MS&E 111X/MS&E211X Suggested Course Project IV: Data-sharing in Predictions Algorithms

November 15, 2021

In this project, we consider the case where each data center i possess s number of observations with same features of dimension p. A decision maker tries to perform estimation algorithms over data sets $(\mathbf{X}_i = [\mathbf{x}_{i,1}; \ldots; \mathbf{x}_{i,j}] \in R^{s\times p}, \mathbf{y}_i = [y_{i,1}; \ldots, y_{i,j}] \in R^{s\times 1})$ across each center $i \in \{1, \ldots, b\}$.

For example, consider there are 2 data centers $b = 2$, and $i \in \{1, 2\}$, each of the data center possess $s = 2$ number of observations with feature space $p = 2$, we have

$$\mathbf{X}_1 = \begin{bmatrix} \mathbf{x}_{1,1} \\ \mathbf{x}_{1,2} \end{bmatrix} = \begin{bmatrix} 0.75, & 0.01 \\ 0.65, & 0.80 \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} \mathbf{x}_{2,1} \\ \mathbf{x}_{2,2} \end{bmatrix} = \begin{bmatrix} 0.76, & 0.02 \\ 0.63, & 0.79 \end{bmatrix} \tag{1}$$

$$\mathbf{y}_1 = \begin{bmatrix} y_{1,1} \\ y_{1,2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{y}_2 = \begin{bmatrix} y_{2,1} \\ y_{2,2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{2}$$

The goal of the decision maker is to find predictor $\boldsymbol{\beta} \in R^{p\times 1}$ over all data sets. In order to find $\boldsymbol{\beta}$, we consider the following optimization problem

$$\sum_{i=1}^{b} \sum_{j=1}^{s} f((\mathbf{x}_{i,j}, y_{i,j}); \boldsymbol{\beta}) \tag{3}$$

When $f((\mathbf{x}_{i,j}, y_{i,j}); \boldsymbol{\beta}) = \frac{1}{2}(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{y}_i)^T(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{y}_i)$, the problem becomes an unconstrained quadratic optimization

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{b} \frac{1}{2}(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{y}_i)^T(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{y}_i) = \sum_{i=1}^{b} \sum_{j=1}^{s} \frac{1}{2}(\mathbf{x}_{i,j}\boldsymbol{\beta} - y_{i,j})^2 \tag{4}$$

Question 1: Optimal solution under full access of data

If the decision maker could have access to share $\mathbf{X}_i$ across all data centers and have access to $\mathbf{X} = [\mathbf{X}_1; \ldots; \mathbf{X}_b]$, and $\mathbf{y} = [\mathbf{y}_1; \ldots; \mathbf{y}_b]$. Write out the optimality condition and verify that the minimizor of (4)

is given by

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Hint: The original problem is equivalent as

$$\min_{\boldsymbol{\beta}} \ \frac{1}{2}\boldsymbol{\beta}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$$

Question 2: Gradient Descend Method

Now suppose that each data center i does not want to provide $\mathbf{X}_i \in R^{s\times p}$ to the decision maker, because $\mathbf{X}_i$ may contain some sensitive information. Instead, the data center allows the decision maker to run gradient descend method with the data set.

Question 2.1 In order to run gradient descent algorithm on across data centers $i \in \{1, \ldots, b\}$, at the k^{th} iteration, the decision maker need to gather the gradient information at each center in order to update

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \rho\mathbf{g}_k$$

where ρ is the step size and $\mathbf{g}_k$ is the gradient evaluated at $\boldsymbol{\beta}^k$. Provide the expression of $\mathbf{g}_k$ as a function of $\mathbf{X}_i$ and $\mathbf{y}_i$.

Hint: $\mathbf{g}_k = \sum_{i=1}^b \mathbf{g}_k^i$, *where* $\mathbf{g}_k^i \in R^{p\times 1}$ *only depends on* $\mathbf{X}_i$ *and* $\mathbf{y}_i$, *you may also find more hints in question 2.*

Question 2.2 Run gradient descent method on the data sets $\mathbf{X}$ and $\mathbf{y}$ provided in (1), (2), fix the starting point $\boldsymbol{\beta}^0 = [0; \ldots; 0]$, and the step size ρ to be constant with $\rho = 1$, report the Euclidean norm $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \sqrt{\sum_{i=1}^p (\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i)^2}$ after 10, 20, 30 number of iterations respectively. Does the algorithm converge? For what range of step size does gradient descent algorithm converge?

Question 3: Primal Distributed ADMM Method

We could also formulate the problem as constrained optimization by introducing local estimators $\boldsymbol{\beta}_i$ to each data centers.

$$\min_{\boldsymbol{\beta}_i, \boldsymbol{\beta}} \sum_{i=1}^{b} \frac{1}{2}(\mathbf{X}_i\boldsymbol{\beta}_i - \mathbf{y}_i)^T(\mathbf{X}_i\boldsymbol{\beta}_i - \mathbf{y}_i) \tag{5}$$

$$s.t. \boldsymbol{\beta}_i - \boldsymbol{\beta} = 0 \qquad \forall\ i \tag{6}$$

Question 3.1 Write down the augmented Lagragian of problem (5).

The algorithm of primal consensus ADMM is given as follows The algorithm of primal consensus ADMM is as follows.

Algorithm 1 Primal Consensus ADMM

Initialization: $t = 0$, step size $\rho \in R^+$ $\boldsymbol{\beta}_t \in R^p$, $\boldsymbol{\lambda}_{t,i} \in R^p$, $\boldsymbol{\beta}_{t,i} \in R^p$ for all $i \in \{1, \ldots, b\}$, and stopping rule τ

while $t \leq \tau$ **do**

Each data center i updates $\boldsymbol{\beta}_{t+1,i}$in parallel by

$\boldsymbol{\beta}_{t+1,i} = argmin_{\boldsymbol{\beta}_i \in R^p} \sum_i \sum_j f((\mathbf{x}_{i,j}, y_{i,j}); \boldsymbol{\beta}_i) + \boldsymbol{\lambda}_{t,i}^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}^t) + \frac{\rho}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_t)^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_t)$

Decision maker updates

$\boldsymbol{\beta}_{t+1} = \frac{1}{b}\sum_j \boldsymbol{\beta}_{t+1,i} + \frac{1}{b\rho}\sum_i \boldsymbol{\lambda}_{t,i}$, $\boldsymbol{\lambda}_{t+1,i} = \boldsymbol{\lambda}_{t,i} + \rho(\boldsymbol{\beta}_{t+1,i} - \boldsymbol{\beta}_{t+1})$

end

Output: $\boldsymbol{\beta}_\tau$ as global estimator

Question 3.2 Run primal consensus ADMM on the data sets $\mathbf{X}$ and $\mathbf{y}$ provided in (1), (2), fix the starting point $\boldsymbol{\beta}^0 = [0; \ldots; 0]$, and the step size ρ to be constant with $\rho = 1$, report the Euclidean norm $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \sqrt{\sum_{i=1}^p (\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i)^2}$ after 10, 20, 30 number of iterations respectively. Compare your algorithm with gradient descent under $\rho = 1$.

Question 5: Benefit of Data Exchange

Now suppose we can swap one entry of the observations. Considering the following data sets

$$\hat{\mathbf{X}}_1 = \begin{bmatrix} \mathbf{x}_{1,1} \\ \mathbf{x}_{2,1} \end{bmatrix} = \begin{bmatrix} 0.75, & 0.01 \\ 0.76, & 0.02 \end{bmatrix}, \hat{\mathbf{X}}_2 = \begin{bmatrix} \mathbf{x}_{1,2} \\ \mathbf{x}_{2,2} \end{bmatrix} = \begin{bmatrix} 0.65, & 0.80 \\ 0.63, & 0.79 \end{bmatrix} \tag{7}$$

$$\hat{\mathbf{y}}_1 = \begin{bmatrix} y_{1,1} \\ y_{2,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \hat{\mathbf{y}}_2 = \begin{bmatrix} y_{1,2} \\ y_{2,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{8}$$

Under this scenario, we exchange the observation $\mathbf{x}_{1,2}$ at data center 1 and $\mathbf{x}_{2,1}$ at data center 2. Note this operation would not influence the optimal solution of $\boldsymbol{\beta}^*$ (why?). Run primal consensus ADMM on the data sets $\hat{\mathbf{X}}$ and $\hat{\mathbf{y}}$ provided in (7), (8), fix the starting point $\boldsymbol{\beta}^0 = [0; \ldots; 0]$, and the step size ρ to be constant with $\rho = 1$, report the Euclidean norm $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \sqrt{\sum_{i=1}^p (\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i)^2}$ after 10, 20, 30 number of iterations respectively. Compare your algorithm with results you had in Question 4.2.

Question 6 (Bonus) : Data sets of UCI Machine Learning Repository

Now consider the real world data sets. The following data set on YearPredictionMSD[1] tries to predict of the release year of a song from audio features. Songs are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s. We focus on the trainning data with $463,715$ number of observations and feature space $p = 90$. Now suppose the data are provided by four audio/entertainment companies, company 1 possesses observations 1 to $115,929$, company 2 possesses observations of $115,930$ to

[1] https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD

231, 858, company 3 possesses observations of 231, 859 to 347, 787 and company 4 possesses observations of 347, 788 to 463, 715.

The four companies do not want to directly share $\mathbf{X}_i$ with you, instead, all of the four companies allows you to run either gradient descend method or ADMM method on the data sets. Compare the gradient descend algorithm, primal consensus ADMM algorithm.

In order to apply data-exchange, consider the following algorithm. Introducing the auxiliary ζ, we have the primal problem could also be formulated as

$$\min_{\boldsymbol{\zeta}} \frac{1}{2}\boldsymbol{\zeta}^T\boldsymbol{\zeta} \tag{9}$$

$$s.t. \mathbf{X}\boldsymbol{\beta} - \mathbf{y} = \boldsymbol{\zeta} \tag{10}$$

And let $\mathbf{t}$ be the dual variables with respect to the primal constraints $\mathbf{X}\boldsymbol{\beta} - \mathbf{y} = \boldsymbol{\zeta}$. Taking the dual with respect to problem (9). Consider the following randomized cyclic updating method on Dual-Randomized-Cyclic ADMM (DRC-ADMM). Let $L(\mathbf{t}, \boldsymbol{\beta})$ be the augmented Lagrangian, at each iteration k, do the following

- random permute $[1, \ldots, n]$ to be $\boldsymbol{\sigma}$. For example with $n = 4$, one random permutation may be $\boldsymbol{\sigma} = [4, 1, 2, 3]$.
- Given number of data center b, separates the data according to $\boldsymbol{\sigma}$. For example, with $n = 4$, and random permutation $\boldsymbol{\sigma} = [4, 1, 2, 3]$, $b = 2$, $\mathbf{t}_{1,\boldsymbol{\sigma}} = [t_4, t_1]$, and $\mathbf{t}_{2,\boldsymbol{\sigma}} = [t_2, t_3]$.
- Cyclic updating

$$\mathbf{t}_{1,\sigma}^{k+1} = argmin_{\mathbf{t}_{1,\boldsymbol{\sigma}}} L(\mathbf{t}_{1,\boldsymbol{\sigma}}, \mathbf{t}_{2,\sigma}^{k}, \boldsymbol{\beta}^k)$$

$$\mathbf{t}_{2,\sigma}^{k+1} = argmin_{\mathbf{t}_{2,\boldsymbol{\sigma}}} L(\mathbf{t}_{1,\boldsymbol{\sigma}}^{k+1}, \mathbf{t}_{2,\sigma}, \boldsymbol{\beta}^k)$$

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \rho\mathbf{X}^T\mathbf{t}^{k+1}$$

Run the Dual-Randomized-Cyclic ADMM (DRC-ADMM) on the data sets on $\mathbf{X}$ and $\mathbf{y}$ for UCI ML data, fix the starting point $\boldsymbol{\beta}^0 = [0; \ldots; 0]$, and the step size ρ to be constant with $\rho = 1$, report the Euclidean norm $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \sqrt{\sum_{i=1}^{p}(\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i)^2}$ after 10, 20, 30 number of iterations respectively. Comparing with algorithms without data exchange, what do you find? Now what if you can only permute 5% of the data?

References

[1] Chen, C., He, B., Ye, Y., and Yuan, X. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Mathematical Programming, 155(1-2), 57-79.

[2] Sun, R., Luo, Z. Q., and Ye, Y. On the efficiency of random permutation for admm and coordinate descent. Mathematics of Operations Research, 45(1), 233-271.

[3] Mihić, K., Zhu, M., and Ye, Y. Managing randomization in the multi-block alternating direction method of multipliers for quadratic optimization. Mathematical Programming Computation, 13(2), 339-413.

[4] Zhu, M., and Ye, Y. Benefit of Data Sharing in Multi-Block Alternating Direction Method of Multipliers Algorithm for Regression Estimation

[5] Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., Zhou, Z. Federated Learning's Blessing: FedAvg has Linear Speedup. arXiv preprint arXiv:2007.05690.

[6] Dua, D. and Graff, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.