LoCoDL: COMMUNICATION-EFFICIENT DISTRIBUTED LEARNING WITH LOCAL TRAINING AND COMPRESSION

King Abdullah University of Science and Technology
KAUST
Saudi Arabia

Laurent Condat, Artavazd Maranjyan, Peter Richtárik

LoCoDL

input: stepsizes $\gamma, \chi, \rho > 0$; probability $p \in (0,1]$; $\omega \geq 0$; initial estimates $x_1^0, \dots, x_n^0, y^0 \in \mathbb{R}^d$ and $u_1^0, \dots, u_n^0, v^0 \in \mathbb{R}^d$ such that $\frac{1}{n}\sum_{i=1}^n u_i^0 + v^0 = 0$.

for $t = 0, 1, \dots$ **do**
  for $i = 1, \dots, n$, at clients in parallel, **do**
    $\hat{x}_i^t := x_i^t - \gamma \nabla f_i(x_i^t) + \gamma u_i^t$
    $\hat{y}^t := y^t - \gamma \nabla g(y^t) + \gamma v^t$ // identical copies at clients
    flip a coin $\theta^t \in \{0,1\}$ with $\mathrm{Prob}(\theta^t = 1) = p$
    if $\theta^t = 1$ **then**
      $d_i^t := \mathcal{C}_i^t(\hat{x}_i^t - \hat{y}^t)$
      send d_i^t to the server
      at server: aggregate $\bar{d}^t := \frac{1}{2n}\sum_{j=1}^n d_j^t$ and send $\bar{d}^t$ to all clients
      $x_i^{t+1} := (1-\rho)\hat{x}_i^t + \rho(\hat{y}^t + \bar{d}^t)$
      $u_i^{t+1} := u_i^t + \frac{p\chi}{\gamma(1+2\omega)}(\bar{d}^t - d_i^t)$
      $y^{t+1} := \hat{y}^t + \rho \bar{d}^t$
      $v^{t+1} := v^t + \frac{p\chi}{\gamma(1+2\omega)}\bar{d}^t$
    else
      $x_i^{t+1} := \hat{x}_i^t$, $y^{t+1} = \hat{y}^t$, $u_i^{t+1} := u_i^t$, $v^{t+1} := v^t$
    end if
  end for
end for

Distributed optimization with n clients + server:

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

f_i: private loss, g: shared loss
Client i calls ∇f_i and ∇g

All f_i and g are L-smooth and μ-strongly convex. $\kappa := \frac{L}{\mu}$

primal-dual optimality conditions:

- $x_1 = \cdots = x_n = y$
- $0 = \nabla f_i(x_i) - u_i, \ \forall i \in [n]$
- $0 = \nabla g(y) - v$
- $0 = u_1 + \cdots + u_n + nv$

General unbiased compressors with relative variance $\omega \geq 0$:

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2, \ \forall x$$

e.g. rand-k: $\omega = \frac{d}{k} - 1$

Theorem (linear convergence). With $\gamma < \frac{2}{L+\mu}$, suitable ρ and χ, then for every $t \geq 0$,

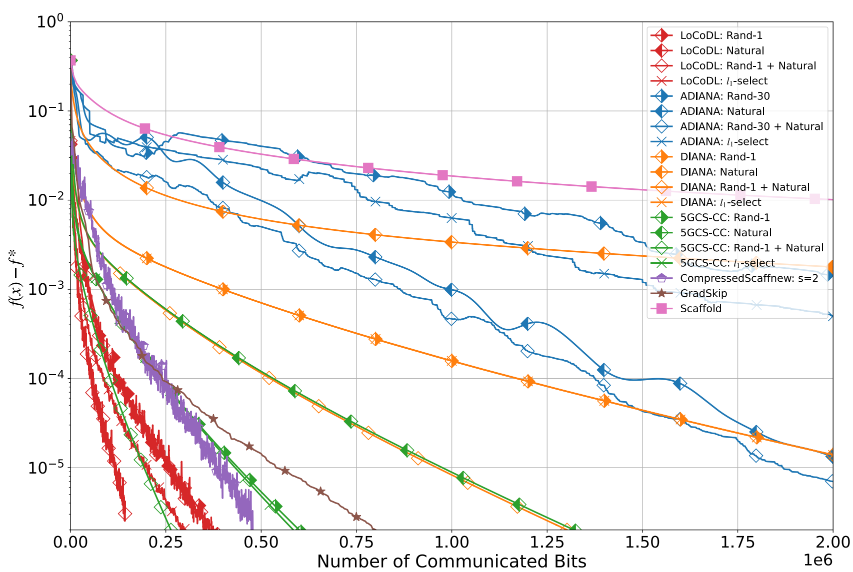
$$\mathbb{E}[\Psi^t] \leq \max\left((1-\gamma\mu)^2, 1 - \frac{p^2\chi}{1+2\omega}\right)^t \Psi^0,$$

where $\Psi^t := \frac{1}{\gamma}\left(\sum_{i=1}^n \left\|x_i^t - x^\star\right\|^2 + n\left\|y^t - x^\star\right\|^2\right) \frac{\gamma(1+2\omega)}{p^2\chi}\left(\sum_{i=1}^n \left\|u_i^t - \nabla f_i(x^\star)\right\|^2 + n\left\|v^t - \nabla g(x^\star)\right\|^2\right)$

Best complexity with independent rand-k compressors, $k = \lceil \frac{d}{n} \rceil$

Uplink communication complexity in #reals:

Algorithm	$\mathcal{O}(\cdot \log \epsilon^{-1})$	if $n = \mathcal{O}(d)$
Scaffold	$d\kappa$	$d\kappa$
Scaffnew	$d\sqrt{\kappa}$	$d\sqrt{\kappa}$
EF21	$d\kappa$	$d\kappa$
DIANA	$(1+\frac{d}{n})\kappa + d$	$\frac{d}{n}\kappa + d$
ADIANA	$\left(1+\frac{d}{\sqrt{n}}\right)\sqrt{\kappa} + d$	$\frac{d}{\sqrt{n}}\sqrt{\kappa} + d$
FedComGate	$d\kappa$	$d\kappa$
5GCS-CC	$\left(\sqrt{d} + \frac{d}{\sqrt{n}}\right)\sqrt{\kappa} + d$	$\frac{d}{\sqrt{n}}\sqrt{\kappa} + d$
C-Scaffnew	$\left(\sqrt{d} + \frac{d}{\sqrt{n}}\right)\sqrt{\kappa} + d$	$\frac{d}{\sqrt{n}}\sqrt{\kappa} + d$
LoCoDL	$\left(\sqrt{d} + \frac{d}{\sqrt{n}}\right)\sqrt{\kappa} + d$	$\frac{d}{\sqrt{n}}\sqrt{\kappa} + d$



Logistic regression, LibSVM a5a, $d = 122$, $n = 288$