1 Adjoint sensitivity method

Consider a problem where our aim is to minimize a function $L(z(t), \theta)$, where $z(t)$ is a solution of an initial value problem and θ are the parameters of the model, including parameters of a differential equation. Namely, with the assumption that the function L can be written as an integral of l function

$$L(z(t), t) = \int_0^T l(z(t), t, \theta) dt.$$

We are trying to find $\min_\theta L(z(t), \theta)$ with the constraints:

$$\begin{cases} \dot{z}(t) = f(z(t), t, \theta) \\ \quad z(0) = z_0(\theta) \end{cases}$$

The practical methods, especially the gradient-based methods which are used to optimize the loss function L usually require calculating the total derivative $\frac{dL}{d\theta}$. It is not an easy task since $z(t)$ implicitly depends on θ through the ODE function f and initial conditions z_0. Firstly, we roughly compute by the definition of total derivative

$$\frac{dL}{d\theta} = \int_0^T \left[\frac{dl(z(t), t, \theta)}{d\theta} \right] dt = \int_0^T \left[\frac{\partial l}{\partial z(t)} \frac{dz(t)}{d\theta} + \frac{\partial l}{\partial \theta} \right] dt$$

The adjoint sensitivity method introduce the new adjoint variable $a(t)$ which appear the following adjusted loss function

$$J = \int_0^T \left[l(z(t), t, \theta) + a(t)^T (f(z(t), t, \theta) - \dot{z}(t)) \right] dt + \mu^T (z(0) - z_0)$$

where $a(t)$ and μ are functions which associated with ODE constrains and will be chosen later. Note that, if both ODE conditions are satisfied then the adjusted loss function J actually is the original function L. So, we will compute the full derivative on the function J

$$\frac{dJ}{d\theta} = \int_0^T \left[\frac{dl}{d\theta} + a(t)^T \frac{d(f(z(t), t, \theta) - \dot{z}(t))}{d\theta} \right] dt + \mu^T \frac{d(z(0) - z_0)}{d\theta}$$

$$= \int_0^T \left[\frac{\partial l}{\partial z(t)} \frac{dz(t)}{d\theta} + \frac{\partial l}{\partial \theta} + a(t)^T \left(\frac{\partial f}{\partial z(t)} \frac{dz(t)}{d\theta} + \frac{\partial f}{\partial \theta} \right) \right] dt - \int_0^T a(t)^T \frac{d\dot{z}(t)}{d\theta} dt + \mu^T \left(\frac{dz(0)}{d\theta} - \frac{dz_0}{d\theta} \right)$$

where we can use the integration by part to compute the second term

$$\int_0^T a(t)^T \frac{d\dot{z}(t)}{d\theta} dt = a(t)^T \frac{dz(t)}{d\theta} \Big|_0^T - \int_0^T \dot{a}(t)^T \frac{dz(t)}{d\theta} dt$$

$$= a(T)^T \frac{dz(T)}{d\theta} - a(0)^T \frac{dz(0)}{d\theta} - \int_0^T \dot{a}(t)^T \frac{dz(t)}{d\theta} dt$$

Hence,

$$
\begin{aligned}
\frac{dJ}{d\theta} &= \int_0^T \left[\frac{\partial l}{\partial z(t)} \frac{dz(t)}{d\theta} + \frac{\partial l}{\partial \theta} + a(t)^T \left(\frac{\partial f}{\partial z(t)} \frac{dz(t)}{d\theta} + \frac{\partial f}{\partial \theta} \right) + \dot{a}(t)^T \frac{dz(t)}{d\theta} \right] dt \\
&\quad - a(T)^T \frac{dz(T)}{d\theta} + a(0)^T \frac{dz(0)}{d\theta} + \mu^T \left(\frac{dz(0)}{d\theta} - \frac{dz_0}{d\theta} \right) \\
&= \int_0^T \left[\left(\frac{\partial l}{\partial z(t)} + a(t)^T \frac{\partial f}{\partial z(t)} + \dot{a}(t)^T \right) \frac{dz(t)}{d\theta} + \frac{\partial l}{\partial \theta} + a(t)^T \frac{\partial f}{\partial \theta} \right] dt \\
&\quad - a(T)^T \frac{dz(T)}{d\theta} + (a(0) + \mu)^T \frac{dz(0)}{d\theta} - \mu^T \frac{dz_0}{d\theta}
\end{aligned}
$$

Chose $a(t), \mu$ such that we can omit the difficult terms

$$
\begin{cases}
a(T) = 0 \\
a(0) + \mu = 0 \\
\dfrac{\partial l}{\partial z(t)} + a(t)^T \dfrac{\partial f}{\partial z(t)} + \dot{a}(t)^T = 0
\end{cases}
$$

So, we can rewrite

$$
\frac{dJ}{d\theta} = \int_0^T \left[\frac{\partial l}{\partial \theta} + a(t)^T \frac{\partial f}{\partial \theta} \right] dt - \mu^T \frac{dz_0}{d\theta} = \frac{\partial L}{\partial \theta} + \int_0^T a(t)^T \frac{\partial f}{\partial \theta} dt + a(0)^T \frac{dz_0}{d\theta}
$$

Problem: There is a process following an unknown ODE and some observations along its trajectory

$$
\begin{cases}
\dfrac{dz(t)}{dt} = f(z(t), t) \\
\{(z_0, t_0), (z_1, t_1), ..., (z_M, t_M)\} - \text{observations}
\end{cases}
$$

Is it possible to find an approximation $\widehat{f}(z, t, \theta)$ of dynamics function $f(z, t)$?

First, consider a somewhat simpler task: there are only 2 observations, at the beginning and at the end of the trajectory, $(z_0, t_0), (z_1, t_1)$. One starts the evolution of the system from z_0, t_0 for time $t_1 - t_0$ with some parameterized dynamics function using any ODE initial value solver. After that, one ends up being at some new state $\hat{z_1}, t_1$, compares it with the observation z_1, and tries to minimize the difference by varying the parameters θ.

Or, more formally, consider optimizing the following loss function $L(\hat{z_1})$

$$
L(z(t_1)) = L\left(z(t_0) + \int_{t_0}^{t_1} f(z(t), t, \theta) dt \right) = L\big(\text{ODESolve}(z(t_0), f, t_0, t_1, \theta)\big)
$$

On this specify problem we can see the loss function does not rely on any parameters directly, so $\dfrac{\partial L}{\partial \theta} = 0$. Note that we consider $L()$ as a scalar-valued loss function, whose input is the result of an ODE solver. The loss depends only on the last point $z(T)$. Therefore, the function $l(z(t), t, \theta)$ is non-zero only at

2

time point T. We have to account for the direct dependency of the loss on the last hidden state. Thus, we add the gradient of the loss to the constraint on the last time point:

$$\begin{cases} a(T) = \dfrac{\partial l}{\partial z(t)} \\ \dot{a}(t)^T = -a(t)^T \dfrac{\partial f}{\partial z(t)} \end{cases}$$

If we denoted the parameters of the ODE function and the parameters of the initial value z_0 as θ. How do we compute the gradient with respect to these parameters separately? If theta are the parameters of the ODE:

$$\frac{dL}{d\theta} = \int_0^T a(t)^T \frac{\partial f}{\partial \theta} dt$$

If theta are the parameters of the initial value:

$$\frac{dL}{d\theta} = a(0)^T \frac{dz_0}{d\theta}$$

From these special conditions, we built the algorithm to compute the gradient of loss function

---

Algorithm 1: Computing the gradient of loss function

1.(Forward propagation) Solve ODE for $z(t)$ from $t = 0$ to T

$$\begin{cases} \dot{z}(t) = f(z(t), t, \theta) \\ z(0) = z_0(\theta) \end{cases}$$

2.(Backward propagation) Solve an adjoint ODE for $a(t)$ from $t = T$ to 0

$$\begin{cases} a(T) = \dfrac{\partial l}{\partial z(t)} \\ \dot{a}(t)^T = -a(t)^T \dfrac{\partial f}{\partial z(t)} \end{cases}$$

3. Compute the gradient of loss function

$$\frac{dL}{d\theta} = \int_0^T a(t)^T \frac{\partial f}{\partial \theta} dt + a(0)^T \frac{dz_0}{d\theta}$$

Go back to the problem, to optimize L one needs to compute the gradients with respect to its parameters: $z(t_0), t_0, t_1, \theta$. We consider the adjoint variable as the gradient with respect to the solution: $a(t) = \frac{\partial L}{\partial z(t)}$.
Thus, the derivative of loss function follows from the standard chain rule:

$$\frac{dL}{d\theta} = a(0)^T \frac{dz_0}{d\theta} = \frac{\partial L}{\partial z(0)} \frac{dz_0}{d\theta}$$

3

We compute the partial derivative of L wrt. $z(t_0)$

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

To compute the gradients wrt. to t and θ one can think of them as if they were part of the augmented state

$$\frac{d}{dt} \begin{bmatrix} z \\ \theta \\ t \end{bmatrix} (t) = f_{\text{aug}}([z, \theta, t]) := \begin{bmatrix} f([z, \theta, t]) \\ 0 \\ 1 \end{bmatrix}$$

Adjoint state to this augmented state is then

$$a_{\text{aug}} := \begin{bmatrix} a \\ a_\theta \\ a_t \end{bmatrix}, a_\theta(t) := \frac{\partial L}{\partial \theta(t)}, a_t(t) := \frac{\partial L}{\partial t}$$

Gradient of the augmented dynamics

$$\frac{\partial f_{\text{aug}}}{\partial [z, \theta, t]} = \begin{bmatrix} \frac{\partial f}{\partial z} & \frac{\partial f}{\partial \theta} & \frac{\partial f}{\partial t} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Adjoint state ODE is then

$$\frac{da_{\text{aug}}}{dt} = - \begin{bmatrix} a \frac{\partial f}{\partial z} & a \frac{\partial f}{\partial \theta} & a \frac{\partial f}{\partial t} \end{bmatrix}$$

By solving this adjoint augmented ODE initial value problem one gets

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

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

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

which, together with,

$$\frac{\partial L}{\partial t_1} = -a(t) \frac{\partial f(z(t), t, \theta)}{\partial t}$$

complements gradients wrt. all the ODESolve parameters.