CoLA: Decentralized Linear Learning

[1] EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Lie He[1*], An Bian[2*], Martin Jaggi[1]

[2] ETH zürich

Overview

We present a decentralized linear learning framework which is communication efficient and has the following advantages:

- ✓ *Convergence Guarantee*
- ✓ *Communication Efficiency and Usability*
- ✓ *Elasticity and Fault Tolerance*
- ✓ *Local Certificate provided*

Setup

Generalized Linear Models: Lasso, Logistic Regression …

$$\min_{\mathbf{x}\in\mathbb{R}^n} F_A(\mathbf{x}) := f(\mathbf{A}\mathbf{x}) + \sum_{i=1}^n g_i(x_i)$$

where $\mathbf{A} := [\mathbf{A}_1; \dots; \mathbf{A}_n] \in \mathbb{R}^{d\times n}$

Distributed over K nodes: Partition $\mathbf{A}$ by columns $\{\mathcal{P}_k\}_{k=1}^K$
Node k has part of data and weights: $\mathbf{A}_{[k]}$, $\mathbf{x}_{[k]}$

Decentralized network topology: $\mathcal{W}$ is a doubly stochastic matrix where a non-zero entry $\mathcal{W}_{ij}$ represents communication between nodes i and j.
1 - β is the spectral gap of $\mathcal{W}$.

Decentralized reformulation: Objectives A, B & Duality gap

Decentralized objective A

$$\min_{\mathbf{x},\,\{\mathbf{v}_k\}_k} \; H_A(\mathbf{x},\{\mathbf{v}_k\}_k) = \tfrac{1}{K}\Sigma_k\, f(\mathbf{v}_k) + g(\mathbf{x}) \quad \text{s.t. } \mathbf{v}_k = \mathbf{A}\mathbf{x},\ k\in[K]$$

Decentralized objective B

$$\min_{\{\mathbf{w}_k\}_k} \; H_B(\{\mathbf{w}_k\}_k) = \tfrac{1}{K}\Sigma_k\, f^*(\mathbf{w}_k) + \Sigma_{i=1}^n g_i^*(-\mathbf{A}_i^{\mathrm{T}}(\tfrac{1}{K}\Sigma_k\, \mathbf{w}_k))$$

Decentralized Duality gap

$$G_H(\mathbf{x}, \{\mathbf{v}_k\}_k, \{\mathbf{w}_k\}_k) = \tfrac{1}{K}\Sigma_k\,(f(\mathbf{v}_k) + f^*(\mathbf{w}_k)) + g(\mathbf{x}) + \Sigma_{i=1}^n g_i^*(-\mathbf{A}_i^{\mathrm{T}}(\tfrac{1}{K}\Sigma_k\, \mathbf{w}_k))$$

f^*, g^*: convex conjugate

Consensus violation

$$\Sigma_k \|\mathbf{v}_k - \mathbf{A}\mathbf{x}\|^2$$

Proved convergence on decentralized duality gap, decentralized objectives and consensus violation for a general topology $\mathcal{W}$

Algorithm and Theory

CoLA: Communication-Efficient Decentralized Linear Learning

Input: Data matrix $\mathbf{A}$. Mixing matrix $\mathcal{W}$. Aggregation parameter $\gamma \in (0,1]$, & subproblem parameter σ'. Starting point $\mathbf{x}^0 := \mathbf{0}$, $\mathbf{v}_k^0 := \mathbf{0}, \forall k = 1,\dots,K$

For $t = 0,1,\dots,T$ **do**:
  For $k \in \{1,\dots,K\}$ in parallel over all nodes **do**: $\mathbf{v}_k$: local estimate of $\mathbf{A}\mathbf{x}$
    compute locally averaged vector $\mathbf{v}_k^{t+1/2} := \Sigma_{l=1}^K \mathcal{W}_{kl}\mathbf{v}_l^t$
    $\Delta\mathbf{x}_{[k]} \leftarrow \Theta$-approximate solution to subproblem at $\mathbf{v}_k^{t+1/2}$
    update local variable $\mathbf{x}_{[k]}^{t+1} := \mathbf{x}_{[k]}^t + \gamma\Delta\mathbf{x}_{[k]}$
    compute update of local estimate $\Delta\mathbf{v}_k := \mathbf{A}_{[k]}\Delta\mathbf{x}_{[k]}$
    $\mathbf{v}_k^{t+1} := \mathbf{v}_k^{t+1/2} + \gamma K\Delta\mathbf{v}_k$
  End
End

Node k solves the **local subproblem** Θ-approximately

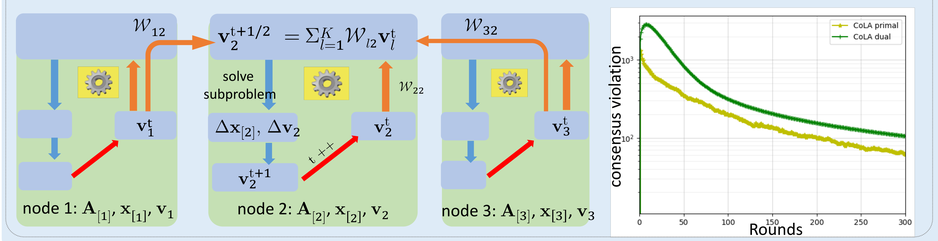
$$\min_{\Delta\mathbf{x}_{[k]}\in\mathbb{R}^n} \mathcal{G}_k^{\sigma'}(\Delta\mathbf{x}_{[k]}; \mathbf{v}_k, \mathbf{x}_{[k]}) := \tfrac{1}{K} f(\mathbf{v}_k) + \nabla f(\mathbf{v}_k)^\top \mathbf{A}_{[k]}\Delta\mathbf{x}_{[k]} + \frac{\sigma'}{2\tau}\left\|\mathbf{A}_{[k]}\Delta\mathbf{x}_{[k]}\right\|^2 + \textstyle\sum_{i\in\mathcal{P}_k} g_i(x_i + (\Delta\mathbf{x}_{[k]})_i).$$

Θ-approximate solution $\Delta\mathbf{x}_{[k]}$

$$\frac{\mathbb{E}[\mathcal{G}_k^{\sigma'}(\Delta\mathbf{x}_{[k]}; \mathbf{v}_k, \mathbf{x}_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\mathbf{x}_{[k]}^\star; \mathbf{v}_k, \mathbf{x}_{[k]})]}{\mathcal{G}_k^{\sigma'}(\mathbf{0}; \mathbf{v}_k, \mathbf{x}_{[k]}) - \mathcal{G}_k^{\sigma'}(\Delta\mathbf{x}_{[k]}^\star; \mathbf{v}_k, \mathbf{x}_{[k]})} \le \Theta,$$

Consensus Violation

Observations: $\frac{1}{K}\Sigma_k\, \mathbf{v}_k = \mathbf{A}\mathbf{x}$; $F_A(\mathbf{x}) \le H_A(\mathbf{x},\{\mathbf{v}_k\}_k) \le F_A(\mathbf{x}) + \frac{1}{2\tau K}\Sigma_k \|\mathbf{v}_k - \mathbf{A}\mathbf{x}\|^2$ (consensus violation)



Case 1: Strongly Convex Objective => Linear Convergence Rate

Theorem (Strongly Convex g_i). *Let g_i be μ_g-strongly convex and let f be $1/\tau$-smooth. After T iterations with*

$$T \ge C_1(\beta,\mu_g,\tau,\gamma,\Theta,\sigma')\log\left(\frac{C_2(\beta,\mu_g,\tau,\gamma,\Theta,\sigma',\epsilon^{(0)})}{\epsilon}\right)$$

the expected duality gap $\mathbb{E}[G_{\mathcal{H}}(\mathbf{x}^{(T)}, \{\sum_{k=1}^K \mathcal{W}_{kl}\mathbf{v}_l^{(T)}\}_{k=1}^K)] \le \epsilon$.

Case 2: General Convex Objective => The decentralized duality gap converges in 1/T rate in expectation

Local Certificates

Theorem (Local Certificates). *Assume g_i has L-bounded support, and let $\mathcal{N}_k := \{j : \mathcal{W}_{jk} > 0\}$ be the set of nodes accessible to node k. Then for any given $\varepsilon > 0$, we have*

$$G_{\mathcal{H}}(\mathbf{x}; \{\mathbf{v}_k\}_{k=1}^K) \le \varepsilon,$$

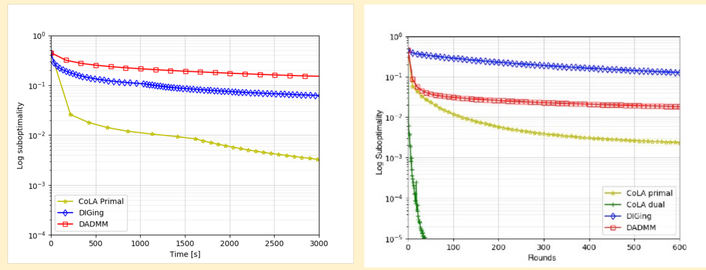
if for all $k = 1,\dots,K$ the following two local conditions are satisfied:

$$\mathbf{v}_k^\top \nabla f(\mathbf{v}_k) + \sum_{i\in\mathcal{P}_k}\left(g_i(\mathbf{x}_i) + g_i^*(-\mathbf{A}_i^\top \nabla f(\mathbf{v}_k))\right) \le \frac{\varepsilon}{2K}$$

$$\left\|\nabla f(\mathbf{v}_k) - \tfrac{1}{|\mathcal{N}_k|}\textstyle\sum_{j\in\mathcal{N}_k}\nabla f(\mathbf{v}_j)\right\|_2 \le \left(\textstyle\sum_{k=1}^K n_k^2\sigma_k\right)^{-1/2}\frac{1-\beta}{2L\sqrt{K}}\varepsilon,$$

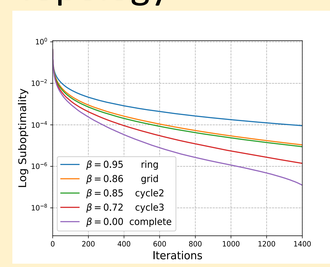
Experiments

Convergence



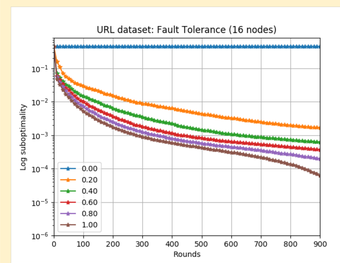
Comparing COLA with DIGing [1] and DADMM [2] on a ring of 16 nodes. (Ridge Regression + URL dataset)

Topology



Run LASSO on RCV1 dataset. All of the graphs have 16 nodes.

Unreliable Nodes



In a round, only certain percent of nodes remain in the network.

References

[1] Nedic et al. 16' Achieving geometric convergence for distributed optimization over time-varying graphs.
[2] Shi et al. 14' Extra: An exact first-order algorithm for decentralized consensus optimization.
[3] Jaggi et al. 14' Communication-efficient distributed dual coordinate ascent