# Reverse Mode of AD I

- Consider

$$\bar{v}_i = \frac{\partial y_j}{\partial v_i}$$

- Note that earlier we considered

$$\dot{v}_i = \frac{\partial v_i}{\partial x_1}$$

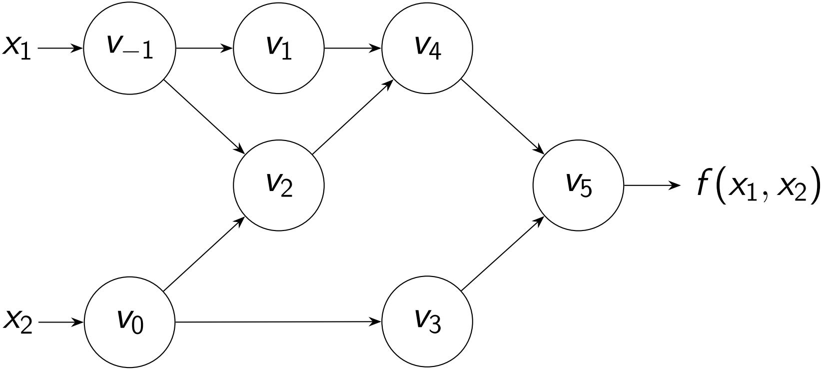
- Consider again

$$f(x_1, x_2) = \ln x_1 + x_1 x_2 - \sin x_2$$

- Let us check the variable $v_0$

# Reverse Mode of AD II

- From the computational graph




$v_0$ can affect $y$ through affecting $v_2$ and $v_3$

# Reverse Mode of AD III

- Thus
$$\frac{\partial y}{\partial v_0} = \frac{\partial y}{\partial v_2}\frac{\partial v_2}{\partial v_0} + \frac{\partial y}{\partial v_3}\frac{\partial v_3}{\partial v_0}$$
or
$$\bar{v}_0 = \bar{v}_2\frac{\partial v_2}{\partial v_0} + \bar{v}_3\frac{\partial v_3}{\partial v_0}$$
- In the practical implementation shown later, this is done in two steps
$$\bar{v}_0 \leftarrow \bar{v}_3\frac{\partial v_3}{\partial v_0}$$
$$\bar{v}_0 \leftarrow \bar{v}_0 + \bar{v}_2\frac{\partial v_2}{\partial v_0}$$

# Reverse Mode of AD IV

- They are part of the following sequence of reverse computation:

# Reverse Mode of AD V

$$
\begin{array}{llll}
\bar{x}_1 & = \bar{v}_{-1} & & = 5.5 \\
\bar{x}_2 & = \bar{v}_0 & & = 1.716 \\
\hline
\bar{v}_{-1} & = \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}} & = \bar{v}_{-1} + \bar{v}_1 / v_{-1} & = 5.5 \\
\bar{v}_0 & = \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0} & = \bar{v}_0 + \bar{v}_2 \times v_{-1} & = 1.716 \\
\bar{v}_{-1} & = \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}} & = \bar{v}_2 \times v_0 & = 5 \\
\bar{v}_0 & = \bar{v}_3 \frac{\partial v_3}{\partial v_0} & = \bar{v}_3 \times \cos v_0 & = -0.284 \\
\bar{v}_2 & = \bar{v}_4 \frac{\partial v_4}{\partial v_2} & = \bar{v}_4 \times 1 & = 1 \\
\bar{v}_1 & = \bar{v}_4 \frac{\partial v_4}{\partial v_1} & = \bar{v}_4 \times 1 & = 1 \\
\bar{v}_3 & = \bar{v}_5 \frac{\partial v_5}{\partial v_3} & = \bar{v}_5 \times (-1) & = -1 \\
\bar{v}_4 & = \bar{v}_5 \frac{\partial v_5}{\partial v_4} & = \bar{v}_5 \times 1 & = 1 \\
\hline
\bar{v}_5 & = \bar{y} & = 1 &
\end{array}
$$

# Reverse Mode of AD VI

- Earlier in the forward process we have

$$y = v_5$$

- Thus in the reverse mode, we begin with

$$\bar{v}_5 = \frac{\partial y}{\partial v_5} = \frac{\partial y}{\partial y} = 1$$

# Reverse Mode of AD VII

- Then because

$$v_4 = \ln x_1 + x_1 x_2$$

affects $y$ only through $v_5$, we have

$$\begin{aligned}\frac{\partial y}{\partial v_4} &= \frac{\partial y}{\partial v_5}\frac{\partial v_5}{\partial v_4} \\ &= \bar{v}_5 \frac{\partial v_5}{\partial v_4} = \bar{v}_5 \times 1\end{aligned}$$

# Reverse Mode of AD VIII

- We continue the process until at the end

$$\frac{\partial y}{\partial x_1} = \bar{x}_1 = \bar{v}_{-1}$$

and

$$\frac{\partial y}{\partial x_2} = \bar{x}_2 = \bar{v}_0$$

are obtained

# Reverse Mode of AD IX

- Note that
$$\frac{\partial y}{\partial x_1} \text{ and } \frac{\partial y}{\partial x_2}$$
are obtained at the same time
- Therefore, an advantage of the reverse mode is that it is suitable for a function with many input variables
- This is useful for calculating the gradient
$$\nabla f = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}^T$$

# Reverse Mode of AD X

- For general

$$f : R^n \rightarrow R^m$$

the Jacobian calculation needs $m$ passes for the $m$ rows:

$$\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ & \ddots & \\ \frac{\partial y_m}{\partial x_1} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- Thus reverse model is better than forward if

$$m \ll n$$

# Transposed Jacobian-vector Products I

- Earlier we talked about Jacobian-vector products
- In optimization another commonly used operation is the

  transposed Jacobian-vector product

- That is

$$J^T \boldsymbol{r} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ & \ddots & \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix}$$

# Transposed Jacobian-vector Products II

- By initializing

$$\bar{\boldsymbol{y}} = \boldsymbol{r}$$

we can calculate $J^T \boldsymbol{r}$ in one pass

# AD and Back-propagation I

- The network itself is a computational graph
- The input of a layer affects $\xi_i$ only through the output
- See the following derivation discussed before

$$
\begin{aligned}
&\frac{\partial \xi_i}{\partial \mathrm{vec}(S^{m,i})^T} \\
=&\frac{\partial \xi_i}{\partial \mathrm{vec}(\sigma(S^{m,i}))^T} \frac{\partial \mathrm{vec}(\sigma(S^{m,i}))}{\partial \mathrm{vec}(S^{m,i})^T} \\
=&\frac{\partial \xi_i}{\partial \mathrm{vec}(Z^{m+1,i})^T} \frac{\partial \mathrm{vec}(Z^{m+1,i})}{\partial \mathrm{vec}(\sigma(S^{m,i}))^T} \frac{\partial \mathrm{vec}(\sigma(S^{m,i}))}{\partial \mathrm{vec}(S^{m,i})^T}
\end{aligned} \tag{1}
$$

# AD and Back-propagation II

In (1), $S^{m,i}$ affects $\xi_i$ only through $\sigma(S^{m,i})$

- Thus back-propagation is a special case of the reverse mode of automatic differentiation