A Discrete Formulation of Neural ODE

Wei Dong
Ann Arbor Algorithms

October 19, 2021

1 Introduction

The mathematics of the original Neural ODE paper is densely presented and it might present a challenge to readers of programming background. In Figure 2 of the original paper, it is mentioned that "if the loss depends directly on the state at multi- ple observation times, the adjoint state must be updated in the direction of the partial derivative of the loss with respect to each observation". However, Algorithm 1 only contains a single observation $z(t_1)$ and it is not obvious how such updates should be done.

This article presents a discrete formulation of Section 2 of the original Neural ODE paper. We intentionally broken down mathematical formuation so it is easier to follow and correspondance to programming implementation is explicitly provided.

The rest of the article follows these convention to minimize ambiguity.

- () always means function application. [] and {} are used for grouping in expressons.
- A variable with subscript, e.g. z_i and f_i, means the values at time point t_i. All values takes the form of a column vector of fixed dimension. These variables usually have counterparts in the computer program.
- A symbol without subscript, e.g. z and $f(z(t), t)$, except for the index i of course, always means a time series or a function. These do not have corresponding program variables.
- When we need to store values of partial derivatives, we allocate variables named in greek letters. For example $\gamma_i = \frac{\partial \mathcal{L}}{\partial z_i}$.
- "$\leftarrow$" is used in place of "=" to emphasize variable assignment or value updates.

2 ODE and Numerical Solutions

An ODE has the form:

$$\frac{dz}{dt} = f(z(t), t, \theta) \tag{1}$$

We assume a simplified form of numerical solution using the following equation.

$$\frac{z_{i+1} - z_i}{t_{i+1} - t_i} = F_i, \quad \text{where } F_i \approx f(z_i, t_i, \theta) \tag{2}$$

Note that the slope F_i is related but does not directly equal $f(z_i, t_i)$. Different methods usually differ in how they use the (linear) combination of f at different timepoint to approximate F_i.

In this article we assume the ODE is solved with Euler's method, i.e.

$$F_i \leftarrow f(z_i, t_i, \theta)$$

The numerical solution can therefore be obtained as follows

$$\begin{aligned}
z_1 &\leftarrow \text{known value} \\
F_1 &\leftarrow f(z_1, t_1, \theta) \\
z_2 &\leftarrow z_1 + F_1[t_2 - t_1] \\
&... \\
F_{i-1} &\leftarrow f(z_{i-1}, t_{i-1}, \theta) \\
z_i &\leftarrow z_{i-1} + F_{i-1}[t_i - t_{i-1}] \\
&...
\end{aligned}$$

3 Problem Formulation

Suppose for any $z_1, z_2, ..., z_n$ we can define a loss function

$$\mathcal{L} = L(z_1, z_2, ..., z_n)$$

Our goal is to find a special set of $z_1^*, z_2^*, ..., z_n^*$ and the parameters θ^*, such that:

- $\mathcal{L}$ is minimized by z_i^*.
- z_i^* and θ^* are consistent to (2).

A typical practical scenario is as follows: noisy observations of z_i are made at certain time points but not others, and we want to predict the real value of z_i at all time points. Let w_i be 1 if observation is made at t_i and 0 otherwise.

Let $\hat{z}_i$ be observed values where available and 0 otherwise. We can, for example, define the following loss function:

$$L(z_1, z_2, \cdots, z_n) = \frac{1}{2} \sum_{i=1}^{n} w_i [z_i - \hat{z}_i]^2.$$

In typical deep-learning frameworks optimization problems are solved with gradient-based methods, so our main task is to obtain the gradients $\frac{\partial \mathcal{L}}{\partial \theta}$ and $\frac{\partial \mathcal{L}}{\partial z_i}$ for $i = 1, 2, ..., n$.

4 Gradiant Calculation

We expand the process of ODE solving with (2) as follows.

$$\begin{aligned}
z_2 &= z_1 + F_1 [t_2 - t_1] && F_1 = F(z_1, \theta, \dots) \\
&\cdots\cdots \\
z_i &= z_{i-1} + F_{i-1} [t_i - t_{i-1}] && F_{i-1} = F(z_{i-1}, \theta, \dots) \quad (3) \\
z_{i+1} &= z_i + F_i [t_{i+1} - t_i] && F_i = F(z_i, \theta, \dots) \quad (4) \\
&\cdots\cdots \\
z_n &= z_{n-1} + F_{n-1} [t_n - t_{n-1}] && F_{n-1} = F(z_{n-1}, \theta, \dots)
\end{aligned}$$

According to (4) so we have

$$\frac{\partial z_{i+1}}{\partial z_i} = 1 + \frac{\partial F_i}{\partial z_i} [t_{i+1} - t_i] \quad (5)$$

By applying the chain rule (note that z_i contributes to $\mathcal{L}$ via z_{i+1} as well)

$$\frac{\partial \mathcal{L}}{\partial z_i} = \frac{\partial \mathcal{L}}{\partial z_{i+1}} \frac{\partial z_{i+1}}{\partial z_i} + \frac{\partial L}{\partial z_i} \quad (6)$$

$$\frac{\partial \mathcal{L}}{\partial F_i} = \frac{\partial \mathcal{L}}{\partial z_{i+1}} [t_{i+1} - t_i] \quad (7)$$

Because θ is involved in the calculation of all F_i, we have

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial F_i} \frac{\partial F_i}{\partial \theta} \quad (8)$$

To assist implementation we assign variables to the following partial derivatives.

$$\alpha_i = \frac{\partial \mathcal{L}}{\partial z_i} \qquad \beta_i = \frac{\partial \mathcal{L}}{\partial F_i}$$
$$\gamma_i = \frac{\partial L}{\partial z_i} \qquad \delta_i = \frac{\partial F}{\partial z_i}$$

We can pre-calculate γ_i and δ_i by the formula of L and F at the present value of z_i. We can solve α_i and β_i using the following reverse-ordered updates.

$$
\begin{aligned}
\alpha_n &\leftarrow \gamma_n \\
\\
\alpha_{n-1} &\leftarrow \alpha_n \{1 + \delta_n[t_{n+1} - t_n]\} + \gamma_{n-1} \\
\beta_{n-1} &\leftarrow \alpha_n[t_{n+1} - t_n] \\
&... \\
\alpha_i &\leftarrow \alpha_{i+1} \{1 + \delta_i[t_{i+1} - t_i]\} + \gamma_i \\
\beta_i &\leftarrow \alpha_{i+1}[t_{i+1} - t_i] \\
&...
\end{aligned}
$$

5 Comparison to the Original Paper

By combining (5) and (6) we have

$$\frac{\partial \mathcal{L}}{\partial z_i} = \frac{\partial \mathcal{L}}{\partial z_{i+1}} \left\{ 1 + \frac{\partial F_i}{\partial z_i}[t_{i+1} - t_i] \right\} + \frac{\partial L}{\partial z_i} \tag{9}$$

In the paper L only depends on z_n so we have $\frac{\partial \mathcal{L}}{\partial z_i} = 0$.

By applying this to (9) and substituting the definition of variable $\alpha_i = \frac{\partial \mathcal{L}}{\partial z_i}$ we have

$$\alpha_i = \alpha_{i+1} \left\{ 1 + \frac{\partial F_i}{\partial z_i}[t_{i+1} - t_i] \right\}$$

Slight reorganization gives us

$$\frac{\alpha_{i+1} - \alpha_i}{t_{i+1} - t_i} = -\alpha_{i+1} \frac{\partial F_i}{\partial z_i} \tag{10}$$

The continuous form of (10) is exactly equation (4) in the original paper:

$$\frac{d\alpha}{dt} = -\alpha \frac{\partial f}{\partial z}$$

Similarly, (refeq:theta) can be written as

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{i=1}^{n} \beta_i \frac{\partial F_i}{\partial \theta} \\
&= \sum_{i=1}^{n} \alpha_{i+1}[t_{i+1} - t_i] \frac{\partial F_i}{\partial \theta}
\end{aligned}
$$

Or

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{n} \alpha_{i+1} \frac{\partial F_i}{\partial \theta} [t_{i+1} - t_i] \tag{11}$$

The continuous form of (11) is equation (5) in the origninal paper.

$$\frac{\partial \mathcal{L}}{\partial \theta} = \int_{t_1}^{t_n} \alpha \frac{\partial f}{\partial \theta} dt$$

Note that in the original paper the integration goes from t_1 to t_0 and hence the negative sign.

The above corresondance to the original paper confirms that in Section 2 the mathematics actually does not handle multiple observations. Neither is it obvious whether adding multiple observation will admit a mathematical formulation that is still concise.