hetFL

shamsiiat.abdurakhmanova

February 12, 2024

1 Numerical experiments

1.1 Datasets

1.1.1 Synthetic Dataset

Experiments were performed on a synthetic dataset whose empirical graph $\mathcal{G}$ is partitioned into 3 equal-sized clusters $\mathcal{P} = \{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \mathcal{C}^{(3)}\}$, with $|\mathcal{C}^{(1)}| = |\mathcal{C}^{(2)}| = |\mathcal{C}^{(3)}|$. We denote the cluster assignment of node $i \in \mathcal{V}$ by $c^{(i)} \in \{1, 2, 3\}$.

For experiments where graph connectivity is known, the edges in $\mathcal{G}$ are generated via realizations of independent binary random variables $b_{i,i'} \in \{0, 1\}$. These random variables are indexed by pairs $i, i^{'}$ of nodes that are connected by an edge $\{i, i^{'}\} \in \mathcal{E}$ if and only if $b_{i,i'} = 1$.
Two nodes in the same cluster are connected with probability $Prob\{b_{i,i'} = 1\} := p_{in}$ if nodes i, i' belong to the same cluster. In contrast, $Prob\{b_{i,i'} = 1\} := p_{out}$ if nodes $i, i^{'}$ belong to different clusters. Every edge in $\mathcal{G}$ has the same weight, $A_e = 1$ for all $e \in \mathcal{E}$.

Each node $i \in \mathcal{V}$ of the empirical graph $\mathcal{G}$ holds a local dataset $\mathcal{D}^{(i)}$ of the form $\mathcal{D}^{(i)} := \{(\mathbf{x}^{(i,1)}, y^{(i,1)}), ..., (\mathbf{x}^{(i,m_i)}, y^{(i,m_i)})\}$. Thus, dataset $\mathcal{D}^{(i)}$ consist of m_i data points, each characterized by a feature vector $\mathbf{x}^{(i,r)} \in \mathbb{R}^d$ and scalar label $y^{(i,r)}$, for $r = 1, ..., m_i$. The feature vectors $\mathbf{x}^{(i,r)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$, are drawn i.i.d. from a standard multivariate normal distribution.

The labels of the data points are generated by a noisy linear model

$$y^{(i,r)} = (\mathbf{w}^{(i)})^T \mathbf{x}^{(i,r)} + \varepsilon^{(i,r)} \tag{1}$$

The noise $\varepsilon^{(i,r)} \sim \mathcal{N}(0, 1)$, for $i \in \mathcal{V}$ and $r = 1, .., m_i$, are i.i.d. realizations of a normal distribution. The true underlying vector $\overline{\mathbf{w}}^{(i)} \sim \mathcal{N}(0, 1)$ is drawn from a standard normal distribution and is the same for nodes from the same cluster, i.e. $\overline{\mathbf{w}}^{(i)} = \overline{\mathbf{w}}^{(i')}$ if $c^{(i)} = c^{(i')}$.

Datasets are divided into training and validation subsets by using resampling with replacement. The size of the validation subset is $m_i^{(val)} = 100$.

1.1.2 Shared Dataset

Dataset $\mathcal{D}^{(test)}$, which predictions are shared across all nodes was formed as follows: the feature, weight and noise vectors are drawn i.i.d. from a standard normal distribution and labels are generated by a noisy linear model. The size of the dataset is $m' = 100$.

1.2 Experiments

1.2.1 Synthetic Dataset, linreg model, graph is known

In these experiments empirical graph $\mathcal{G}$ consist of $N = 15$ nodes partitioned into three clusters ($|\mathcal{C}^{(i)}| = 5$). Two nodes in the same cluster are connected with probability $p_{in} = 0.8$ if nodes $i, i^{'}$ belong to the same cluster and $p_{out} = 0.2$ if nodes $i, i^{'}$ belong to different clusters.

Each node $i \in \mathcal{V}$ of the empirical graph $\mathcal{G}$ holds a local dataset $\mathcal{D}^{(i)}$ consisting of m_i data points, each characterized by a feature vector $\mathbf{x}^{(i,r)} \in \mathbb{R}^d$ and scalar label $y^{(i,r)}$, for $r = 1, ..., m_i$, where $d = 10$. The sample size of the shared dataset $\mathcal{D}^{(test)}$ is $m' = 100$.

To learn the local parameters $\mathbf{w}^{(i)}$, we use Algorithm X with local loss

$$L_{(i)}(h^{(i)}) = \frac{1}{m_i} \sum_{r=1}^{m_i} \Big(y^{(i,r)} - h^{(i)}(\mathbf{x}^{(i,r)}) \Big)^2 \tag{2}$$

and regularizer

$$\frac{\lambda}{2m'} \sum_{i' \in \mathcal{V}} A_{i,i'} \sum_{r=1}^{m'} \Big(h^{(i)}(\mathbf{x}^{(r)}) - h^{(i^{'})}(\mathbf{x}^{(r)}) \Big)^2 \tag{3}$$

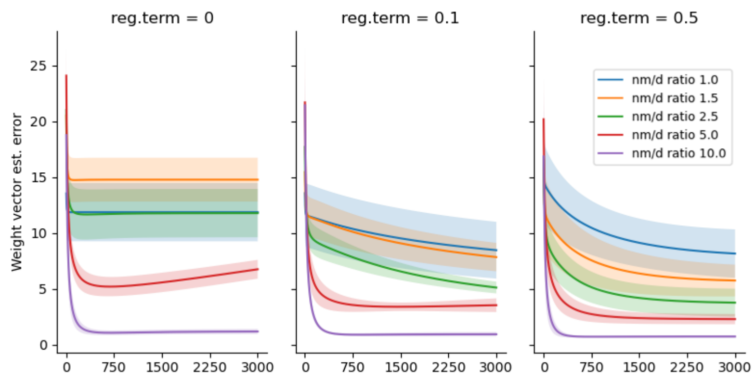
For the experiment we use linear model implemented with pytorch (no bias, optimizer SGD learning rate 0.01). We try different local dataset sizes $m_i = \{2,\ 3,\ 5,\ 10,\ 20\}$ and regularization strength $\lambda = \{0,\ 0.1,\ 0.5\}$. Tried ratio $|\mathcal{C}^{(i)}| m_i/d = \{1, 1.5, 2.5, 5, 10\}$ or $d/m_i = \{5, 3.3, 2, 1, 0.5\}$

As stopping criterion in Algorithm 2, we use a fixed number of R = 3000 iterations. For pytorch models one iteration is equivalent to one gradient step.

Below is the plot of mean estimation error (mean over 10 repetitions of the experiment for each pair of $\{\lambda, m_i\}$).

$$\frac{1}{N} \sum_{i=1}^{N} ||\overline{\mathbf{w}}^{(i)} - \widehat{\mathbf{w}}^{(i)}||_2^2 \tag{4}$$

On each repetition new local $\mathcal{D}^{(i)}$ and shared $\mathcal{D}^{(test)}$ datasets were generated. The shaded region is $\pm$ one standard deviation.



reg.term = 0
reg.term = 0.1
reg.term = 0.5
Weight vector est. error
0
5
10
15
20
25
0
750
1500
2250
3000
nm/d ratio 1.0
nm/d ratio 1.5
nm/d ratio 2.5
nm/d ratio 5.0
nm/d ratio 10.0

Algorithm 1 Least-Square Regression (Adjacency matrix is known)

Input: empirical graph $\mathcal{G}$ with edge weights A_{ij}; local loss $L_{(i)}(\cdot)$; shared dataset $D^{(test)} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(m')}\}$; GTV parameter λ;

Initialize: $k := 0; \widehat{h}_0^{(i)} \equiv$ for all nodes $i \in \mathcal{V}$.

while stopping criterion is not met do **do**

  for all nodes $i \in \mathcal{V}$ in parallel **do**

    share predictions $\{\widehat{h}_k^{(i)}(\mathbf{x})\}_{\mathbf{x} \in \mathcal{D}^(test)}$, with neighbours $i' \in \mathcal{N}^{(i)}$

    update local hypothesis $\widehat{h}_k^{(i)}$ by

$$\widehat{h}_{k+1}^{(i)} \in \arg \min_{h^{(i)} \in \mathcal{H}^{(i)}} \Bigg[\frac{1}{m_i} \sum_{r=1}^{m_i} \Big(y^{(i,r)} - h^{(i)}(\mathbf{x}^{(i,r)}) \Big)^2 + \frac{\lambda}{m'} \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} \sum_{r=1}^{m'} \Big(h^{(i)}(\mathbf{x}^{(r)}) - \widehat{h}_k^{(i')}(\mathbf{x}^{(r)}) \Big)^2 \Bigg]$$

  end for

  $k := k + 1$

end while

Ensure: local $\widehat{h}^{(i)} := \widehat{h}_{k+1}^{(i)}$ for all nodes $i \in \mathcal{V}$

1.2.2 Synthetic Dataset, linreg model, graph is not known

Algorithm 2 Least-Square Regression (Adjacency matrix is not known)

Input: empirical graph $\mathcal{G}$; local loss $L_{(i)}(\cdot)$; shared dataset $D^{(test)} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(m')}\}$; GTV parameter λ;

Initialize: $k := 0; A := 0; \widehat{h}_0^{(i)} \equiv$ for all nodes $i \in \mathcal{V}$.

while stopping criterion is not met do **do**

  for all nodes $i \in \mathcal{V}$ in parallel **do**

    share predictions $\{\widehat{h}_k^{(i)}(\mathbf{x})\}_{\mathbf{x} \in \mathcal{D}^{(test)}}$, with neighbours $i' \in \mathcal{N}^{(i)}$

    update local hypothesis $\widehat{h}_k^{(i)}$ by

$$\widehat{h}_{k+1}^{(i)} \in \arg \min_{h^{(i)} \in \mathcal{H}^{(i)}} \Bigg[\frac{1}{m_i} \sum_{r=1}^{m_i} \Big(y^{(i,r)} - h^{(i)}(\mathbf{x}^{(i,r)}) \Big)^2 + \frac{\lambda}{m'} \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} \sum_{r=1}^{m'} \Big(h^{(i)}(\mathbf{x}^{(r)}) - \widehat{h}_k^{(i')}(\mathbf{x}^{(r)}) \Big)^2 \Bigg]$$

  end for

  $k := k + 1$

  for all nodes $i \in \mathcal{V}$ in parallel **do**

    find p-neighbours for the node by selecting nodes with p-smallest values of

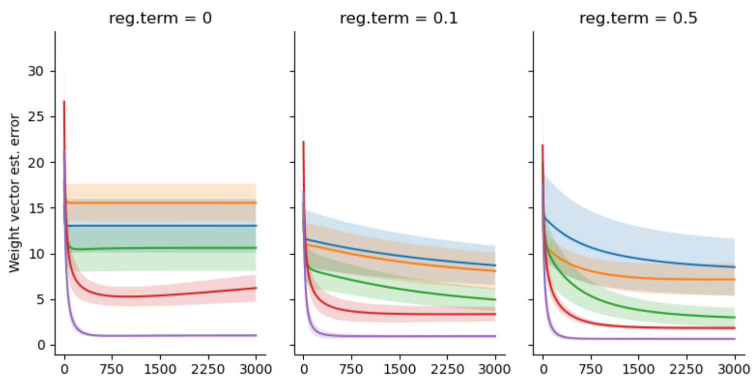
$$\frac{1}{m'} \sum_{r=1}^{m'} \Big(h^{(i)}(\mathbf{x}^{(r)}) - \widehat{h}_k^{(i')}(\mathbf{x}^{(r)}) \Big)^2 \tag{5}$$

  end for

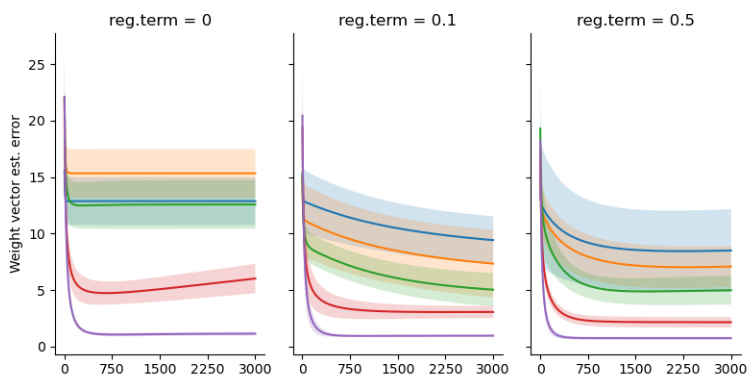
end while

Ensure: local $\widehat{h}^{(i)} := \widehat{h}_{k+1}^{(i)}$ for all nodes $i \in \mathcal{V}$

number of neighbors p=3.



number of neighbors p=5



1.2.3 TODO Synthetic Dataset, models of mixed type, graph is not known