base distribution at t_0 into the target distribution at t_1 by integrating over learned continuous dynamics.

- Essentially a better Continuous Normalizing Flow.
- Makes a better estimate for the log determinant term.
- "We demonstrate our approach on high-dimensional density estimation, image generation, and variational inference, achieving the state-of-the-art among exact likelihood methods with efficient sampling."

https://arxiv.org/pdf/1810.01367.pdf

Augmented Neural ODEs (Dupont et al., NeurIPS 2019)



- Shows that Neural ODEs cannot model non-homeomorphisms (non-flows)
- **Augments** the state with additional dimensions to cover non-homeomorphisms
- Performs ablation study on toy examples and image classification

https://arxiv.org/pdf/1904.01681.pdf

ANODEV2: A Coupled Neural ODE Evolution Framework

(Zhang et al., NeurIPS 2019)

$$\begin{cases} z(1) = z(0) + \int_0^1 f(z(t), \theta(t))dt & \text{“parent network”}, \\ \theta(t) = \theta(0) + \int_0^t q(\theta(t), p)dt, \quad \theta(0) = \theta_0 & \text{“weight network”}. \end{cases}$$

- Network weights are also a function of time
- Separate “weight network” generates the weights of the function network at a given time

https://arxiv.org/pdf/1906.04596.pdf

Mila

Latent ODEs for Irregularly-Sampled Time Series

(Rubanova et al., NeurIPS 2019)

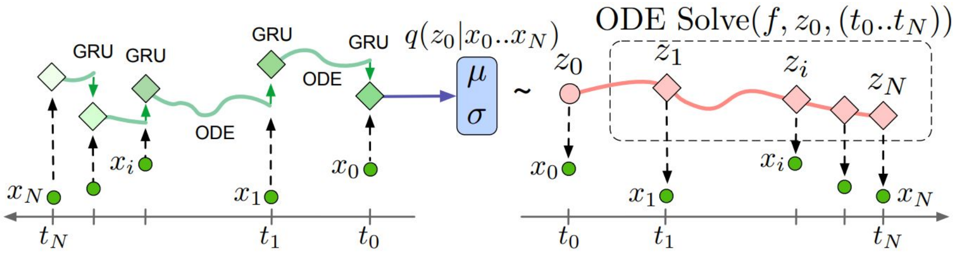


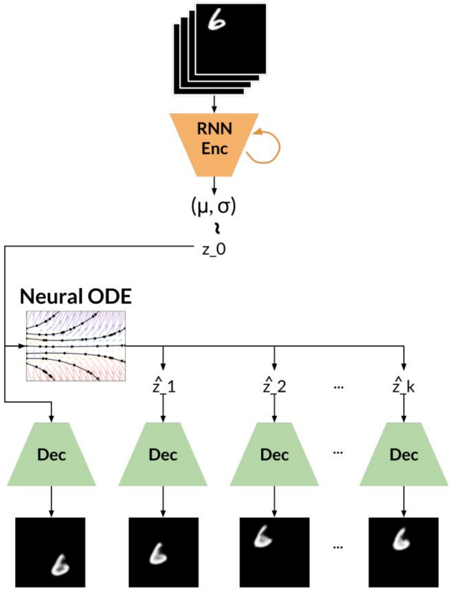
Figure 2: The Latent ODE model with an ODE-RNN encoder. To make predictions in this model, the ODE-RNN encoder is run backwards in time to produce an approximate posterior over the initial state: $q(z_0|\{x_i, t_i\}_{i=0}^N)$. Given a sample of z_0, we can find the latent state at any point of interest by solving an ODE initial-value problem. Figure adapted from Chen et al. [2018].

- Improves the generative latent variable framework for irregularly-sampled time series
- Essentially uses an ODE in the encoder where samples are missing
- Shows results on toy data, MuJoCo, PhysioNet

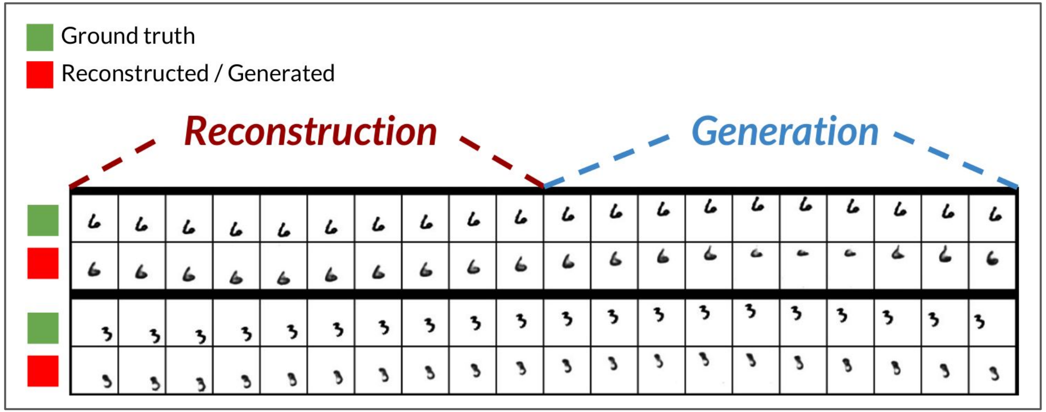
https://arxiv.org/pdf/1907.03907.pdf

Simple Video Generation using Neural ODEs

(David Kanaa*, Vikram Voleti*, Samira Kahou, Christopher Pal; NeurIPS 2019 Workshop)



- Video generation as a generative latent variable model using Neural ODEs



https://sites.google.com/view/neurips2019lire/accepted-papers?authuser=0

ODE2VAE: Deep generative second order ODEs with Bayesian neural networks (Yildiz et al., NeurIPS 2019)

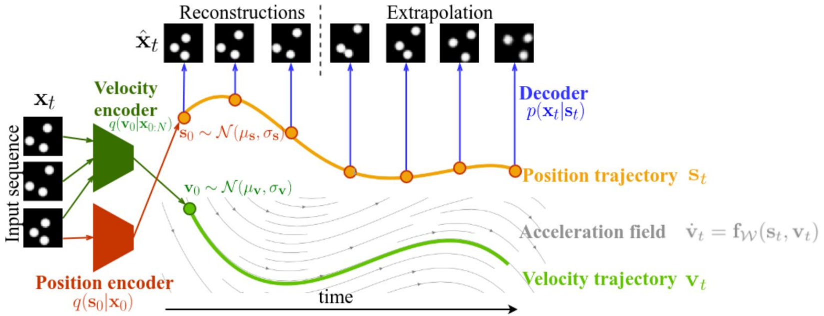


Figure 2: A schematic illustration of ODE^{2}VAE model. Position encoder $(\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\sigma}_\mathbf{s})$ maps the first item $\mathbf{x}_0$ of a high-dimensional data sequence into a distribution of the initial position $\mathbf{s}_0$ in a latent space. Velocity encoder $(\boldsymbol{\mu}_\mathbf{v}, \boldsymbol{\sigma}_\mathbf{v})$ maps the first m high-dimensional data items $\mathbf{x}_{0:m}$ into a distribution of the initial velocity $\mathbf{v}_0$ in a latent space. Probabilistic latent dynamics are implemented by a second order ODE model $\tilde{\mathrm{f}}_\mathcal{W}$ parameterised by a Bayesian deep neural network ($\mathcal{W}$). Data points in the original data domain are reconstructed by a decoder.

- Uses 2nd-order Neural ODE
- Uses a Bayesian Neural Network
- Showed results modelling video generation as a generative latent variable model using (2nd-order Bayesian) Neural ODE

https://papers.nips.cc/paper/9497-ode2vae-deep-generative-second-order-odes-with-bayesian-neural-networks.pdf

On Robustness of Neural Ordinary Differential Equations

(Yan et al., ICLR 2020)

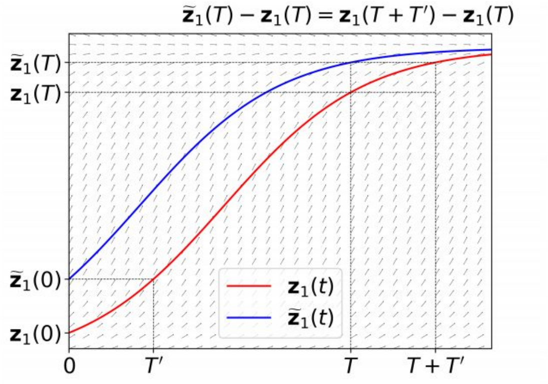
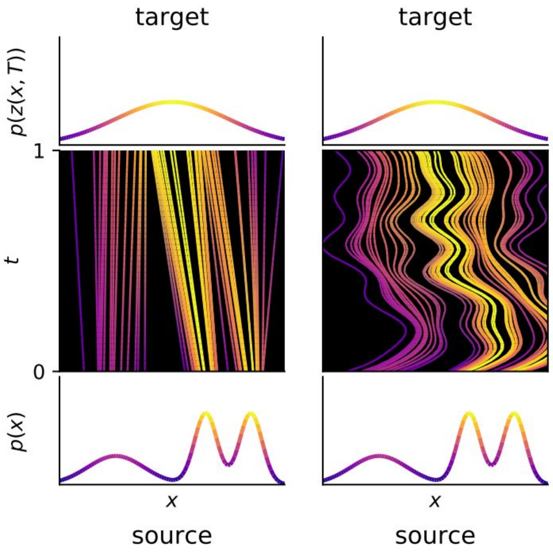


Figure 3: An illustration of the time-invariant property of ODEs. We can see that the curve $\tilde{\mathbf{z}}_1(t)$ is exactly the horizontal translation of $\mathbf{z}_1(t)$ on the interval $[T', \infty)$.

- Ablation study on adversarial attacks on ODE-Nets
- Introduces new regularization term to improve robustness

https://arxiv.org/pdf/1910.05513.pdf, https://openreview.net/pdf?id=B1e9Y2NYvS

How to Train Your Neural ODE (Finlay et al., 2020)



(a) Optimal transport map (b) generic flow

Figure 1. Optimal transport map and a generic normalizing flow.

https://arxiv.org/pdf/2002.02798.pdf

- Makes a link between the flow in Neural ODEs and optimal transport
- Introduces two new regularization terms to constrain flows to straight lines
- Speeds up training of Neural ODEs

Additional References

- http://faculty.bard.edu/belk/math213/InitialValueProblems.pdf
- https://math.temple.edu/~queisser/assets/files/Talk3.pdf
- Textbook : https://users.math.msu.edu/users/gnagy/teaching/ode.pdf
- https://lpsa.swarthmore.edu/NumInt/NumIntFirst.html
- http://homepages.cae.wisc.edu/~blanchar/eps/ivp/ivp
- Excellent blog post on ODE solvers: https://blogs.mathworks.com/loren/2015/09/23/ode-solver-selection-in-matlab/
- Autodiff tutorial: http://www.cs.toronto.edu/~rgrosse/courses/csc421_2019/readings/L06%20Automatic%20Differentiation.pdf
- Course on Neural Networks & Deep Learning by Roger Grosse & Jimmy Ba, University of Toronto - http://www.cs.toronto.edu/~rgrosse/courses/csc421_2019/

Thank you!

voletiv.github.io