we have a sparse matrix 'A' described by a list of 'n' entries 'a' and lists 'ik' and 'jk' of corresponding indices.

to construct A, we define the matrices

$$I = \begin{pmatrix} 1 & 1 \\ ik_1 & ik_2 \cdots \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \cdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

--- example ...

$$J = \begin{pmatrix} 1 & 1 \\ jk_1 & jk_2 \cdots \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \cdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the k-th column is a one hot vector into the row/column of A. for entry 'k'.

Then, we may construct A as

$$A_{ij} = I_{ik} J_{jk} a_k$$

or, in matrix notation:

$$A = I \, \mathrm{diag}(a) \, J^T$$

where $\mathrm{diag}(a) = \begin{pmatrix} a_1 & & \\ & a_2 & \\ & & \ddots \\ & & & a_n \end{pmatrix}$

for convenience, we define $O = J^T$.

$$A = I \, \mathrm{diag}(a) \, O$$

One can confirm that to strip entries 'a' from matrix 'A', the process is:

$$a_k = I_{ik} A_{ij} O_{kj}$$

or, in matrix notation:

$$q = \text{diag}(I^T A O^T),$$

where 'diag(M)' returns the diagonal entries in 'M'.

We have function $b = f(a)$ that takes entries into A and returns entries into $B = AX$.

our goal is to find the Jacobian

$$\frac{\partial b}{\partial a} = \frac{\partial f}{\partial a}$$

because $f()$ returns the entries into B, it can be written as

$$f(a) = \text{diag}(I_b^T B O_b^T)$$

$$= \text{diag}\left(I_b^T A X O_b^T\right)$$

we then insert the form of A.

$$= \text{diag}\left(I_b^T I_a \, \text{diag}(a) \, O_a X O_b^T\right)$$

The derivative of diag(v) is just diag of the derivative of v.

So $$\frac{db}{da} = \text{diag}\left(I_b^T I_a \frac{d\,\text{diag}(a)}{da} O_a X O_b^T\right)$$

the term $\frac{d\,\text{diag}(a)}{da}$ is a rank 3 tensor with ijk elements $\delta_{ij}\delta_{jk}$

then, we can write other VJPs

as follows.

$$\frac{db}{da}^T v = \mathrm{diag}\left(I_a^T \, I_b \, \mathrm{diag}(v) \, O_b \, X^T \, O_a^T \right)$$

$$\frac{db}{da} g = \mathrm{diag}\left(I_a \, g \, O_a \, X \right)$$

I_b^T O_b^T

these we need to implement in primitives.

for backprop:

$$\frac{db}{da}^T v = \mathrm{diag}\left(I_a^T \cdot I_b \, \mathrm{diag}(v) \, O_b \cdot X^T \, O_a^T \right)$$

for forward prop:

$$\frac{db}{da} g = \mathrm{diag}\left(I_b^T \cdot I_a \, \mathrm{diag}(g) \, O_a \cdot X \cdot O_b^T \right).$$

note that $I_c \, \mathrm{diag}(v) \, O_c \cdot U$

is equivalent to multiplying a matrix VU where V uses entries of V but indices of C.