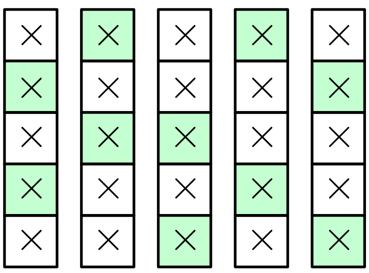
TAMUNA: Doubly Accelerated Federated Learning with Local Training, Compression, and Partial Participation

King Abdullah University of Science and Technology
KAUST
Saudi Arabia

Laurent Condat, Ivan Agarský, Grigory Malinovsky, Peter Richtárik

TAMUNA

input: stepsizes $\gamma > 0$, $\eta > 0$;
number of participating clients $c \in \{2, \dots, n\}$
sparsity index $s \in \{2, \dots, n\}$ for compression
initial model estimate $\bar{x}^0 \in \mathbb{R}^d$
initial control variates $h_i^0 \in \mathbb{R}^d$ s.t. $\sum_{i=1}^n h_i^0 = 0$
for $r = 0, 1, \dots$ (rounds) **do**
  choose a subset $\Omega^r \in [n]$ of size c
  choose the number of local steps L^r
  for clients $i \in \Omega^r$, in parallel, **do**
    $x^{r,0} := \bar{x}^r$
    for $l = 0, \dots, L^r$ **do**
      $x_i^{r,l} := x_i - \gamma g_i^{r,l} + \gamma h_i^r$ with $g_i^{r,l} \approx \nabla f_i^t(x_i^{r,l})$
    end for
    send $v_i^r := \mathcal{C}_i^r(x_i^{r,L^r})$ to server   // uplink comm.
  end for
  at server: $\bar{x}^{r+1} := \frac{1}{s} \sum_{i \in \Omega^r} v_i^r$ // model update
  $\bar{x}^{r+1}$ is sent to clients $i \in \Omega^r \cup \Omega^{r+1}$ // downlink comm.
  for clients $i \in \Omega^r$, in parallel, **do**   // update of control variates
    $h_i^{r+1} := h_i^r + \frac{\eta}{\gamma}\left(\mathcal{C}_i^r(\bar{x}^{r+1}) - v_i^r\right)$
  end for
  for clients $i \notin \Omega^r$, in parallel, **do**
    $h_i^{r+1} := h_i^r$ // idle clients
  end for
end for



$\mathcal{C}_i^r$: random selectors with
- s (=2) comm. values per coordinate
- $\leq \lceil \frac{sd}{c} \rceil$ (=2) comm. values per active client

With $\gamma \approx \frac{1}{L}$, $s \approx \max\left(2, \frac{c}{d}, \alpha c\right)$, $L^r \approx \max\left(\sqrt{\frac{s\kappa}{n}}, 1\right)$, $\eta \approx \frac{1}{2L^r}$, $g_i = \nabla f_i$, TotalCom (= UpCom + α.DownCom) complexity of TAMUNA in #floats:

$$\mathcal{O}\left(\left(\sqrt{d}\sqrt{\kappa}\sqrt{\frac{n}{c}} + d\sqrt{\kappa}\frac{\sqrt{n}}{c} + d\frac{n}{c} + \sqrt{\alpha}\, d\sqrt{\kappa}\sqrt{\frac{n}{c}}\right)\log \epsilon^{-1}\right)$$

→ New SOTA with double acceleration

Distributed optimization with n clients + server:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$$

Every f_i is $\mathcal{L}$-smooth and μ-strongly convex. $\kappa := \frac{\mathcal{L}}{\mu}$

Algorithms with Local Training (LT) or Compressed Communication (CC), in case of full participation:

Algorithm	LT	CC	uplink comm.
DIANA	✗	✓	$\tilde{\mathcal{O}}((1+\frac{d}{n})\kappa + d)$
EF21	✗	✓	$\tilde{\mathcal{O}}(d\kappa)$
Scaffold	✓	✗	$\tilde{\mathcal{O}}(d\kappa)$
FedLin	✓	✗	$\tilde{\mathcal{O}}(d\kappa)$
S-Local-GD	✓	✗	$\tilde{\mathcal{O}}(d\kappa)$
Scaffnew	✓	✗	$\tilde{\mathcal{O}}(d\sqrt{\kappa})$
5GCS	✓	✗	$\tilde{\mathcal{O}}(d\sqrt{\kappa})$
FedCOMGATE	✓	✓	$\tilde{\mathcal{O}}(d\kappa)$
TAMUNA	✓	✓	$\tilde{\mathcal{O}}\left(\sqrt{d}\sqrt{\kappa} + d\frac{\sqrt{\kappa}}{\sqrt{n}} + d\right)$

(full participation: TAMUNA reverts to CompressedScaffnew [Condat et al. 2022])

Algorithms with LT or CC, and allowing for Partial Participation:

Algorithm	LT	CC	uplink communication
DIANA-PP	✗	✓	$\tilde{\mathcal{O}}((1+\frac{d}{c})\kappa + d\frac{n}{c})$
Scaffold	✓	✗	$\tilde{\mathcal{O}}(d\kappa + d\frac{n}{c})$
5GCS	✓	✗	$\tilde{\mathcal{O}}\left(d\sqrt{\kappa}\sqrt{\frac{n}{c}} + d\frac{n}{c}\right)$
TAMUNA	✓	✓	$\tilde{\mathcal{O}}\left(\sqrt{d}\sqrt{\kappa}\sqrt{\frac{n}{c}} + d\sqrt{\kappa}\frac{\sqrt{n}}{c} + d\frac{n}{c}\right)$