Why does $\text{AllGather}_X(A[I_X]) \to A[I]$ in forward imply $\text{ReduceScatter}_X(A'[I]\{U_X\}) \to A'[I_X]$ in backward?

Imagine: $A[I_X]$ represents model parameters sharded along X axis
and $data[I_X]$ is also sharded.

Forward pass:

$$data[I_X] \cdot A[I_X]$$

$$\downarrow \text{AllGather}$$

$$\to \underbrace{data[I_X] \cdot A[I]}_{\| \; \|}$$

$$\to loss[I_X] \; (value - label)$$

$$\to loss = \sum_{I_X} loss[I_X]$$

$$\frac{\partial loss}{\partial A} = \underbrace{\sum_{I_X} \frac{\partial loss[I_X]}{\partial A}}$$

↳ This means if we just compute $\frac{\partial loss[I_X]}{\partial A}$ on each device it will result in $\underbrace{A'[I]\{U_X\}}_{\frac{\partial loss}{\partial A}}$ (U_X: unreduced)

To obtain the true gradient $\frac{\partial loss}{\partial A}$, we could

$$AllReduce(A'[I]\{U_x\}) \longrightarrow A'[I]$$

But the goal is to keep $A[I_x]$ on each device.

We can just do:

$$ReduceScatter_x(A'[I]\{U_x\}) \longrightarrow A'[I_x]$$

$A'[I_x]$ already gives us enough info to do

$$A[I_x] \longleftarrow A[I_x] - lr \cdot A'[I_x]$$

Why does $\text{ReduceScatter}_X(A[I]\{U_X\}) \to A[I_V]$ in forward pass imply

$$\text{AllGather}_X(A'[I_X]) \to A'[I]$$

Scenario:

$$\text{Input}[B, D],\ W_1[D, F_X],\ W_2[F_X, G]$$

$$\text{Input}[B, D] @ W_1[D, F_X]$$

$$= \text{out}_1[B, F_X]$$

$$\Rightarrow \text{out}_1\{B, F_X\} @ W_2[F_X, G] \xrightarrow{\text{locally}} \text{out}_2[B, G]\{U_X\}$$

In theory, if we want to make $\text{out}_2[B, G]$ fully replicated, we need to do

$$\text{AllReduce} \overset{\text{which is}}{=} \text{ReduceScatter} \to \text{AllGather}$$

In practice, we don't have to do that, we can

$$\text{ReduceScatter}_X(\text{out}_2[B, G]\{U_X\}) = \text{out}_2[B, G_X]$$

and don't do AllGather because not necessary needed

In this case, when we receive $out_2'[B, G_x]$,

to compute $W_2'[F_x, G]$

By chain-rule: $\underset{\mathbb{R}^{1\times G}}{z} = \underset{\mathbb{R}^{1\times|}}{x}\ \underset{\mathbb{R}^{F\times G}}{W}$

If we already have: $\frac{\partial L}{\partial z} \in \mathbb{R}^{1\times G}$

$$\frac{\partial L}{\partial W} = \underset{\mathbb{R}^{F\times 1}}{x}^{T} \underset{\mathbb{R}^{1\times G}}{\frac{\partial L}{\partial z}}$$

So to compute: $W_2'[F_x, G]$

we need $out_1[B, F]$ and $out_2'[B, G]$

More precisely: $\frac{1}{B}(out_1^T[F, B] @ out_2'[B, G])$

In this case, $out_1[B, F_x]$ already matches $W_2'[F_x, G]$

sharded match

but we need G fully sharded.

Therefore, during the backward pass,

we need to $AllGather_x(out_2'[B, G_x])$

and then multiply this with $out_1[B, F_x]$