Neural ODE A Journey in Learning

Allen C.

August 2022

1 Introduction: Adjoint Method for ODE

Question: with given function including parameter p: $F(x,p)$ where

$$F(x,p) = \int_0^T f(x,p,t)dt$$

we minimize the objective

$$\min_p F(x,p)$$

subject to the constrained condition

$$h(x,\dot{x},p,t) = 0$$

and the initial condition

$$g(x(0),p) = 0.$$

To solve this, we define the Lagrangian:

$$L = \int_0^T \Big[f(x,t,p) + \lambda^{\mathrm{T}} h(x,\dot{x},t,p)\Big] + \mu^{\mathrm{T}} g(x(0),p)$$

since the constrained condition and the initial condition are equal to 0, then we have:

$$\frac{dL}{dp} = \frac{dF}{dp} \tag{1}$$

In order to solve $\frac{dF}{dp}$ to obtain the minimal solution, we solve the $\frac{dL}{dp}$:

$$\begin{aligned}\frac{dF}{dp} = \frac{\partial L}{\partial p} = \int_0^T \frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial p} + \frac{\partial f}{\partial p} + \lambda^{\mathrm{T}} \Big(\frac{\partial h}{\partial x} \cdot \frac{\partial x}{\partial p} + \frac{\partial h}{\partial \dot{x}} \cdot \frac{\partial \dot{x}}{\partial p} + \frac{\partial h}{\partial p}\Big) dt \\ + \mu^{\mathrm{T}} \Big(\frac{\partial g}{\partial x(0)} \cdot \frac{\partial x(0)}{\partial p} + \frac{\partial g}{\partial p}\Big)\end{aligned} \tag{2}$$

Sine $\frac{\partial \dot{x}}{\partial p}$ is difficult to compute, then we apply integration by parts to avoid this term:

$$\begin{aligned}
\frac{dF}{dp} = \frac{\partial L}{\partial p} &= \int_0^T \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial p} + \frac{\partial f}{\partial p} + \lambda^{\mathrm{T}}\Big(\frac{\partial h}{\partial x}\cdot\frac{\partial x}{\partial p} + \frac{\partial h}{\partial p}\Big)dt \\
&+ \int_0^T \lambda^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}}\cdot\frac{\partial \dot{x}}{\partial p}dt + \mu^{\mathrm{T}}\Big(\frac{\partial g}{\partial x(0)}\cdot\frac{\partial x(0)}{\partial p} + \frac{\partial g}{\partial p}\Big) \\
&= \int_0^T \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial p} + \frac{\partial f}{\partial p} + \lambda^{\mathrm{T}}\Big(\frac{\partial h}{\partial x}\cdot\frac{\partial x}{\partial p} + \frac{\partial h}{\partial p}\Big)dt \\
&+ \mu^{\mathrm{T}}\Big(\frac{\partial g}{\partial x(0)}\cdot\frac{\partial x(0)}{\partial p} + \frac{\partial g}{\partial p}\Big) \\
&+ \lambda^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}}\cdot\frac{\partial x}{\partial p}\Big|_0^T - \int_0^T \Big(\dot{\lambda}^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}} + \lambda^{\mathrm{T}}\frac{d}{dt}(\frac{\partial h}{\partial \dot{x}})\Big)\cdot\frac{\partial x}{\partial p}dt
\end{aligned} \tag{3}$$

Assume $\lambda^{\mathrm{T}}|_T = 0$: we have

$$\begin{aligned}
\frac{dF}{dp} =& \frac{\partial L}{\partial p} \\
=& \int_0^T \left(\frac{\partial f}{\partial x} + \lambda^{\mathrm{T}}\Big(\frac{\partial h}{\partial x} - \frac{d}{dt}(\frac{\partial h}{\partial \dot{x}})\Big) - \dot{\lambda}^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}}\right)\cdot\frac{\partial x}{\partial p}dt \\
&+ \int_0^T \Big(\frac{\partial f}{\partial p} + \lambda^{\mathrm{T}}\frac{\partial h}{\partial p}\Big)dt \\
&+ \mu^{\mathrm{T}}\frac{\partial g}{\partial p} + \mu^{\mathrm{T}}\frac{\partial g}{\partial x(0)}\cdot\frac{\partial x(0)}{\partial p} - \lambda^{\mathrm{T}}\frac{\partial x(0)}{\partial p}\cdot\frac{\partial h}{\partial \dot{x}}\Big|_{t=0}
\end{aligned} \tag{4}$$

Let

$$\frac{\partial f}{\partial x} + \lambda^{\mathrm{T}}\Big(\frac{\partial h}{\partial x} - \frac{d}{dt}(\frac{\partial h}{\partial \dot{x}})\Big) - \dot{\lambda}^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}} = 0$$

and

$$\mu^{\mathrm{T}}\frac{\partial g}{\partial x(0)} = \lambda^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}}\Big|_{t=0}$$

Then we have the following form for $\frac{dF}{dp}$:

$$\frac{dF}{dp} = \int_0^T \Big(\frac{\partial f}{\partial p} + \lambda^{\mathrm{T}}\frac{\partial h}{\partial p}\Big)dt + \lambda^{\mathrm{T}}\frac{\partial h}{\partial \dot{x}}\Big|_{t=0}\frac{\partial g}{\partial p}\cdot\frac{\partial g^{-1}}{\partial x(0)} \tag{5}$$

Here $\frac{\partial f}{\partial p}$, $\frac{\partial h}{\partial p}$, $\frac{\partial h}{\partial \dot{x}}$, $\frac{\partial g}{\partial p}$ and $\frac{\partial g^{-1}}{\partial x(0)}$ are easily to compute. To better understand the methodology, we illustrate the methodology with the following examples:

Set $F(x,p) = \int_0^T x dt$, and $F(x,t)$ is subject to $\dot{x} = bx \quad x(0) = a$. Then we can it in the above form:

$$\begin{aligned}F(x,p,t) &= \int_0^T f(x,p,t)dt \\ f(x,p,t) &= x \\ h(\dot{x},x,p,t) &= \dot{x} - bx \\ g(x(0),p) &= x(0) - a \\ p &= \begin{pmatrix} a \\ b \end{pmatrix}\end{aligned} \tag{6}$$

To obtain the derivative with respect a and b for $F(x,p)$, we apply the adjoint method as described above. We can define the equations for λ and μ as following:

$$\begin{aligned}1 - b\lambda - \dot{\lambda} &= 0 \\ \mu &= \lambda\end{aligned} \tag{7}$$

We have $\lambda = \mu = b^{-1}(1 - \mathrm{e}^{b(T-t)})$, and $x = a\mathrm{e}^{bt}$. Then the derivatives $\frac{\partial F}{\partial a}$ and $\frac{\partial F}{\partial b}$ are calculated as:

$$\begin{aligned}\frac{\partial F}{\partial a} &= \int_0^T \Big(\frac{\partial f}{\partial a} + \lambda \frac{\partial h}{\partial a}\Big)dt + \lambda \frac{\partial h}{\partial \dot{x}}\Big|_{t=0} \frac{\partial g}{\partial x(0)}^{-1} \cdot \frac{\partial g}{\partial a} \\ &= b^{-1}(\mathrm{e}^{bT} - 1)\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial F}{\partial b} &= \int_0^T \Big(\frac{\partial f}{\partial b} + \lambda \frac{\partial h}{\partial b}\Big)dt + \lambda \frac{\partial h}{\partial \dot{x}}\Big|_{t=0} \frac{\partial g}{\partial x(0)}^{-1} \cdot \frac{\partial g}{\partial b} \\ &= \frac{aT}{b}\mathrm{e}^{bT} - \frac{a}{b^2}(\mathrm{e}^{bT} - 1)\end{aligned} \tag{9}$$

2 Neural ODE

$$\frac{dz(t)}{dt} = f(z(t), t, \theta) \tag{10}$$

where $z(t)$ is the state, t is time, and θ is the weight or we can call it parameter. To optimize the loss function L with respect to θ, we define the loss function as

$$L = L(z(t_1)) \tag{11}$$

and calculate the gradient with respect to θ:

$$\frac{dL}{d\theta} = \frac{dL}{dz(t_1)} \frac{z(t_1)}{d\theta} \tag{12}$$

How to calculate the gradient? In the neural ODE paper, the authors define

$$\frac{dz(t)}{dt} = f(z(t), t, \theta) \tag{13}$$

with $z(t_0)$ known. Then the authors derive the following ODEs:

$$\frac{da(t)}{dt} = -a(t)\frac{\partial f(z(t),t,\theta)}{\partial z(t)} \tag{14}$$

with $a(t) = \frac{dL}{dz(t)}$ and $a(t_1)$ known. And

$$\frac{da_\theta(t)}{dt} = -a_\theta(t)\frac{\partial f(z(t),t,\theta)}{d\theta} \quad a_\theta(t) = \frac{dL}{d\theta} \tag{15}$$

Then they present

$$\frac{dL}{d\theta} = a_\theta(t_0) = a_\theta(t_1) + \int_{t_1}^{t_0} a(t)\frac{\partial f}{\partial\theta}dt \tag{16}$$

with $a_\theta(t_1) = 0$. **Why??**

In the following, we derive the adjoint equation and corresponding answer the 'Why'.

Objective question:

$$\min_\theta L(z(t_1)) \tag{17}$$

subject to

$$F(\dot(z)(t), z(t), t, \theta) = \dot(z)(t) - f(z(t), t, \theta) = 0, \tag{18}$$

with $z(t_0)$ known. We apply Lagrange for the above problem:

$$\psi = L(z(t_1)) - \int_{t_0}^{t_1} \lambda(t)F(\dot(z)(t), t, \theta)dt \tag{19}$$

Since $F(\dot(z)(t), z(t), t, \theta) = 0$, then we have

$$\frac{d\psi}{d\theta} = \frac{dL(z(t_1))}{d\theta} = \frac{dL}{dz(t_1)} \cdot \frac{dz(t_1)}{d\theta} \tag{20}$$

Here we see $\frac{dz(t_1)}{d\theta}$ is time consuming to calculate. Then in the following we apply the Lagrange multiplier to the above problem to avoid calculating some terms.
First:

$$\begin{aligned}
\int_{t_0}^{t_1} \lambda(t)Fdt &= \int_{t_0}^{t_1} \lambda(t)\Big(\dot(z)(t) - f\Big) \\
&= \int_{t_0}^{t_1} \lambda(t)\dot(z)(t)dt - \int_{t_0}^{t_1} \lambda(t)fdt \\
&= \lambda(t)z(t)|_{t_0}^{t_1} - \int_{t_0}^{t_1} z(t)\dot(\lambda)(t)dt - \int_{t_0}^{t_1} \lambda(t)fdt \\
&= \lambda(t_1)z(t_1) - \lambda(t_0)z(t_0) - \int_{t_0}^{t_1} z(t)\dot{\lambda}(t) + \lambda(t)fdt
\end{aligned}$$

Second:

$$\frac{d}{d\theta}\Big[\int_{t_0}^{t_1}\lambda(t)Fdt\Big] = \lambda(t_1)\frac{dz(t_1)}{d\theta} - \lambda(t_0)\frac{dz(t_0)}{d\theta} - \int_{t_0}^{t_1}\Big(\frac{dz}{d\theta}\dot{\lambda}(t) + \lambda(t)\frac{df}{d\theta}\Big)$$

where $\lambda(t_0)\frac{dz(t_0)}{d\theta} = 0$ by the initial condition. By applying the chain rule, we have

$$\frac{df}{d\theta} = \frac{\partial f}{\partial\theta} + \frac{df}{dz}\cdot\frac{dz}{d\theta} \tag{21}$$

Then we obtain:

$$\begin{aligned}&\frac{d}{d\theta}\Big[\int_{t_0}^{t_1}\lambda(t)Fdt\Big]\\ =&\lambda(t_1)\frac{dz(t_1)}{d\theta} - \int_{t_0}^{t_1}(\dot{\lambda}(t) + \lambda(t)\frac{\partial f}{\partial z})\frac{dz}{d\theta}dt - \int_{t_0}^{t_1}\lambda(t)\frac{\partial f}{\partial\theta}dt\end{aligned}$$

Finally:

$$\begin{aligned}\frac{dL}{d\theta} &= \frac{\partial L}{\partial z(t_1)}\cdot\frac{dz(t_1)}{d\theta} - \frac{d}{d\theta}\Big[\int_{t_0}^{t_1}\lambda(t)Fdt\Big]\\ &= \Big[\frac{\partial L}{\partial z(t_1)} - \lambda(t_1)\Big]\frac{dz(t_1)}{d\theta} + \int_{t_0}^{t_1}(\dot{\lambda}(t) + \lambda(t)\frac{\partial f}{\partial z})\frac{dz}{d\theta}dt + \int_{t_0}^{t_1}\lambda\frac{\partial f}{\partial\theta}dt\end{aligned}$$

$\frac{\partial L}{\partial z(t_1)}$, $\dot{\lambda}(t)$, $\lambda(t)\frac{\partial f}{\partial z}$, $\lambda(t)\frac{\partial f}{\partial\theta}$ are easy to solve, but $\frac{dz(t_1)}{d\theta}$ and $\frac{dz(t)}{d\theta}$ are time consuming to calculate. Then here we eliminate these two terms by letting

$$\dot{\lambda}(t) + \lambda(t)\frac{\partial f}{\partial z} = 0 \tag{22}$$

with $\lambda(t_1) = \frac{\partial L}{\partial z(t_1)}$. Then we have

$$\frac{dL}{d\theta} = \int_{t_0}^{t_1}\lambda(t)\frac{\partial f}{\partial\theta}dt = -\int_{t_1}^{t_0}\lambda(t)\frac{\partial f}{\partial\theta}dt. \tag{23}$$

Here we can see that

$$\dot{\lambda}(t) + \lambda(t)\frac{\partial f}{\partial z} = 0 \tag{24}$$

is exactly the same as $a(t)$ defined in the Neural ODE paper and the equation to calculate $\frac{dL}{d\theta}$ is the same as mentioned in the paper.