Vectorized Adjoint Sensitivity Method for Graph Convolutional Neural Ordinary Differential Equations

Jack Cai[1]

[1]University of Toronto, Division of Engineering Science

August 9, 2021

Abstract

This document, as the title stated, is meant to provide a vectorized implementation of adjoint dynamics calculation for Graph Convolutional Neural Ordinary Differential Equations (GCDE). The adjoint sensitivity method is the gradient approximation method for neural ODEs that replaces the back propagation. When implemented on libraries such as PyTorch or Tensorflow, the adjoint can be calculated by autograd functions without the need for a hand-derived formula. In applications such as edge computing and in memristor crossbars, however, autograds are not available, and therefore we need a vectorized derivation of adjoint dynamics to efficiently map the system on hardware. This document will go over the basics, then move on to derive the vectorized adjoint dynamics for GCDE.

1 Introduction and Preliminaries

Neural Ordinary Differential Equations (ODE) is a class of residual neural networks that frames the layer-wise propagation of hidden states into an initial value problem using differential equation solvers.

$$h_{t+1} = h_t + f(h_t, \theta_t) \tag{1}$$

A typical residual network may have layer wise transformation of hidden states looks like Equation (1). If we add more layers and take smaller time step, then the entire problem can be framed as an ODE defined by Equation (2), and a solution at time t_1 defined by Equation (3).

$$\frac{dh(t)}{dt} = f(h(t), t, \theta) \tag{2}$$

$$h(t_1) = h(t_0) + \int_{t_0}^{t_1} f(h(t), t, \theta)dt \tag{3}$$

Compared to conventional deep neural networks (DNNs), neural ODE uses less parameters applied more times to the hidden states, thereby achieves the depth of DNN while having higher memory efficiency.

1.1 Adjoint sensitivity method

The adjoint sensitivity method is used for automatic differentiation for neural ODE. Unlike traditional backpropagation, a quantity named adjoint is calculated for each t and another ODE solver is used to integrate the overall gradient. Let $L()$ be the scaler loss function for the neural ODE, the adjoint is defined as $a(t) = \frac{\partial L}{\partial h(t)}$, and its dynamic is given by Equation (4):

$$\frac{da(t)}{dt} = -a(t)^T \frac{\partial f(h(t), t, \theta)}{\partial h} \tag{4}$$

The adjoint at each instant is calculated by a backward ODE solver, and the overall gradient of the parameters $\frac{\partial L}{\partial \theta}$ is given by Equation (5), which is integrated by another ODE solver:

$$\frac{\partial L}{\partial \theta} = -\int_{t_1}^{t_0} a(t)^T \frac{\partial f(h(t), t, \theta)}{\partial \theta} dt \tag{5}$$

The proof and intuition behind adjoint sensitivity method is presented in the original paper. For the scope of this paper, we are only focused on how Equation (4) and (5) can be vectorized for GCDE.

1.2 GCDE

GCDE, in short, is the neural ODE version of the state of the art graph learning method Graph Convolutional Neural Network (GCN). For GCN and GCDE, hidden state is represented by a matrix H of dimension $\mathbb{R}^{N\times C}$, representing N nodes with C features. Each layer of GCN or each step of GCDE involves node-wise exchange of information and convolution on nodes. The dynamic of GCDE (which is the same as the layer-wise propagation of GCN) is given by Equation (6), where $A \in \mathbb{R}^{N\times N}$ is symmetric that represents the graph topology and $W \in \mathbb{R}^{C\times C}$ represents the convolution filters:

$$\frac{dH(t)}{dt} = f(H(t), A, W) = ReLU(AH(t)W) \tag{6}$$

To calculate the adjoint, we need to calculate calculate the Jacobian of the function $f(H(t), A, W)$, namely $\frac{\partial f(H(t),A,W)}{\partial H(t)}$ and $\frac{\partial f(H(t),A,W)}{\partial W}$. This poses us a challenge as we will have to unroll the matrix into vectors. While this can be done easily with built-in autograd functions in PyTorch and Tensorflow, it becomes messy when we write it out to obtain an analytical vectorized solution.

1.3 Matrix conventions

Before we delve into deriving the adjoint dynamics, I would like to discuss some convention used in this document. All matrices are represented by capital letters, all vectors are represented by bold lower case letter, all entries within a matrix are represented by the *i,j* subscripts which denotes the i_{th} and j_{th} entry of the matrix, and all entries within a vector are represented by a single subscript denoting the index. For example, A is a matrix, $\mathbf{b}$ is a vector, $a_{1,2}$ denotes the $1,2$ entry of A, and b_2 denotes the second entry of $\mathbf{b}$. We use " : " to denote the entire indices along a row or column, so that we represent the i_{th} row of A as $\mathbf{a}_{i,:}$, and we represent j_{th} column of A as $\mathbf{a}_{:,j}$.

1.4 Multivariable calculus conventions

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ where it takes in a vector $\mathbf{x} \in \mathbb{R}^n$ and output a vector $\mathbf{y} \in \mathbb{R}^m$, i.e.:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}) \tag{7}$$

Then the Jacobian matrix $J_f \in \mathbb{R}^{m \times n}$ is defined as:

$$J_f = f(\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}) \tag{8}$$

Following this convention, let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$, then $g \circ f : \mathbb{R}^n \rightarrow \mathbb{R}^p$. The chain rule is defined as:

$$J_{g \circ f} = J_g \cdot J_f \tag{9}$$

1.5 Partial derivatives of a matrix

Finding the Jacobian for GCDE is challenging not only because it is a composite function, but also because the input is a matrix, we need to unroll the matrix into vectors to match our matrix Jacobian convention. Let *roll*() and *unroll*() to be such operations:

$$\mathbb{R}^{m \times n} \ni A = roll(\mathbf{a}), \mathbf{a} \in \mathbb{R}^{mn} \tag{10}$$

$$\mathbb{R}^{mn} \ni \mathbf{a} = unroll(A), A \in \mathbb{R}^{m \times n} \tag{11}$$

The example below illustrate how they can be used. Suppose $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times n}$ defined by $f(A) = BA$, where $B \in \mathbb{R}^{p \times m}$. Using Equation (10) and (11), we can convert f into $g : \mathbb{R}^{mn} \rightarrow \mathbb{R}^{pn}$, such that:

$$g(\mathbf{a}) = unroll(B(roll(\mathbf{a}))) \tag{12}$$

As a result, we can obtained a Jacobian matrix $J_g \in \mathbb{R}^{pn \times mn}$, which follows the convention defined in section 1.4. The way we unroll a matrix could be by rows or by columns – it really depends on the situation.

2 Adjoint derivation for GCDE

The key of finding the Jacobian matrix for $f(H(t), A, W)$ is to find the Jacobian matrix for the three-matrix multiplication step, as the derivative for *ReLU* is just an elementwise binary step function. So before we moving on solving the Jacobian matrix for the whole thing, let's consider the general case function $f(A) = XAY : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{m \times q}$ where $X \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{n \times p}$, and $Y \in \mathbb{R}^{p \times q}$. We can think of it as composite function $f(A) = h \circ g(A)$, where:

$$g(A) = XA = B, B \in \mathbb{R}^{m \times p} \tag{13}$$

$$h(B) = BY = C, C \in \mathbb{R}^{m \times q} \tag{14}$$

According to our convention, we need to unroll the matrix into vectors in order to calculate the Jacobian matrix. Let $\widehat{g}$ and $\widehat{h}$ to denote them and $\mathbf{a} = unroll(A)$ and $\mathbf{b} = unroll(B)$:

$$\widehat{g}(\mathbf{a}) = unroll(X(roll(\mathbf{a}))) \tag{15}$$

$$\widehat{h}(\mathbf{b}) = unroll((roll(\mathbf{b}))Y) \tag{16}$$

$$\widehat{f}(\mathbf{a}) = \widehat{h} \circ \widehat{g}(\mathbf{a})$$

Hence, according to chain rule, the Jacobian matrix $J_{\widehat{h} \circ \widehat{g}} = J_{\widehat{h}} \cdot J_{\widehat{g}}$ and we will calculate it step by step below.

2.1 Jacobian matrix $J_{\widehat{g}}$ for $g(A) = XA$

We will start by breaking X into rows and A into columns. According to our convention outlined in section 1.3, Equation (13) becomes:

$$g(A) = XA = \begin{bmatrix} \mathbf{x}_{1,:} \\ \vdots \\ \mathbf{x}_{1,:} \end{bmatrix} \begin{bmatrix} \mathbf{a}_{:,1} & \dots & \mathbf{a}_{:,p} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1} & \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2} & \dots & \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,p} \\ \mathbf{x}_{2,:} \cdot \mathbf{a}_{:,1} & \ddots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{m,:} \cdot \mathbf{a}_{:,1} & \dots & \dots & \mathbf{x}_{m,:} \cdot \mathbf{a}_{:,p} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,p} \\ b_{2,1} & \ddots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ b_{m,1} & \dots & \dots & b_{m,p} \end{bmatrix}$$

Since we used rows of X and columns of A, for our convenience, we will unroll B by rows and A by columns. Therefore our Jacobian matrix $J_{\widehat{g}}$ becomes:

$$J_{\widehat{g}} = \begin{bmatrix} \frac{\partial b_{1,1}}{\partial a_{1,1}} & \frac{\partial b_{1,1}}{\partial a_{2,1}} & \dots & \frac{\partial b_{1,1}}{\partial a_{1,2}} & \dots & \frac{\partial b_{1,1}}{\partial a_{n,p}} \\ \frac{\partial b_{1,2}}{\partial a_{1,1}} & \ddots & \dots & \frac{\partial b_{1,2}}{\partial a_{1,2}} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ \frac{\partial b_{2,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_{m,p}}{\partial a_{1,1}} & \dots & \dots & \dots & \dots & \frac{\partial b_{m,p}}{\partial a_{n,p}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{1,1}} & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{2,1}} & \dots & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{1,2}} & \dots & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{n,p}} \\ \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2}}{\partial a_{1,1}} & \ddots & \dots & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2}}{\partial a_{1,2}} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ \frac{\partial \mathbf{x}_{2,:} \cdot \mathbf{a}_{:,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{x}_{m,:} \cdot \mathbf{a}_{:,m}}{\partial a_{1,1}} & \dots & \dots & \dots & \dots & \frac{\partial \mathbf{x}_{m,:} \cdot \mathbf{a}_{:,p}}{\partial a_{n,p}} \end{bmatrix}$$

Notice that for $\frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{2,1}}$, the partial derivative is non-zero because $\mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}$ depends on the entry $a_{2,1}$; whereas for $\frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2}}{\partial a_{1,1}}$, the partial derivative is zero because $a_{1,1}$ is not an entry inside the column vector $\mathbf{a}_{:,2}$, therefore change in $a_{1,1}$ do not impact $\mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2}$. *The entry must be within the column vector to have a non-zero partial derivative.* Following this pattern, our Jacobian matrix $J_{\widehat{g}}$ becomes:

$$J_{\widehat{g}} = \begin{bmatrix} \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{1,1}} & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,1}}{\partial a_{2,1}} & \dots & 0 & \dots & 0 \\ 0 & \ddots & \dots & \frac{\partial \mathbf{x}_{1,:} \cdot \mathbf{a}_{:,2}}{\partial a_{1,2}} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ \frac{\partial \mathbf{x}_{2,:} \cdot \mathbf{a}_{:,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \frac{\partial \mathbf{x}_{m,:} \cdot \mathbf{a}_{:,p}}{\partial a_{n,p}} \end{bmatrix}$$

Taking the derivative of the dot product, vectorize it and we get the Jacobian matrix $J_{\widehat{g}}$:

$$J_{\widehat{g}} = \begin{bmatrix} x_{1,1} & \dots & x_{1,n} & 0 & \dots & 0 \\ 0 & \ddots & 0 & x_{1,1} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ x_{2,1} & \vdots & x_{2,n} & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & x_{m,n} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} & & & \\ & \mathbf{x}_{2,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{2,:} \\ & \dots & \dots & \\ \mathbf{x}_{m,:} & & & \\ & \mathbf{x}_{m,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{m,:} \end{bmatrix} \in \mathbb{R}^{mp \times np} \tag{17}$$

2.2 Jacobian matrix $J_{\widehat{h}}$ for $h(B) = BY$

Similarly to section 2.1, we will start by breaking B and Y into rows and columns. So Equation (14) becomes:

$$h(B) = BY = \begin{bmatrix} \mathbf{b}_{1,:} \\ \vdots \\ \mathbf{b}_{m,:} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{:,1} & \dots & \mathbf{y}_{:,q} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1} & \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,2} & \dots & \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,p} \\ \mathbf{b}_{2,:} \cdot \mathbf{y}_{:,1} & \ddots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}_{m,:} \cdot \mathbf{y}_{:,1} & \dots & \dots & \mathbf{b}_{m,:} \cdot \mathbf{y}_{:,p} \end{bmatrix} = \begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,p} \\ c_{2,1} & \ddots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & \dots & \dots & c_{m,p} \end{bmatrix}$$

We will unroll B by rows and C by columns to construct our Jacobian matrix:

$$J_{\widehat{h}} = \begin{bmatrix} \frac{\partial c_{1,1}}{\partial b_{1,1}} & \frac{\partial c_{1,1}}{\partial b_{1,2}} & \cdots & \frac{\partial c_{1,1}}{\partial b_{2,1}} & \cdots & \frac{\partial c_{1,1}}{\partial b_{m,p}} \\ \frac{\partial c_{1,2}}{\partial b_{1,1}} & \ddots & \cdots & \frac{\partial c_{1,2}}{\partial b_{2,1}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial c_{2,1}}{\partial b_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_{m,q}}{\partial b_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial c_{m,q}}{\partial b_{m,p}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{1,1}} & \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{1,2}} & \cdots & \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{2,1}} & \cdots & \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{m,p}} \\ \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,2}}{\partial b_{1,1}} & \ddots & \cdots & \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,2}}{\partial b_{2,1}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial \mathbf{b}_{2,:} \cdot \mathbf{y}_{:,1}}{\partial b_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{b}_{m,:} \cdot \mathbf{y}_{:,q}}{\partial b_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial \mathbf{b}_{m,:} \cdot \mathbf{y}_{:,q}}{\partial b_{m,p}} \end{bmatrix}$$

By the pattern that non-pairing entry and row vectors have zero partial derivatives, the Jacobian matrix can be reduced to:

$$J_{\widehat{h}} = \begin{bmatrix} \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{1,1}} & \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,1}}{\partial b_{1,2}} & \cdots & 0 & \cdots & 0 \\ \frac{\partial \mathbf{b}_{1,:} \cdot \mathbf{y}_{:,2}}{\partial b_{1,1}} & \ddots & \cdots & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \frac{\partial \mathbf{b}_{m,:} \cdot \mathbf{y}_{:,q}}{\partial b_{m,p}} \end{bmatrix}$$

Taking the partial derivatives of the dot product:

$$J_{\widehat{h}} = \begin{bmatrix} y_{1,1} & y_{2,1} & \cdots & y_{p,1} & 0 & \cdots & 0 \\ y_{1,2} & \vdots & \ddots & y_{p,2} & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots & \vdots \\ y_{1,q} & y_{2,q} & \vdots & y_{p,q} & 0 & \cdots & \vdots \\ 0 & 0 & \vdots & 0 & y_{1,1} & \cdots & \vdots \\ 0 & 0 & \vdots & 0 & y_{1,2} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & 0 & y_{1,q} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & y_{p,q} \end{bmatrix}$$

Notice that the diagonal $q \times p$ blocks are Y^T, therefore:

$$J_{\widehat{h}} = \begin{bmatrix} Y^T & & & & \\ & Y^T & & & \\ & & Y^T & & \\ & & & \ddots & \\ & & & & Y^T \end{bmatrix} \in \mathbb{R}^{mq \times mp} \tag{18}$$

2.3 Calculating $J_{\widehat{h} \circ \widehat{g}}$ from the chain rule

By the chain rule, we calculate $J_{\widehat{h} \circ \widehat{g}} \in \mathbb{R}^{mq \times np}$ by $J_{\widehat{h}} \cdot J_{\widehat{g}}$:

$$J_{\widehat{h}} \cdot J_{\widehat{g}} = \begin{bmatrix} Y^T & & & & \\ & Y^T & & & \\ & & Y^T & & \\ & & & \ddots & \\ & & & & Y^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \\ & \cdots & \cdots & \\ \mathbf{x}_{m,:} & & & \\ & \mathbf{x}_{m,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{m,:} \end{bmatrix}$$

To simplify the matrix multiplication, we break done Y^T to (note that $\mathbf{y}_{i,:}^T \in \mathbb{R}^{q \times 1}$ is the transpose of the first row of Y):

$$Y^T = \begin{bmatrix} \mathbf{y}_{1,:}^T & \mathbf{y}_{2,:}^T & \cdots & \mathbf{y}_{p,:}^T \end{bmatrix}$$

And we break the matrix multiplication blocks by blocks:

$$J_{\widehat{h}} \cdot J_{\widehat{g}} = \begin{bmatrix} \begin{bmatrix} Y^T & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \cdot J_{\widehat{g}} \\ \begin{bmatrix} \mathbf{0} & Y^T & \cdots & \mathbf{0} \end{bmatrix} \cdot J_{\widehat{g}} \\ \cdots \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & Y^T \end{bmatrix} \cdot J_{\widehat{g}} \end{bmatrix}$$

Let's look at the first block:

$$\begin{bmatrix} Y^T & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \cdot J_{\widehat{g}} = \begin{bmatrix} Y^T & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \\ & \dots & \dots & \\ \mathbf{x}_{m,:} & & & \\ & \mathbf{x}_{m,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{m,:} \end{bmatrix} \tag{19}$$

Note that since $Y^T \in q \times p$, only the first p rows of $J_{\widehat{g}}$ matters. Hence Equation (19) becomes:

$$\begin{aligned} & Y^T \cdot \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{y}_{1,:}^T & \mathbf{y}_{2,:}^T & \cdots & \mathbf{y}_{p,:}^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{1,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{1,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{1,:} \end{bmatrix} \in \mathbb{R}^{q \times pn} \end{aligned}$$

If we repeat for other blocks, we would get the Jacobian matrix for $J_{\widehat{h} \circ \widehat{g}}$ to be:

$$J_{\widehat{h} \circ \widehat{g}} = \begin{bmatrix} \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{1,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{1,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{1,:} \\ \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{2,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{2,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{2,:} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{m,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{m,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{m,:} \end{bmatrix} \in \mathbb{R}^{qm \times pn} \tag{20}$$

It is also important to keep in mind what each entry within $J_{\widehat{h} \circ \widehat{g}}$ mean:

$$\begin{aligned} J_{\widehat{h} \circ \widehat{g}} = J_{\widehat{h}} \cdot J_{\widehat{g}} &= \begin{bmatrix} \frac{\partial c_{1,1}}{\partial b_{1,1}} & \frac{\partial c_{1,1}}{\partial b_{1,2}} & \cdots & \frac{\partial c_{1,1}}{\partial b_{2,1}} & \cdots & \frac{\partial c_{1,1}}{\partial b_{m,p}} \\ \frac{\partial c_{1,2}}{\partial b_{1,1}} & \ddots & \cdots & \frac{\partial c_{1,2}}{\partial b_{2,1}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial c_{2,1}}{\partial b_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_{m,q}}{\partial b_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial c_{m,q}}{\partial b_{m,p}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial b_{1,1}}{\partial a_{1,1}} & \frac{\partial b_{1,1}}{\partial a_{2,1}} & \cdots & \frac{\partial b_{1,1}}{\partial a_{1,2}} & \cdots & \frac{\partial b_{1,1}}{\partial a_{n,p}} \\ \frac{\partial b_{1,2}}{\partial a_{1,1}} & \ddots & \cdots & \frac{\partial b_{1,2}}{\partial a_{1,2}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial b_{2,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_{m,p}}{\partial a_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial b_{m,p}}{\partial a_{n,p}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial c_{1,1}}{\partial a_{1,1}} & \frac{\partial c_{1,1}}{\partial a_{2,1}} & \cdots & \frac{\partial c_{1,1}}{\partial a_{1,2}} & \cdots & \frac{\partial c_{1,1}}{\partial a_{n,p}} \\ \frac{\partial c_{1,2}}{\partial a_{1,1}} & \ddots & \cdots & \frac{\partial c_{1,2}}{\partial a_{1,2}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial c_{2,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_{m,q}}{\partial a_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial c_{m,q}}{\partial a_{n,p}} \end{bmatrix} \end{aligned} \tag{21}$$

2.4 Adjoint calculation

Now we have successfully found the general solution of Jacobian matrix for $f(A) = XAY$ and $g(A) = XA$ (or more precisely, the unrolled version of them, namely $f(\mathbf{a})$ and $\widehat{g}(\mathbf{a})$, and it is time to apply them in the adjoint calculation for GCDE, as they are the special cases of the general solutions outlined in section 2.1-2.3. Back in section 1.2 I stated that we need to find the Jacobian matrix $\frac{\partial f(H(t),A,W)}{\partial H(t)}$ and $\frac{\partial f(H(t),A,W)}{\partial W}$; well that does not actually make sense in our convention since (1) $H(t)$ is not a vector and (2) the adjoint must be a vector, which means $\frac{\partial L}{\partial H(t)}$ should also be unrolled. Therefore, a GCDE version of adjoint dynamics for Equation (4) and (5) is:

$$-a(t)^T \frac{\partial f(h(t),t,\theta)}{\partial h} \to -\frac{\partial L}{\partial \mathbf{h}(t)}^T \frac{\partial \widehat{f}(\mathbf{h}(t),A,W)}{\partial \mathbf{h}(t)} \tag{22}$$

$$-a(t)^T \frac{\partial f(h(t),t,\theta)}{\partial \theta} \to -\frac{\partial L}{\partial \mathbf{h}(t)}^T \frac{\partial \widehat{f}(\mathbf{h}(t),A,W)}{\partial \mathbf{w}} \tag{23}$$

where $\mathbf{h}(t) = unroll(H(t))$, $\mathbf{w} = unroll(W)$, and $\widehat{f}(\mathbf{h}(t),A,W) = unroll(f(roll(\mathbf{h}(t)),A,W))$. However, this does not stop us to find a vectorized equivalence for Equation (4) and (5) for GCDE implementation on hardware, as keeping the matrices unrolled greatly increases the dimensions and could not take the parallel in-memory computing advantages brought by memristor crossbars.

2.5 Vectorized Equation (22)

We start the vectorization from the general case $f(A) = XAY$ we discussed in the earlier sections. Let $\mathbb{R}^{m\times q} \ni C = f(A)$, and $\mathbb{R}^{mq} \ni \mathbf{c} = \mathit{unroll}(C)$, we unroll C row by row so that its partial derivatives with respect to a scalar loss function L has the form:

$$\frac{\partial L}{\partial \mathbf{c}} = \begin{bmatrix} \frac{\partial L}{\partial c_{1,1}} \\ \vdots \\ \frac{\partial L}{\partial c_{1,q}} \\ \frac{\partial L}{\partial c_{2,1}} \\ \vdots \\ \frac{\partial L}{\partial c_{m,q}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}}^T \\ \frac{\partial L}{\partial \mathbf{c}_{2,:}}^T \\ \vdots \\ \frac{\partial L}{\partial \mathbf{c}_{m,:}}^T \end{bmatrix} \in \mathbb{R}^{mq}, \frac{\partial L}{\partial \mathbf{c}_{i,:}} \in \mathbb{R}^{1\times q}$$

We want to find (a reminder that $\widehat{f}(\mathbf{a})$ is a unrolled version of $f(A)$):

$$-\frac{\partial L}{\partial \mathbf{c}}^T \frac{\partial \widehat{f}(\mathbf{a})}{\partial \mathbf{a}} = -\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{f}} = -\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h}\circ\widehat{g}}$$

Expand and plug in Equation (20):

$$-\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h}\circ\widehat{g}} = -\begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} & \frac{\partial L}{\partial \mathbf{c}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{1,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{1,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{1,:} \\ \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{2,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{2,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{2,:} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{m,:} & \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{m,:} & \cdots & \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{m,:} \end{bmatrix}$$

Since $\frac{\partial L}{\partial \mathbf{c}_{i,:}} \in \mathbb{R}^{1\times q}$ and $\mathbf{y}_{i,:}^T \cdot \mathbf{x}_{j,:} \in \mathbb{R}^{q\times n}$, the blocks defined above have matching dimensions. Hence we can take their dot product directly:

$$-\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h}\circ\widehat{g}} = -\begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{1,:} + \ldots + \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{m,:} & \cdots & \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{1,:} + \ldots + \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{m,:} \end{bmatrix}$$

Note that this dot product is a long $1 \times np$ vector, and we can simplify it by rolling it into a $p \times n$ matrix:

$$-\mathit{roll}(\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h}\circ\widehat{g}}) = -\begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{1,:} + \ldots + \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \cdot \mathbf{x}_{m,:} \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{1,:} + \ldots + \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{2,:}^T \cdot \mathbf{x}_{m,:} \\ \vdots \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{1,:} + \ldots + \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{p,:}^T \cdot \mathbf{x}_{m,:} \end{bmatrix} \tag{24}$$

Equation (24) can be further simplified:

$$-\mathit{roll}(\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h}\circ\widehat{g}}) = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{2,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{2,:}^T \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{p,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{p,:}^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{m,:} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{2,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{2,:}^T \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{p,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{p,:}^T \end{bmatrix} \cdot X$$

Let:

$$E = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{2,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{2,:}^T \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{p,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{p,:}^T \end{bmatrix}$$

We can simplify E further by looking at it row by row. The first row of E is:

$$\mathbf{e}_{1,:} = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} \cdot \mathbf{y}_{1,:}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \cdot \mathbf{y}_{1,:}^T \end{bmatrix} \tag{25}$$

Notice that each entry within Equation (25) is a scalar because $\frac{\partial L}{\partial \mathbf{c}_{i,:}} \in \mathbb{R}^{1\times q}$ and $\mathbf{y}_{1,:}^T \in \mathbb{R}^{q\times 1}$. Therefore, we can manipulate Equation (25) such that:

$$\mathbf{e}_{1,:} = (\mathbf{y}_{1,:}^T)^T \cdot \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}}^T & \frac{\partial L}{\partial \mathbf{c}_{2,:}}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}}^T \end{bmatrix} = \mathbf{y}_{1,:} \cdot \begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}}^T & \frac{\partial L}{\partial \mathbf{c}_{2,:}}^T & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}}^T \end{bmatrix}$$

We define $\frac{\partial L}{\partial C}$ having matching entries to C:

$$\frac{\partial L}{\partial C} = \begin{bmatrix} \frac{\partial L}{\partial c_{1,1}} & \frac{\partial L}{\partial c_{1,2}} & \cdots & \frac{\partial L}{\partial c_{1,q}} \\ \frac{\partial L}{\partial c_{2,1}} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial c_{m,1}} & \cdots & \cdots & \frac{\partial L}{\partial c_{m,q}} \end{bmatrix}$$

Therefore:

$$\mathbf{e}_{1,:} = \mathbf{y}_{1,:} \cdot \frac{\partial L}{\partial C}^T$$

If we repeat for all rows of E, then:

$$E = \begin{bmatrix} \mathbf{y}_{1,:} \cdot \frac{\partial L}{\partial C}^T \\ \mathbf{y}_{2,:} \cdot \frac{\partial L}{\partial C}^T \\ \vdots \\ \mathbf{y}_{p,:} \cdot \frac{\partial L}{\partial C}^T \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{1,:} \\ \mathbf{y}_{2,:} \\ \vdots \\ \mathbf{y}_{p,:} \end{bmatrix} \cdot \frac{\partial L}{\partial C}^T = Y \cdot \frac{\partial L}{\partial C}^T$$

As a result:

$$-\,roll(\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h} \circ \widehat{g}}) = -E \cdot X = -Y \cdot \frac{\partial L}{\partial C}^T \cdot X \tag{26}$$

We need to keep in mind what this matrix actually represents. So we go back to Equation (21):

$$-roll(\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h} \circ \widehat{g}}) = -roll(\begin{bmatrix} \frac{\partial L}{\partial \mathbf{c}_{1,:}} & \frac{\partial L}{\partial \mathbf{c}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{c}_{m,:}} \end{bmatrix} \cdot = \begin{bmatrix} \frac{\partial c_{1,1}}{\partial a_{1,1}} & \frac{\partial c_{1,1}}{\partial a_{2,1}} & \cdots & \frac{\partial c_{1,1}}{\partial a_{1,2}} & \cdots & \frac{\partial c_{1,1}}{\partial a_{n,p}} \\ \frac{\partial c_{1,2}}{\partial a_{1,1}} & \ddots & \cdots & \frac{\partial c_{1,2}}{\partial a_{1,2}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial c_{2,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_{m,q}}{\partial a_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial c_{m,q}}{\partial a_{n,p}} \end{bmatrix})$$

$$= -roll(\begin{bmatrix} \frac{\partial L}{\partial a_{1,1}} & \frac{\partial L}{\partial a_{2,1}} & \cdots & \frac{\partial L}{\partial a_{n,1}} & \frac{\partial L}{\partial a_{1,2}} & \cdots & \frac{\partial L}{\partial a_{n,p}} \end{bmatrix})$$

$$= -\begin{bmatrix} \frac{\partial L}{\partial a_{1,1}} & \frac{\partial L}{\partial a_{2,1}} & \cdots & \frac{\partial L}{\partial a_{n,1}} \\ \frac{\partial L}{\partial a_{1,2}} & \frac{\partial L}{\partial a_{2,2}} & \cdots & \frac{\partial L}{\partial a_{n,2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial a_{1,p}} & \frac{\partial L}{\partial a_{2,p}} & \cdots & \frac{\partial L}{\partial a_{n,p}} \end{bmatrix} = -\begin{bmatrix} - \frac{\partial L}{\partial \mathbf{a}_{:,1}}^T - \\ - \frac{\partial L}{\partial \mathbf{a}_{:,2}}^T - \\ \vdots \\ - \frac{\partial L}{\partial \mathbf{a}_{:,p}}^T - \end{bmatrix} \tag{27}$$

Notice that the entries of this matrix matches the transpose of A. Hence, in order to match the entries of A, we will take Equation (27)'s transpose – and here we found the general vectorized adjoint solution for $f(A)$:

$$vectorized(-\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h} \circ \widehat{g}}) = -\begin{bmatrix} - \frac{\partial L}{\partial \mathbf{a}_{:,1}}^T - \\ - \frac{\partial L}{\partial \mathbf{a}_{:,2}}^T - \\ \vdots \\ - \frac{\partial L}{\partial \mathbf{a}_{:,p}}^T - \end{bmatrix}^T = -(Y \cdot \frac{\partial L}{\partial C}^T \cdot X)^T$$

$$vectorized(-\frac{\partial L}{\partial \mathbf{c}}^T J_{\widehat{h} \circ \widehat{g}}) = -X^T \cdot \frac{\partial L}{\partial C} \cdot Y^T \tag{28}$$

The adjoint dynamics shown in Equation (22) for GCDE is just a special of Equation (28). The derivative of the $ReLU()$ activation function is an element-wise binary step function:

$$step(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{29}$$

Hence, let $\odot$ denote the Hadamard product between matrices, by chain rule and Equation (28), the vectorized Equation (22) is:

$$vectorized(-\frac{\partial L}{\partial \mathbf{h}(t)}^T \frac{\partial \widehat{f}(\mathbf{h}(t), A, W)}{\partial \mathbf{h}(t)}) = -A^T \cdot (\frac{\partial L}{\partial H(t)} \odot step(H(t))) \cdot W^T$$

where:

$$\frac{\partial L}{\partial H(t)} = roll(\frac{\partial L}{\partial \mathbf{h}(t)}) = \begin{bmatrix} \frac{\partial L}{\partial h_{1,1}} & \frac{\partial L}{\partial h_{1,2}} & \cdots & \frac{\partial L}{\partial h_{1,q}} \\ \frac{\partial L}{\partial h_{2,1}} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial h_{m,1}} & \cdots & \cdots & \frac{\partial L}{\partial h_{m,q}} \end{bmatrix}$$

Since by definition of GCDE, the graph topology matrix A is always symmetric, the final vectorized Equation (22) is:

$$vectorized(-\frac{\partial L}{\partial \mathbf{h}(t)}^T \frac{\partial \widehat{f}(\mathbf{h}(t), A, W)}{\partial \mathbf{h}(t)}) = -A \cdot (\frac{\partial L}{\partial H(t)} \odot step(H(t))) \cdot W^T \tag{30}$$

2.6 Vectorized Equation (23)

We will again vectorize Equation (23) from a general case, $g(A) = XA$. This is because we can treat $AH(t)$ in Equation (6) as a single matrix that linearly transforms W. Let $\mathbb{R}^{m \times p} \ni B = g(A)$, and $\mathbb{R}^{mp} \ni \mathbf{b} = unroll(B)$, we unroll B row by row so that its partial derivatives with respect to a scalar loss function L has the form:

$$\frac{\partial L}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial L}{\partial b_{1,1}} \\ \vdots \\ \frac{\partial L}{\partial b_{1,p}} \\ \frac{\partial L}{\partial b_{2,1}} \\ \vdots \\ \frac{\partial L}{\partial b_{m,p}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{b}_{1,:}}^T \\ \frac{\partial L}{\partial \mathbf{b}_{2,:}}^T \\ \vdots \\ \frac{\partial L}{\partial \mathbf{b}_{m,:}}^T \end{bmatrix} \in \mathbb{R}^{mp}, \frac{\partial L}{\partial \mathbf{b}_{i,:}} \in \mathbb{R}^{1\times p}$$

In addition, we define

$$\frac{\partial L}{\partial B} = \begin{bmatrix} \frac{\partial L}{\partial b_{1,1}} & \frac{\partial L}{\partial b_{1,2}} & \cdots & \frac{\partial L}{\partial b_{1,q}} \\ \frac{\partial L}{\partial b_{2,1}} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial b_{m,1}} & \cdots & \cdots & \frac{\partial L}{\partial b_{m,q}} \end{bmatrix}$$

Since we are considering the general case $g(A) = XA$ and we are given an arbitury adjoint $\mathbf{b}$, we can change Equation (17) to:

$$-\frac{\partial L}{\partial \mathbf{h}(t)}^T \frac{\partial \widehat{f}(\mathbf{h}(t), A, W)}{\partial \mathbf{w}} \to -\frac{\partial L}{\partial \mathbf{b}}^T \frac{\partial \widehat{g}(\mathbf{a})}{\partial \mathbf{a}} = -\frac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}}$$

Expand and plug in Equation (17):

$$-\frac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}} = -\begin{bmatrix}\frac{\partial L}{\partial \mathbf{b}_{1,:}} & \frac{\partial L}{\partial \mathbf{b}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{b}_{m,:}}\end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_{1,:} & & & \\ & \mathbf{x}_{1,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{1,:} \\ & \cdots & \cdots & \\ \mathbf{x}_{m,:} & & & \\ & \mathbf{x}_{m,:} & & \\ & & \ddots & \\ & & & \mathbf{x}_{m,:} \end{bmatrix}$$

We again break it into blocks:

$$-\frac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}} = -\begin{bmatrix} \begin{bmatrix}\frac{\partial L}{\partial \mathbf{b}_{1,:}} & \frac{\partial L}{\partial \mathbf{b}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{b}_{m,:}}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{x}_{1,:} \\ \vdots \\ \mathbf{0} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{m,:} \\ \vdots \\ \mathbf{0}\end{bmatrix} \cdots \begin{bmatrix}\frac{\partial L}{\partial \mathbf{b}_{1,:}} & \frac{\partial L}{\partial \mathbf{b}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{b}_{m,:}}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{0} \\ \vdots \\ \mathbf{x}_{1,:} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \vdots \\ \mathbf{x}_{m,:}\end{bmatrix} \end{bmatrix}$$

If we expand the first block:

$$\begin{bmatrix}\frac{\partial L}{\partial \mathbf{b}_{1,:}} & \frac{\partial L}{\partial \mathbf{b}_{2,:}} & \cdots & \frac{\partial L}{\partial \mathbf{b}_{m,:}}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{x}_{1,:} \\ \vdots \\ \mathbf{0} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{m,:} \\ \vdots \\ \mathbf{0}\end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial b_{1,1}} \cdot x_{1,1} + \frac{\partial L}{\partial b_{2,1}} \cdot x_{2,1} + \ldots + \frac{\partial L}{\partial b_{m,1}} \cdot x_{m,1} \\ \frac{\partial L}{\partial b_{1,1}} \cdot x_{1,2} + \frac{\partial L}{\partial b_{2,1}} \cdot x_{2,2} + \ldots + \frac{\partial L}{\partial b_{m,1}} \cdot x_{m,2} \\ \vdots \\ \frac{\partial L}{\partial b_{1,1}} \cdot x_{1,p} + \frac{\partial L}{\partial b_{2,1}} \cdot x_{2,p} + \ldots + \frac{\partial L}{\partial b_{m,1}} \cdot x_{m,p} \end{bmatrix}^T$$

$$= (\begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{m,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,p} & x_{2,p} & \cdots & x_{m,p} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial L}{\partial b_{1,1}} \\ \frac{\partial L}{\partial b_{2,1}} \\ \vdots \\ \frac{\partial L}{\partial b_{m,1}} \end{bmatrix})^T = (X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,1}})^T$$

If we repeat for other blocks, we would get:

$$-\frac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}} = \begin{bmatrix} X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,1}} \\ X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,2}} \\ \vdots \\ X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,p}} \end{bmatrix}^T \quad \text{or } (-\frac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}})^T = \begin{bmatrix} X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,1}} \\ X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,2}} \\ \vdots \\ X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,p}} \end{bmatrix}$$

We will make it vectorized by rolling it into a $n \times p$ matrix:

$$roll((-\tfrac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}})^T) = -\begin{bmatrix} X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,1}} & X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,2}} & \cdots & X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,p}} \end{bmatrix}$$

Further vectorization could be done:

$$-\begin{bmatrix} X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,1}} & X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,2}} & \cdots & X^T \cdot \frac{\partial L}{\partial \mathbf{b}_{:,p}} \end{bmatrix} = -X^T \cdot \begin{bmatrix} \frac{\partial L}{\partial \mathbf{b}_{:,1}} & \frac{\partial L}{\partial \mathbf{b}_{:,2}} & \cdots & \frac{\partial L}{\partial \mathbf{b}_{:,p}} \end{bmatrix}$$

$$= -X^T \cdot \frac{\partial L}{\partial B} \tag{31}$$

Of course, we need to keep in mind what $-X^T \cdot \frac{\partial L}{\partial B}$ actually represents:

$$roll((-\tfrac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}})^T) = -(\begin{bmatrix} \frac{\partial L}{\partial b_{1,1}} & \cdots & \frac{\partial L}{\partial b_{1,p}} & \frac{\partial L}{\partial b_{2,1}} & \cdots & \frac{\partial L}{\partial b_{m,p}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial b_{1,1}}{\partial a_{1,1}} & \frac{\partial b_{1,1}}{\partial a_{2,1}} & \cdots & \frac{\partial b_{1,1}}{\partial a_{1,2}} & \cdots & \frac{\partial b_{1,1}}{\partial a_{n,p}} \\ \frac{\partial b_{1,2}}{\partial a_{1,1}} & \ddots & \cdots & \frac{\partial b_{1,2}}{\partial a_{1,2}} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial b_{2,1}}{\partial a_{1,1}} & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_{m,p}}{\partial a_{1,1}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial b_{m,p}}{\partial a_{n,p}} \end{bmatrix})^T$$

$$= -roll(\begin{bmatrix} \frac{\partial L}{\partial a_{1,1}} & \frac{\partial L}{\partial a_{2,1}} & \cdots & \frac{\partial L}{\partial a_{n,1}} & \frac{\partial L}{\partial a_{1,2}} & \cdots & \frac{\partial L}{\partial a_{n,p}} \end{bmatrix}^T)$$

$$= -\begin{bmatrix} \frac{\partial L}{\partial a_{1,1}} & \frac{\partial L}{\partial a_{1,2}} & \cdots & \frac{\partial L}{\partial a_{1,p}} \\ \frac{\partial L}{\partial a_{2,1}} & \frac{\partial L}{\partial a_{2,2}} & \cdots & \frac{\partial L}{\partial a_{2,p}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial a_{n,1}} & \frac{\partial L}{\partial a_{n,2}} & \cdots & \frac{\partial L}{\partial a_{n,p}} \end{bmatrix} \tag{32}$$

$$= -X^T \cdot \tfrac{\partial L}{\partial B}$$

Since the entries of $-X^T \cdot \frac{\partial L}{\partial B}$ matches A, Equation (31) is a good vectorized solution. Therefore:

$$vectorized(-\tfrac{\partial L}{\partial \mathbf{b}}^T J_{\widehat{g}}) = -X^T \cdot \tfrac{\partial L}{\partial B}$$

Now if we plug Equation (31) into our GCDE special case, and if we define:

$$\tfrac{\partial L}{\partial W} = roll(\tfrac{\partial L}{\partial \mathbf{w}}) = \begin{bmatrix} \frac{\partial L}{\partial w_{1,1}} & \frac{\partial L}{\partial w_{1,2}} & \cdots & \frac{\partial L}{\partial w_{1,q}} \\ \frac{\partial L}{\partial w_{2,1}} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial w_{m,1}} & \cdots & \cdots & \frac{\partial L}{\partial w_{m,q}} \end{bmatrix}$$

Then the vectorized Equation (23) is:

$$vectorized(-\frac{\partial L}{\partial \mathbf{w}}^T \frac{\partial \widehat{f}(\mathbf{h}(t), A, W)}{\partial \mathbf{w}}) = -(AH(t))^T \cdot (\frac{\partial L}{\partial W} \odot step(H(t))) \tag{33}$$

2.7 Summary

In conclusion, we have found the vectorized adjoint dynamics for GCDE. Using the vectorized adjoint dynamics, we do not need to unroll the hidden states and parameters, and we could take the advantages of the in-memory matrices programmed on memristor crossbar.

Acknowledgement

I thank Louis Primeau for his thorough tutorial on multivariable calculus and matrix unrolling for Jacobian calculation. I thank Nafiseh Ghoroghchian for her guide on matrix differentiation and proofread of the work.

References

[1] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, & D. Duvenaud, ‘Neural Ordinary Differential Equations’. arXiv, 2018. https://arxiv.org/pdf/1806.07366.pdf
[2] T. N. Kipf & M. Welling, ‘Semi-Supervised Classification with Graph Convolutional Networks’. arXiv, 2016. https://arxiv.org/pdf/1911.07532.pdf
[3] M. Poli, S. Massaroli, J. Park, A. Yamashita, H. Asama, & J. Park, ‘Graph Neural Ordinary Differential Equations’. arXiv, 2019. https://arxiv.org/abs/1911.07532
[4] Louis Primeau, Neural ODE for Memristor Crossbar.