Adjoint Sensitivity Equation and NeuralODE

A/Prof Richard Yi Da Xu
richardxu.com

University of Technology Sydney (UTS)

July 26, 2020

Content and reference

This notes is in an elabrated attempt to explain:

- Qiqi Wang's YouTube lecture `https://www.youtube.com/watch?v=7CZP6dHIkNE`
- NeuralODE paper: `https://arxiv.org/abs/1806.07366`
- I did **not** try to unify notations, i.e., I keep notations identical to original reference
- but I will have a page to explain how they convert from one to the other

Motivation: Solving ODE by Separation of Variables

- looking at the simplest ODE example:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y(t)) \qquad y(t_0) = y_0$$

- solution

$$y(t) = y_0 + \int_{t=t_0}^{t} f(t, y(t))\mathrm{d}t$$

- example $\frac{\mathrm{d}y}{\mathrm{d}t} = y(t) \qquad y(0) = 0$

$$\begin{aligned}
\frac{\mathrm{d}y}{\mathrm{d}t} &= y(t) \\
\frac{\mathrm{d}y}{y(t)} &= \mathrm{d}t \\
\int \frac{1}{y(t)}\mathrm{d}y &= \int \mathrm{d}t \\
\ln(y) + C_Y &= t + C_t \\
\ln(y) &= t + C_1 \\
y &= \exp(t + C_1) \\
y &= C\exp(t)
\end{aligned}$$

- substitute $y(0) = 1$:

$$\begin{aligned}
1 &= C\exp(0) \\
\implies C &= 1
\end{aligned}$$

- solution:

$$y = \exp(t)$$

Motivation: Solving ODE by approximation: Euler's method

$$\frac{dy}{dt} = f(t, y(t)), \qquad y(t_0) = y_0$$

- Euler's method:

$$y_{n+1} = y_n + hf(t_n, y_n)$$

- compare with solutions on $\frac{dy}{dt} = f(t, y(t)) = y \qquad y(t_0) = y_0$ and let $h = 1$:

$$\begin{aligned}
y_1 &= y_0 + hf(y_0) = 1 + 1 \cdot 1 = 2 \qquad \exp(1) = 2.71 \\
y_2 &= y_1 + hf(y_1) = 2 + 1 \cdot 2 = 4 \qquad \exp(2) = 7.38 \\
y_3 &= y_2 + hf(y_2) = 4 + 1 \cdot 4 = 8 \qquad \exp(3) = 20.08 \\
y_4 &= y_3 + hf(y_3) = 8 + 1 \cdot 8 = 16 \qquad \exp(4) = 54.59
\end{aligned}$$

- if we substitue $f(t_n, y_n) \equiv \sigma(W_n^\top y_n + B_n)$

$$y_{n+1} = y_n + \sigma(W_n^\top y_n + B_n)$$

tihs is ResNet!

More efficient ODE solver: Adjoint method

- Euler's method is inefficnet and can be replaced by modern solver
- also, Neural netwrks's gradient descent requires to compute $\frac{\partial \mathcal{L}}{\partial \theta}$
- how can we do that when we are given:

$$\frac{\mathrm{d}h}{\mathrm{d}t} \equiv f_\theta(t, h(t))$$

- we need to use **adjoint** method

Adjoint method: motivation through simple example

- look at the problem:

$$\min_{s \in \mathcal{S}} C^\top x$$
$$\text{s.t.} \quad Ax = b(s)$$

- naive way to solve:

1. somehow get a collection of all $s \in \mathcal{S}$:
2. then for each s, one finds a corresponding x using:

$$Ax = b(s)$$

3. substitute $\{x\}$ into $C^\top x$ to see which is smallest

Using Adjoint equation/method:

$$
\begin{aligned}
\min_s \; & C^\top x \\
\text{s.t.} \quad & Ax = b(s)
\end{aligned}
$$

- Adjoint solution

$$
\begin{aligned}
x &= A^{-1} b(s) \\
C^\top x &= \underbrace{C^\top A^{-1}}_{\hat{x}} b(s) = \hat{x} b(s)
\end{aligned}
$$

- $\hat{x}$ is the Adjoint:

$$\hat{x} = C^\top A^{-1} \qquad \text{which is easy, as it does not depend on } s$$

- finally,

$$\min_s C^\top x \;\rightarrow\; \min_s \hat{x} b(s)$$

Apply to discrete difference equation:

$$\min_{s} J(u_n)$$
$$\text{s.t.} \quad u_n = u_n(u_{n-1}, s)$$

- use Lagrange multiplier $\{\hat{u}_n\}$ make the evolution of $u_1 \to u_2 \cdots \to u_n$ satisfy:

$$J = J(u_n) + \hat{u}_n^\top \big(\underbrace{u_n - u_n(u_{n-1}, s)}_{u_n = u_n(u_{n-1}, s)} \big) + \hat{u}_{n-1}^\top \big(\underbrace{u_{n-1} - u_{n-1}(u_{n-2}, s)}_{u_{n-1} = u_{n-1}(u_{n-2}, s)} \big) + \cdots + \hat{u}_0^\top \big(\underbrace{u_0 - u_0(s)}_{u_0 = u_0(s)} \big)$$

- define $\delta_s F \equiv \delta F$ is response of F to an infinitesimal perturbation of s
- solution is:

$$\begin{aligned} \delta J = \frac{\partial J}{\partial u_n} \delta u_n &+ \hat{u}_n^\top \left(\delta u_n - \frac{\partial u_n}{\partial u_{n-1}} \delta u_{n-1} - \frac{\partial u_n}{\partial s} \delta s \right) \quad \text{both } u_n \text{ and } s \text{ is function of } s \\ &+ \hat{u}_{n-1}^\top \left(\delta u_{n-1} - \frac{\partial u_{n-1}}{\partial u_{n-2}} \delta u_{n-2} - \frac{\partial u_{n-1}}{\partial s} \delta s \right) \\ &+ \ldots \\ &+ \hat{u}_0^\top \left(\delta u_{n-1} - \frac{\partial u_0}{\partial s} \delta s \right) \end{aligned}$$

- s is parameter, i.e., θ

- one way to look at:

$$\begin{aligned}\delta J = \frac{\partial J}{\partial u_n}\delta u_n &+ \hat{u}_n^\top\left(\delta u_n - \frac{\partial u_n}{\partial u_{n-1}}\delta u_{n-1} - \frac{\partial u_n}{\partial s}\delta s\right) \quad \text{both } u_n \text{ and } s \text{ is function of } s\\ &+ \hat{u}_{n-1}^\top\left(\delta u_{n-1} - \frac{\partial u_{n-1}}{\partial u_{n-2}}\delta u_{n-2} - \frac{\partial u_{n-1}}{\partial s}\delta s\right)\\ &+ \dots\\ &+ \hat{u}_0^\top\left(\delta u_0 - \frac{\partial u_0}{\partial s}\delta s\right)\end{aligned}$$

- following expression is the usual expression for $\frac{\partial J}{\partial s}$:

$$\begin{aligned}\frac{\partial J}{\partial s} = \frac{\partial J}{\partial u_n}\frac{\partial u_n}{\partial s} &+ \hat{u}_n^\top\left(\frac{\partial u_n}{\partial s} - \frac{\partial u_n}{\partial u_{n-1}}\frac{\partial u_{n-1}}{\partial s} - \frac{\partial u_n}{\partial s}\right) \quad \text{both } u_n \text{ and } s \text{ is function of } s\\ &+ \hat{u}_{n-1}^\top\left(\frac{\partial u_{n-1}}{\partial s} - \frac{\partial u_{n-1}}{\partial u_{n-2}}\frac{\partial u_{n-2}}{\partial s} - \frac{\partial u_{n-1}}{\partial s}\right)\\ &+ \dots\\ &+ \hat{u}_0^\top\left(\frac{\partial u_0}{\partial s} - \frac{\partial u_0}{\partial s}\right)\end{aligned}$$

Clever assignment of Lagrange/Adjoint variables: $\hat{u}_n^\top = -\frac{\partial J}{\partial u_n}$

$$
\begin{aligned}
\delta J &= \frac{\partial J}{\partial u_n}\delta u_n + \hat{u}_n^\top \left(\delta u_n - \frac{\partial u_n}{\partial u_{n-1}}\delta u_{n-1} - \frac{\partial u_n}{\partial s}\delta s\right) + \dots \\
&= \frac{\partial J}{\partial u_n}\delta u_n - \frac{\partial J}{\partial u_n}\left(\delta u_n - \frac{\partial u_n}{\partial u_{n-1}}\delta u_{n-1} - \frac{\partial u_n}{\partial s}\delta s\right) + \dots \qquad \text{let } \hat{u}_n^\top = -\frac{\partial J}{\partial u_n} \\
&= \frac{\partial J}{\partial u_n}\delta u_n - \frac{\partial J}{\partial u_n}\delta u_n + \frac{\partial J}{\partial u_n}\frac{\partial u_n}{\partial u_{n-1}}\delta u_{n-1} + \frac{\partial J}{\partial u_n}\frac{\partial u_n}{\partial s}\delta s + \dots \\
&= \underbrace{-\hat{u}_n^\top \frac{\partial u_n}{\partial s}\delta s + \frac{\partial J}{\partial u_{n-1}}\delta u_{n-1} + \dots} \\
&= -\hat{u}_n^\top \frac{\partial u_n}{\partial s}\delta s + \frac{\partial J}{\partial u_{n-1}}\delta u_{n-1} + \hat{u}_{n-1}^\top\left(\delta u_{n-1} - \frac{\partial u_{n-1}}{\partial u_{n-2}}\delta u_{n-2} - \frac{\partial u_{n-1}}{\partial s}\delta s\right) + \dots \qquad \text{bring second term} \\
&= -\hat{u}_n^\top \frac{\partial u_n}{\partial s}\delta s - \hat{u}_{n-1}^\top \delta u_{n-1} + \hat{u}_{n-1}^\top \delta u_{n-1} - \hat{u}_{n-1}^\top \frac{\partial u_{n-1}}{\partial u_{n-2}}\delta u_{n-2} - \hat{u}_{n-1}^\top \frac{\partial u_{n-1}}{\partial s}\delta s + \dots \qquad \frac{\partial J}{\partial u_{n-1}} = -\hat{u}_{n-1}^\top \\
&= \underbrace{-\hat{u}_n^\top \frac{\partial u_n}{\partial s}\delta s - \hat{u}_{n-1}^\top \frac{\partial u_{n-1}}{\partial s}\delta s - \hat{u}_{n-1}^\top \frac{\partial u_{n-1}}{\partial u_{n-2}}\delta u_{n-2} + \dots} \\
&= \sum_{i=0}^{n} -\hat{u}_i^\top \frac{\partial u_i}{\partial s}\delta s
\end{aligned}
$$

- so we have an expression for $\delta J = \sum_{i=0}^{n} -\hat{u}_i^\top \frac{\partial u_i}{\partial s}\delta s$
- we must solve for $\{\hat{u}_i\}$

Order to solve for Adjoint equations

$$\hat{u}_n^\top = -\frac{\partial J}{\partial u_n} \qquad \frac{\partial J}{\partial u_{n-1}} = -\hat{u}_{n-1}^\top$$

$$\implies \frac{\partial J}{\partial u_{n-1}} = \frac{\partial J}{\partial u_n} \frac{\partial u_n}{\partial u_{n-1}} = -\hat{u}_{n-1}^\top$$

$$\implies -\hat{u}_n^\top \frac{\partial u_n}{\partial u_{n-1}} = -\hat{u}_{n-1}^\top$$

$$\implies \hat{u}_{n-1}^\top = \hat{u}_n^\top \frac{\partial u_n}{\partial u_{n-1}}$$

- meaning we use back-propagation to solve for $\{\hat{u}_i\}$

Continous case with ODE constraint (1)

- Difference equations:

$$\min_s J(u_n)$$
$$\text{s.t.} \quad u_n = u_n(u_{n-1}, s)$$
$$\implies J = J(u_n) + \hat{u}_n^\top (u_n - u_n(u_{n-1}, s)) + \hat{u}_{n-1}^\top (u_{n-1} - u_{n-1}(u_{n-2}, s)) + \cdots + \hat{u}_0^\top (u_0 - u_0(s))$$

- Continous, ODE constrained problem:

$$\min_s J_f(u(T)) + \int_0^T J_c(u(t))\mathrm{d}t$$
$$\text{s.t.} \quad \frac{\mathrm{d}u}{\mathrm{d}t} = f(u, s(t))$$
$$\implies J = J_f(u(T)) + \int_0^T J_c(u(t)\mathrm{d}t + \int_0^T \hat{u}(t)\left(\frac{\mathrm{d}u}{\mathrm{d}t} - f(u,s)\right)\mathrm{d}t \quad \text{continuous Lagrange}$$

- difficulty arise as having a derivative function in the Lagrange

Simplify objective equation

- instead of including $\int_0^T J_c(u(t))\mathrm{d}t$ (for aerodynamics application)

$$\min_s J_f(u(T)) + \int_0^T J_c(u(t))\mathrm{d}t$$

$$\text{s.t.} \quad \frac{\mathrm{d}u}{\mathrm{d}t} = f(u, s(t))$$

$$\implies J = J_f(u(T)) + \int_0^T J_c(u(t)\mathrm{d}t + \int_0^T \hat{u}(t)\left(\frac{\mathrm{d}u}{\mathrm{d}t} - f(u, s)\right)\mathrm{d}t$$

- we simplify by removing $\int_0^T J_c(u(t))\mathrm{d}t$:

$$\min_s J_f(u(T))$$

$$\text{s.t.} \quad \frac{\mathrm{d}u}{\mathrm{d}t} = f(u, s(t))$$

$$\implies J = J_f(u(T)) + \int_0^T \hat{u}(t)\left(\frac{\mathrm{d}u}{\mathrm{d}t} - f(u, s)\right)\mathrm{d}t$$

Adjoint equation for ODE (1):

- substitute: $\hat{u}(T) = -\frac{\partial J_f}{\partial u}$:

$$
\begin{aligned}
J &= J_f(u(T)) + \int_0^T \hat{u}(t)\left(\frac{\mathrm{d}u}{\mathrm{d}t} - \underbrace{f(u,s)}\right)\mathrm{d}t \\
\Longrightarrow \quad \delta J &= \frac{\partial J_f}{\partial u}\delta u(T)\mathrm{d}t + \int_0^T \hat{u}(t)\left(\frac{\mathrm{d}\delta u}{\mathrm{d}t} \underbrace{- \frac{\partial f}{\partial u}\delta u - \frac{\partial f}{\partial s}\delta s}\right)\mathrm{d}t \\
&= \frac{\partial J_f}{\partial u}\delta u(T) + \int_0^T \hat{u}(t)\frac{\mathrm{d}\delta u}{\mathrm{d}t}\,\mathrm{d}t - \int_0^T \left(\frac{\partial f}{\partial u}\right)^\top \hat{u}(t)\delta u\,\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial s}\delta s\,\mathrm{d}t \\
&= \frac{\partial J_f}{\partial u}\delta u(T) + \left[\underbrace{\hat{u}(t)\delta u\Big|_0^T} - \int_0^T \delta u\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\mathrm{d}t\right] - \int_0^T \left(\frac{\partial f}{\partial u}\right)^\top \hat{u}(t)\delta u\,\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial s}\delta s\,\mathrm{d}t \\
&= \frac{\partial J_f}{\partial u}\delta u(T) + \underbrace{\hat{u}(T)\delta u(T)} - \int_0^T \delta u\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\mathrm{d}t - \int_0^T \left(\frac{\partial f}{\partial u}\right)^\top \hat{u}(t)\delta u\,\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial s}\delta s\,\mathrm{d}t \\
&= \cancel{\frac{\partial J_f}{\partial u}\delta u(T)} + \cancel{-\frac{\partial J_f}{\partial u}\delta u(T)} - \int_0^T \delta u\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\mathrm{d}t - \int_0^T \left(\frac{\partial f}{\partial u}\right)^\top \hat{u}(t)\delta u\,\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial s}\delta s\,\mathrm{d}t \\
&= -\int_0^T \delta u\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial u}\delta u\,\mathrm{d}t - \int_0^T \hat{u}(t)\frac{\partial f}{\partial s}\delta s\,\mathrm{d}t
\end{aligned}
$$

- integration by parts: $\int_a^b u(x)v'(x)\,dx = \Big[u(x)v(x)\Big]_a^b - \int_a^b u'(x)v(x)\,\mathrm{d}x$

Adjoint equation for ODE (2)

- we already let $\hat{u}(T) = -\frac{\partial J_f}{\partial u(T)}$
- further, we let $\hat{u}(t)$: instead of let $\hat{u}(t) = -\frac{\partial J_f}{\partial u(t)} \equiv -\frac{\partial J_f}{\partial u}(t)$:
- **important: if we can prove**:

$$\hat{u}(t) = -\frac{\partial J_f}{\partial u(t)} \quad \Longrightarrow \quad \frac{\mathrm{d}\hat{u}}{\mathrm{d}t} = -\left(\frac{\partial f}{\partial u(t)}\right)^\top \hat{u}(t)$$

- then we can solve δJ:

$$\begin{aligned}
\delta J &= -\int_0^T \delta u \frac{\mathrm{d}\hat{u}}{\mathrm{d}t} \mathrm{d}t - \int_0^T \left(\frac{\partial f}{\partial u(t)}\right)^\top \hat{u}(t) \delta u \, \mathrm{d}t - \int_0^T \hat{u}(t) \frac{\partial f}{\partial s} \delta s \, \mathrm{d}t \\
&= \cancel{\int_0^T \left(\frac{\partial f}{\partial u(t)}\right)^\top} \hat{u}(t) \delta u(t) \, \mathrm{d}t - \cancel{\int_0^T \left(\frac{\partial f}{\partial u}\right)^\top} \hat{u}(t) \delta u \, \mathrm{d}t - \int_0^T \hat{u}(t) \frac{\partial f}{\partial s} \delta s \, \mathrm{d}t \\
&= -\int_0^T \hat{u}(t) \frac{\partial f}{\partial s} \delta s \, \mathrm{d}t
\end{aligned}$$

- by running another ODE to solve $\hat{u}(t)$, from dynamic equation $\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}$

Compare with discrete/difference equation

- continous case:

$$\delta J = -\int_0^T \hat{u}(t) \frac{\partial f}{\partial s} \delta s \, \mathrm{d}t$$

- compare with discrete case:

$$\delta J = \sum_{i=0}^{n} -\hat{u}_i^\top \frac{\partial u_i}{\partial s} \delta s$$

Proof for $\hat{u}(t) = -\frac{\partial J_f}{\partial u(t)} \implies \frac{\mathrm{d}\hat{u}}{\mathrm{d}t} = -\left(\frac{\partial f}{\partial u(t)}\right)^\top \hat{u}(t)$ (1)

- here we change notions to use Neural ODE:

$$
\begin{aligned}
\hat{u}(t) &= -\frac{\partial J_f}{\partial u(t)} \implies \frac{\mathrm{d}\hat{u}}{\mathrm{d}t} = -\left(\frac{\partial f}{\partial u(t)}\right)^\top \hat{u}(t) \\
\mathbf{a}(t) &= \frac{\partial L}{\partial \mathbf{z}(t)} \implies \frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t} = -\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)^\top \mathbf{a}(t)
\end{aligned}
$$

- similary

$$
\begin{aligned}
\delta J &= -\int_0^T \hat{u}(t) \frac{\partial f}{\partial s} \delta s \, \mathrm{d}t \\
\frac{\partial L}{\partial \theta} \equiv \frac{\partial L}{\partial \theta}(t_0) &= -\int_{t_N}^{t_0} \mathbf{a}(t) \frac{\partial f}{\partial \theta} \, \mathrm{d}t
\end{aligned}
$$

- sign change may be caused by swapping the integrand t_0 and t_N

Proof for $\mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{z}(t)} \implies \frac{d\mathbf{a}}{dt} = -\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)^\top \mathbf{a}(t)$

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

$$\mathbf{z}(t+\epsilon) = \int_t^{t+\epsilon} f(\mathbf{z}(t), t)dt = T_\epsilon(\mathbf{z}(t), t)$$

$$\frac{dL}{\partial \mathbf{z}(t)} = \frac{dL}{d\mathbf{z}(t+\epsilon)} \frac{d\mathbf{z}(t+\epsilon)}{d\mathbf{z}(t)} \quad \implies \quad \mathbf{a}(t) = \mathbf{a}(t+\epsilon) \frac{\partial T_\epsilon(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)}$$

$$\begin{aligned}
\frac{d\mathbf{a}(t)}{dt} &= \lim_{\epsilon \to 0+} \frac{\mathbf{a}(t+\epsilon) - \mathbf{a}(t)}{\epsilon} \\
&= \lim_{\epsilon \to 0+} \frac{\mathbf{a}(t+\epsilon) - \mathbf{a}(t+\epsilon)\frac{\partial}{\partial \mathbf{z}(t)} T_\epsilon(\mathbf{z}(t), t)}{\epsilon} \\
&\approx \lim_{\epsilon \to 0+} \frac{\mathbf{a}(t+\epsilon) - \mathbf{a}(t+\epsilon)\frac{\partial}{\partial \mathbf{z}(t)} (\mathbf{z}(t) + \epsilon f(\mathbf{z}(t), t))}{\epsilon} \quad \text{Taylor } T_\epsilon(\mathbf{z}(t), t) \approx \mathbf{z}(t) + \epsilon f(\mathbf{z}(t), t) \\
&= \lim_{\epsilon \to 0+} \frac{\mathbf{a}(t+\epsilon) - \mathbf{a}(t+\epsilon)\left(\frac{\partial \mathbf{z}(t)}{\partial \mathbf{z}(t)} + \frac{\partial \epsilon f(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)}\right)}{\epsilon} = \lim_{\epsilon \to 0+} \frac{\mathbf{a}(t+\epsilon) - \mathbf{a}(t+\epsilon)\epsilon\left(I + \frac{\partial f(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)}\right)}{\epsilon} \\
&= \lim_{\epsilon \to 0+} \frac{-\mathbf{a}(t+\epsilon)\epsilon \frac{\partial f(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)}}{\epsilon} = \lim_{\epsilon \to 0+} -\mathbf{a}(t+\epsilon) \frac{\partial f(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)} \\
&= -\mathbf{a}(t) \frac{\partial f(\mathbf{z}(t), t)}{\partial \mathbf{z}(t)} \\
&= -\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)^\top \mathbf{a}(t) \quad \text{multi-dimensional case}
\end{aligned}$$

How to compute adjoint $\mathbf{a}(t)$ from its dynamics $\frac{\mathrm{d}\mathbf{a}(t)}{\mathrm{d}t}$

- just like $\hat{x} = C^\top x$, but complicated.
- initial condition: $\mathbf{a}(t_N) = \frac{\mathrm{d}L}{\mathrm{d}\mathbf{z}(t_N)}$, and run backwards to obtain $\mathbf{a}(t_0)$:

$$\begin{aligned}
\mathbf{a}(t_0) &= \mathbf{a}(t_N) + \int_{t_N}^{t_0} \frac{\mathrm{d}\mathbf{a}(t)}{\mathrm{d}t}\,\mathrm{d}t \\
&= \mathbf{a}(t_N) - \int_{t_N}^{t_0} \left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)^\top \mathbf{a}(t)\,\mathrm{d}t \quad \text{substitute}
\end{aligned}$$

main learning task $\frac{\partial L}{\partial \theta}$ and $\frac{\partial L}{\partial t}$ (1)

- $\frac{\partial L}{\partial \theta}$:

$$\frac{\partial L}{\partial \theta} \equiv \frac{\partial L}{\partial \theta}(t_0) = \frac{\partial L}{\partial \theta}(t_N) - \int_{t_N}^{t_0} \mathbf{a}(t) \frac{\partial f}{\partial \theta}(t) \, \mathrm{d}t$$

dynamic equation: $-\mathbf{a}(t) \frac{\partial f}{\partial \theta}(t)$ initial condition: $\frac{\partial L}{\partial \theta}(t_N) = \mathbf{0}$

- $\frac{\mathrm{d}L}{\mathrm{d}t_0}$:

$$\frac{\mathrm{d}L}{\mathrm{d}t_0} \equiv \frac{\mathrm{d}L}{\mathrm{d}t}(t_0) = \frac{\mathrm{d}L}{\mathrm{d}t}(t_N) - \int_{t_N}^{t_0} \mathbf{a}(t) \frac{\mathrm{d}f}{\mathrm{d}t}(t) \, \mathrm{d}t$$

dynamic equation: $-\mathbf{a}(t) \frac{\mathrm{d}f}{\mathrm{d}t}(t)$ initial condition: $\frac{\mathrm{d}L}{\mathrm{d}t}(t_N) = \mathbf{a}(t_N) f(\mathbf{z}(t_N, t_N))$

- combine with $\mathbf{a}(t_0)$:

$$\mathbf{a}(t_0) = \mathbf{a}(t_N) - \int_{t_N}^{t_0} \mathbf{a}(t)\frac{\partial f}{\partial \mathbf{z}}(t)\mathrm{d}t$$

dynamic equation: $-\mathbf{a}(t)\frac{\partial f}{\partial \mathbf{z}}(t)$ initial condition: $\frac{\partial L}{\partial \mathbf{z}(t_N)} = \mathbf{a}(t_N)$

- combine with $\mathbf{z}(t_0)$:

$$\mathbf{z}(t_0) = \mathbf{z}(t_N) - \int_{t_N}^{t_0} f(\mathbf{z}, t)\mathrm{d}t$$

dynamic equation: $f(\mathbf{z}, t)$ initial condition: $\mathbf{z}(t_N)$

Combine everything in a backward $t_N \to t_0$ pass

- combined back dynamics:

$$-\left[f(\mathbf{z}, t) \quad \mathbf{a}\frac{\partial f}{\partial \mathbf{z}} \quad \mathbf{a}\frac{\partial f}{\partial \theta} \quad \mathbf{a}\frac{\mathrm{d}f}{\mathrm{d}t} \right](t)$$

- to solve:

$$\left[\mathbf{z}(t_0) \quad \mathbf{a}(t_0) \quad \frac{\partial L}{\partial \theta}(t_0) \quad \frac{\mathrm{d}L}{\mathrm{d}t}(t_0) \right]$$

- with initial condition at t_N:

$$\left[\mathbf{z}(t_N) \quad \mathbf{a}(t_N) \quad \mathbf{0} \quad \frac{\mathrm{d}f}{\mathrm{d}t}(t_N) \right]$$

Neural ODE: feed-forward much easier

- Inference using feed-forward:

$$\mathcal{L}(\mathbf{z}(t_1)) = \mathcal{L}\left(\mathbf{z}(t_0) + \int_{t=0}^{t_1} \underbrace{f(\mathbf{z}(t), t, \theta)}_{\mathrm{NN}^t(z)} \mathrm{d}t\right)$$

- $f(\mathbf{z}(t), t, \theta) \equiv \mathrm{NN}^t(z)$ is Neural Network at t^{th} infinitesimal layer
- we can solve this by in a Solver, to specify:
  1. boundary condition: $\mathbf{z}(t_0)$
  2. start and end time t_0 and t_1
  3. dynamic equation: $f(\mathbf{z}(t), t, \theta)$
- the solver looks like:

$$\mathcal{L}(\mathbf{z}(t_1)) = \mathcal{L}\Big(\mathrm{ODESolve}(\mathbf{z}(t_0), t_0, t_1, f(\mathbf{z}(t), t, \theta))\Big)$$