

ECE59500NL Lecture 28: Automatic Differentiation—III

Jeffrey Mark Siskind

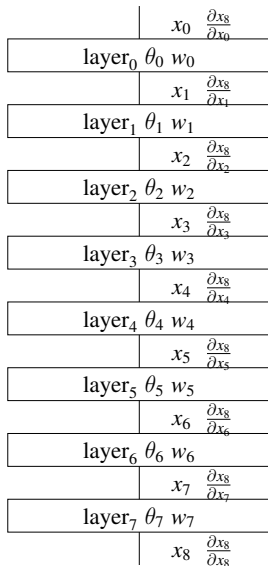
School of Electrical and Computer Engineering

Spring 2021

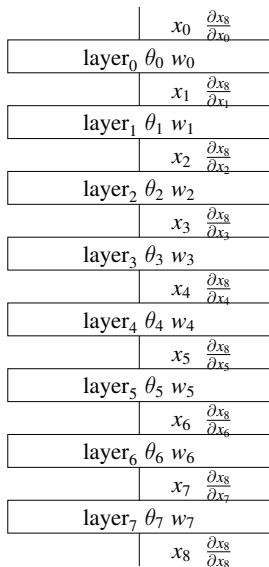


© 2021 Jeffrey Mark Siskind. All rights reserved.

A Neural Network



A Neural Network is a (Functional) Program



```

net  $[\theta_0, \dots, \theta_7] [w_0, \dots, w_7] x_0 \triangleq$ 
  let  $x_1 = \text{layer}_0 \theta_0 w_0 x_0$ 
       $x_2 = \text{layer}_1 \theta_1 w_1 x_1$ 
       $x_3 = \text{layer}_2 \theta_2 w_2 x_2$ 
       $x_4 = \text{layer}_3 \theta_3 w_3 x_3$ 
       $x_5 = \text{layer}_4 \theta_4 w_4 x_4$ 
       $x_6 = \text{layer}_5 \theta_5 w_5 x_5$ 
       $x_7 = \text{layer}_6 \theta_6 w_6 x_6$ 
       $x_8 = \text{layer}_7 \theta_7 w_7 x_7$ 
  in  $x_8$ 
    
```

A (Functional) Program

$$f [w_0, w_1] [x_0, x_1] \triangleq$$

let t_0 = $w_0 \times x_0$
 t_1 = $w_1 \times x_1$
 y = $t_0 + t_1$
in y

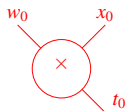
A (Functional) Program is a (Neural) Network

$$f [w_0, w_1] [x_0, x_1] \triangleq$$
$$\begin{array}{lcl} \mathbf{let} & t_0 & = w_0 \times x_0 \\ & t_1 & = w_1 \times x_1 \\ & y & = t_0 + t_1 \\ \mathbf{in} & y & \end{array}$$

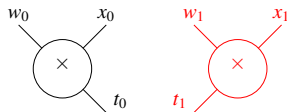
A (Functional) Program is a (Neural) Network

$$f [w_0, w_1] [x_0, x_1] \triangleq$$

let $t_0 = w_0 \times x_0$
 $t_1 = w_1 \times x_1$
 $y = t_0 + t_1$
in y



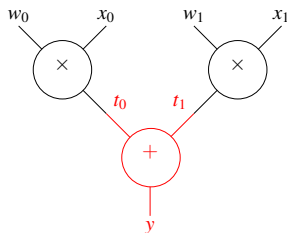
A (Functional) Program is a (Neural) Network

$$\begin{aligned} f [w_0, w_1] [x_0, x_1] &\triangleq \\ \text{let } t_0 &= w_0 \times x_0 \\ t_1 &= w_1 \times x_1 \\ y &= t_0 + t_1 \\ \text{in } y \end{aligned}$$


A (Functional) Program is a (Neural) Network

$$f [w_0, w_1] [x_0, x_1] \triangleq$$

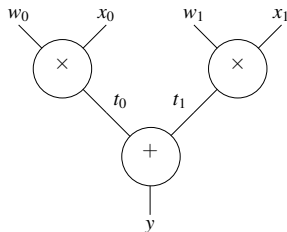
let $t_0 = w_0 \times x_0$
 $t_1 = w_1 \times x_1$
 $y = t_0 + t_1$
in y



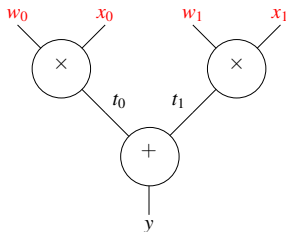
A (Functional) Program is a (Neural) Network

$$f [w_0, w_1] [x_0, x_1] \triangleq$$

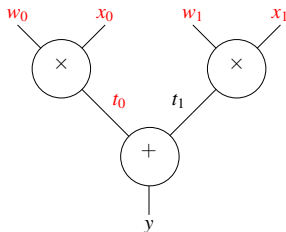
let $t_0 = w_0 \times x_0$
 $t_1 = w_1 \times x_1$
 $y = t_0 + t_1$
in y



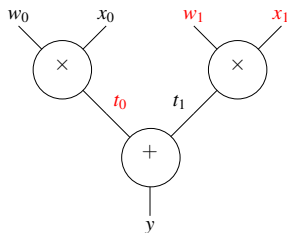
Evaluating a Network



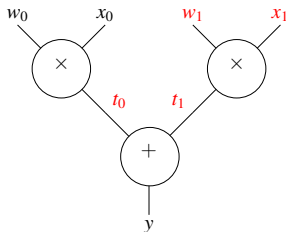
Evaluating a Network



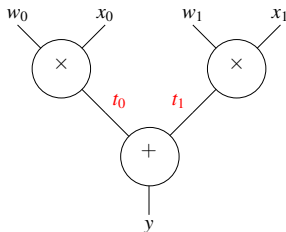
Evaluating a Network



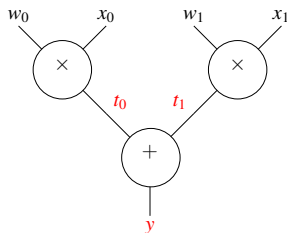
Evaluating a Network



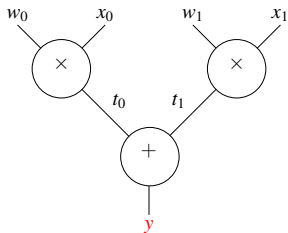
Evaluating a Network



Evaluating a Network



Evaluating a Network



Matrix multiplication and transposition

$$(\mathbf{X}_1 \times \mathbf{X}_2)^\top = \mathbf{X}_2^\top \times \mathbf{X}_1^\top$$

Programs as function composition

$$f = f_1 \circ \cdots \circ f_n$$

$$\mathbf{x}_1 = f_1 \mathbf{x}_0$$

$$\vdots$$

$$\mathbf{x}_n = f_n \mathbf{x}_{n-1}$$

The chain rule applied to programs

$$\begin{aligned}\mathcal{J} f \mathbf{x}_0 &= (\mathcal{J} f_n \mathbf{x}_{n-1}) \times \cdots \times (\mathcal{J} f_1 \mathbf{x}_0) \\ (\mathcal{J} f \mathbf{x}_0)^\top &= (\mathcal{J} f_1 \mathbf{x}_0)^\top \times \cdots \times (\mathcal{J} f_n \mathbf{x}_{n-1})^\top\end{aligned}$$

Computing the Jacobian

$$\overline{\mathbf{X}}_1' = (\mathcal{J} f_1 \mathbf{x}_0)$$

$$\overline{\mathbf{X}}_2' = (\mathcal{J} f_2 \mathbf{x}_1) \times \overline{\mathbf{X}}_1'$$

$$\vdots$$

$$\overline{\mathbf{X}}_n' = (\mathcal{J} f_n \mathbf{x}_{n-1}) \times \overline{\mathbf{X}}_{n-1}'$$

Computing the transpose of the Jacobian

$$\begin{aligned}\overline{\mathbf{X}}_{n-1} &= (\mathcal{J} f_n \mathbf{x}_{n-1})^\top \\ \overline{\mathbf{X}}_{n-2} &= (\mathcal{J} f_{n-1} \mathbf{x}_{n-2})^\top \times \overline{\mathbf{X}}_{n-1} \\ &\vdots \\ \overline{\mathbf{X}}_0 &= (\mathcal{J} f_1 \mathbf{x}_0)^\top \times \overline{\mathbf{X}}_1\end{aligned}$$

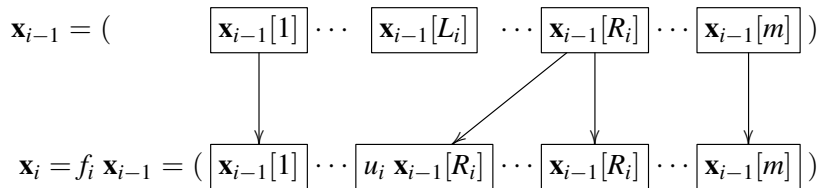
Unary machine-state transition functions

$$\mathbf{x}[L_i] := u_i \mathbf{x}[R_i]$$

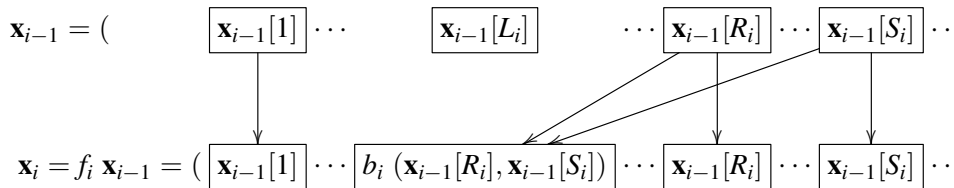
Binary machine-state transition functions

$$\mathbf{x}[L_i] := b(\mathbf{x}[R_i], \mathbf{x}[S_i])$$

What a unary machine-state transition function does



What a binary machine-state transition function does



Computing a Jacobian-vector product

$$\overrightarrow{\mathbf{x}}_n = (\mathcal{J} f \mathbf{x}_0) \times \overrightarrow{\mathbf{x}}_0$$

$$\overrightarrow{\mathbf{x}}_1 = (\mathcal{J} f_1 \mathbf{x}_0) \times \overrightarrow{\mathbf{x}}_0$$

$$\vdots$$

$$\overrightarrow{\mathbf{x}}_n = (\mathcal{J} f_n \mathbf{x}_{n-1}) \times \overrightarrow{\mathbf{x}}_{n-1}$$

Computing a vector-Jacobian product

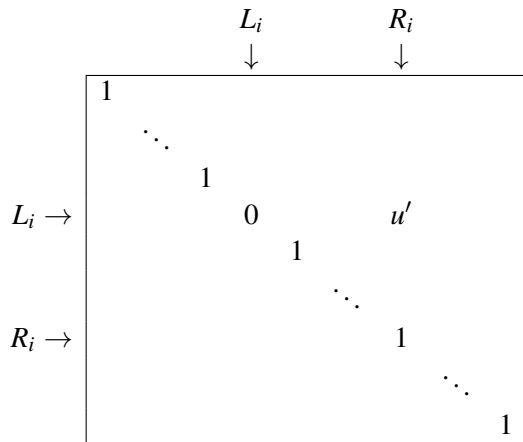
$$\overline{\mathbf{x}_0} = (\mathcal{J} f \mathbf{x}_0)^\top \times \overline{\mathbf{x}_n}$$

$$\overline{\mathbf{x}_{n-1}} = (\mathcal{J} f_n \mathbf{x}_{n-1})^\top \times \overline{\mathbf{x}_n}$$

$$\vdots$$

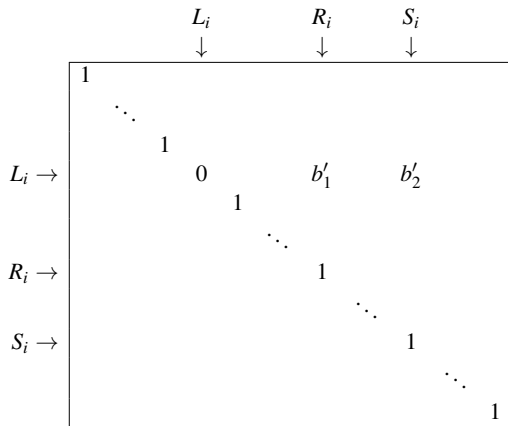
$$\overline{\mathbf{x}_0} = (\mathcal{J} f_1 \mathbf{x}_0)^\top \times \overline{\mathbf{x}_1}$$

Jacobian of a unary primitive



$$u' = \mathcal{D} u_i \mathbf{x}_{i-1}[R_i]$$

Jacobian of a binary primitive



$$b'_1 = \mathcal{D}_1 b_i (\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

$$b'_2 = \mathcal{D}_2 b_i (\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

Unary single-step Jacobian-vector product

$$\begin{pmatrix} \overrightarrow{\mathbf{x}_{i-1}}[1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i - 1] \\ u' \times \overrightarrow{\mathbf{x}_{i-1}}[R_i] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i + 1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[R_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[m] \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & & u' & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \overrightarrow{\mathbf{x}_{i-1}}[1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i - 1] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i + 1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[R_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[m] \end{pmatrix}$$

$$u' = \mathcal{D} u_i \mathbf{x}_{i-1}[R_i]$$

Binary single-step Jacobian-vector product

$$\begin{pmatrix} \overrightarrow{\mathbf{x}_{i-1}}[1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i - 1] \\ b'_1 \times \overrightarrow{\mathbf{x}_{i-1}}[R_i] + b'_2 \times \overrightarrow{\mathbf{x}_{i-1}}[S_i] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i + 1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[R_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[S_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[m] \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & & b'_1 & & b'_2 & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & & & & & \ddots & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \overrightarrow{\mathbf{x}_{i-1}}[1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i - 1] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i] \\ \overrightarrow{\mathbf{x}_{i-1}}[L_i + 1] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[R_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[S_i] \\ \vdots \\ \overrightarrow{\mathbf{x}_{i-1}}[m] \end{pmatrix}$$

$$b'_1 = \mathcal{D}_1 b_i(\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

$$b'_2 = \mathcal{D}_2 b_i(\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

Unary single-step vector-Jacobian product

$$\begin{pmatrix} \overline{\mathbf{x}}_i[1] \\ \vdots \\ \overline{\mathbf{x}}_i[L_i - 1] \\ 0 \\ \overline{\mathbf{x}}_i[L_i + 1] \\ \vdots \\ u' \times \overline{\mathbf{x}}_i[L_i] + \overline{\mathbf{x}}_i[R_i] \\ \vdots \\ \overline{\mathbf{x}}_i[m] \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & & & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & u' & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{x}}_i[1] \\ \vdots \\ \overline{\mathbf{x}}_i[L_i - 1] \\ \overline{\mathbf{x}}_i[L_i] \\ \overline{\mathbf{x}}_i[L_i + 1] \\ \vdots \\ \overline{\mathbf{x}}_i[R_i] \\ \vdots \\ \overline{\mathbf{x}}_i[m] \end{pmatrix}$$

$$u' = \mathcal{D} \, u_i \, \mathbf{x}_{i-1}[R_i]$$

Binary single-step vector-Jacobian product

$$\begin{pmatrix} \overline{\mathbf{x}}_i[1] \\ \vdots \\ \overline{\mathbf{x}}_i[L_i - 1] \\ 0 \\ \overline{\mathbf{x}}_i[L_i + 1] \\ \vdots \\ b'_1 \times \overline{\mathbf{x}}_i[L_i] + \overline{\mathbf{x}}_i[R_i] \\ \vdots \\ b'_2 \times \overline{\mathbf{x}}_i[L_i] + \overline{\mathbf{x}}_i[S_i] \\ \vdots \\ \overline{\mathbf{x}}_i[m] \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & & & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & b'_1 & & & & 1 & & \\ & & & & & & & \ddots & \\ & & b'_2 & & & & & & 1 \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{x}}_i[1] \\ \vdots \\ \overline{\mathbf{x}}_i[L_i - 1] \\ \overline{\mathbf{x}}_i[L_i] \\ \overline{\mathbf{x}}_i[L_i + 1] \\ \vdots \\ \overline{\mathbf{x}}_i[R_i] \\ \vdots \\ \overline{\mathbf{x}}_i[S_i] \\ \vdots \\ \overline{\mathbf{x}}_i[m] \end{pmatrix}$$

$$b'_1 = \mathcal{D}_1 \, b_i \, (\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

$$b'_2 = \mathcal{D}_2 \, b_i \, (\mathbf{x}_{i-1}[R_i], \mathbf{x}_{i-1}[S_i])$$

$$f = f_1 \circ \cdots \circ f_n$$

$$f = f_1 \circ \cdots \circ f_n$$
$$\mathcal{J}(f)(x_0) = \mathcal{J}(f_n)(x_{n-1}) \times \cdots \times \mathcal{J}(f_1)(x_0)$$

$$\begin{aligned}f &= f_1 \circ \cdots \circ f_n \\ \mathcal{J}(f)(x_0) &= \mathcal{J}(f_n)(x_{n-1}) \times \cdots \times \mathcal{J}(f_1)(x_0) \\ \acute{x}_n &= \mathcal{J}(f)(x_0) \times \acute{x}_0\end{aligned}$$

$$\begin{aligned}f &= f_1 \circ \cdots \circ f_n \\ \mathcal{J}(f)(x_0) &= \mathcal{J}(f_n)(x_{n-1}) \times \cdots \times \mathcal{J}(f_1)(x_0) \\ \acute{x}_n &= \mathcal{J}(f)(x_0) \times \acute{x}_0\end{aligned}$$

$$\begin{aligned}x_1 &= f_1(x_0) \\ \acute{x}_1 &= \mathcal{J}(f_1)(x_0) \times \acute{x}_0 \\ &\vdots \\ x_n &= f_n(x_{n-1}) \\ \acute{x}_n &= \mathcal{J}(f_n)(x_{n-1}) \times \acute{x}_{n-1}\end{aligned}$$

$$f = f_1 \circ \cdots \circ f_n$$

$$f = f_1 \circ \cdots \circ f_n$$
$$\mathcal{J}(f)(x_0)^\top = \mathcal{J}(f_1)(x_0)^\top \times \cdots \times \mathcal{J}(f_n)(x_{n-1})^\top$$

$$\begin{aligned}f &= f_1 \circ \cdots \circ f_n \\ \mathcal{J}(f)(x_0)^\top &= \mathcal{J}(f_1)(x_0)^\top \times \cdots \times \mathcal{J}(f_n)(x_{n-1})^\top \\ \dot{x}_0 &= \mathcal{J}(f)(x_0)^\top \times \dot{x}_n\end{aligned}$$

$$\begin{aligned}f &= f_1 \circ \cdots \circ f_n \\ \mathcal{J}(f)(x_0)^\top &= \mathcal{J}(f_1)(x_0)^\top \times \cdots \times \mathcal{J}(f_n)(x_{n-1})^\top \\ \dot{x}_0 &= \mathcal{J}(f)(x_0)^\top \times \dot{x}_n\end{aligned}$$

$$\begin{aligned}x_1 &= f_1(x_0) \\ &\vdots \\ x_n &= f_n(x_{n-1}) \\ \dot{x}_{n-1} &= \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_n \\ &\vdots \\ \dot{x}_0 &= \mathcal{J}(f_1)(x_0) \times \dot{x}_1\end{aligned}$$

Forward Mode by Overloading

$$x_1 = f_1(x_0)$$

$$\dot{x}_1 = \mathcal{J}(f_1)(x_0) \times \dot{x}_0$$

$$\vdots$$

$$x_n = f_n(x_{n-1})$$

$$\dot{x}_n = \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_{n-1}$$

Forward Mode by Overloading

$$x_1 = f_1(x_0)$$

$$\dot{x}_1 = \mathcal{J}(f_1)(x_0) \times \dot{x}_0$$

$$\vdots$$

$$x_n = f_n(x_{n-1})$$

$$\dot{x}_n = \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_{n-1}$$

$$x_i = f_i(x_{i-1})$$

Forward Mode by Overloading

$$x_1 = f_1(x_0)$$

$$\dot{x}_1 = \mathcal{J}(f_1)(x_0) \times \dot{x}_0$$

$$\vdots$$

$$x_n = f_n(x_{n-1})$$

$$\dot{x}_n = \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_{n-1}$$

$$x_i = f_i(x_{i-1})$$

$$\langle x_i, \dot{x}_i \rangle = \langle f_i(x_{i-1}), \mathcal{J}(f_i)(x_{i-1}) \times \dot{x}_{i-1} \rangle$$

Forward Mode by Overloading

$$x_1 = f_1(x_0)$$

$$\dot{x}_1 = \mathcal{J}(f_1)(x_0) \times \dot{x}_0$$

$$\vdots$$

$$x_n = f_n(x_{n-1})$$

$$\dot{x}_n = \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_{n-1}$$

$$x_i = f_i(x_{i-1})$$

$$\langle x_i, \dot{x}_i \rangle = \langle f_i(x_{i-1}), \mathcal{J}(f_i)(x_{i-1}) \times \dot{x}_{i-1} \rangle$$

$$\overrightarrow{x_i} = \overrightarrow{f_i}(\overrightarrow{x_{i-1}})$$

$$x_1 = f_1(x_0)$$

$$\vdots$$

$$x_n = f_n(x_{n-1})$$

$$\dot{x}_{n-1} = \mathcal{J}(f_n)(x_{n-1}) \times \dot{x}_n$$

$$\vdots$$

$$\dot{x}_0 = \mathcal{J}(f_1)(x_0) \times \dot{x}_1$$