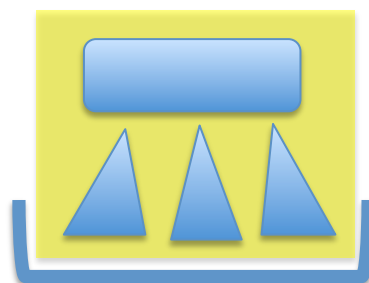


Efficient Large-Scale Distributed Training of Conditional Maxent

Gideon Mann, Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker

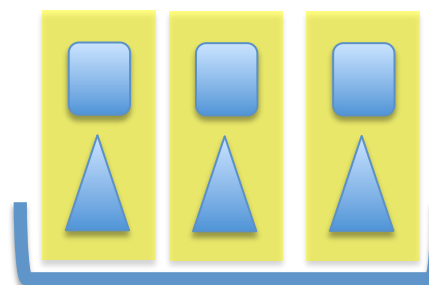
Distributed
training
methods

Distributed Gradient



$O_{\text{network}}(NT)$

Mixture Weight



$O_{\text{network}}(N)$

- **Theory:** Convergence bounds for maxent parameters.
- **Experiments:** Data sets with 1M to 1B examples.

Theory and experiments show the mixture weight method has about the same accuracy, but requires less than $1/100^{\text{th}}$ of the network usage needed by the distributed gradient and typically is somewhat faster.