DEMYSTIFYING MODEL AVERAGING FOR COMMUNICATION-EFFICIENT FEDERATED MATRIX FACTORIZATION

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Shuai Wang[1,2], **Richard Cornelius Suwandi**[2], **Tsung-Hui Chang**[1,2]

[1] Shenzhen Research Institute of Big Data, [2] The Chinese Univ. of Hong Kong, Shenzhen

SRIBD

Summary

The technique of model averaging (MA) has not been considered for the important matrix factorization (MF) model under the scenario of federated learning (FL).

- Propose a new MA based algorithm, named FedMAvg, by judiciously combining the alternating minimization technique and MA.
- Local GD with diminishing steps and partial client communication can greatly reduce the communication cost, even under non-i.i.d. data.

Federated Matrix Factorization Model

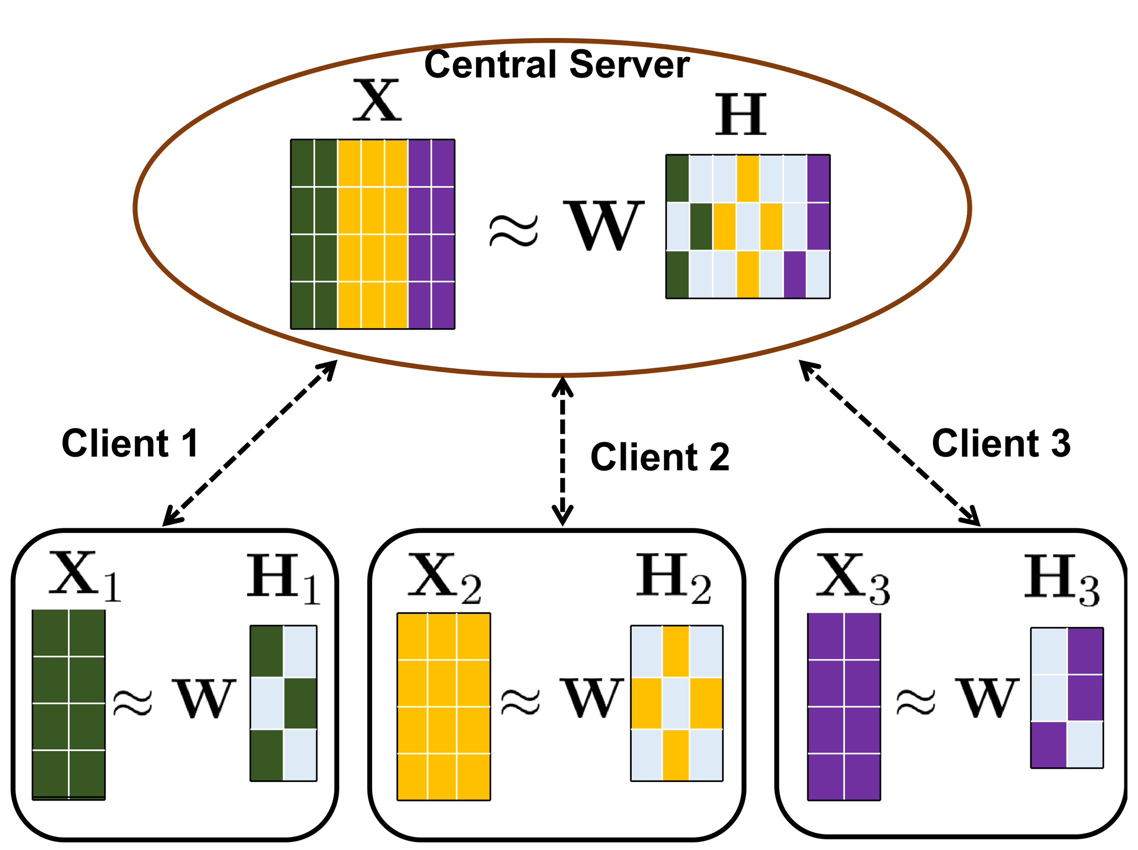
The data samples are partitioned as $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_P]$ and respectively owned by P distributed clients. Each client p owns non-overlapping data $\boldsymbol{X}_p \in \mathbb{R}^{M \times N_p}$, where N_p is the number of samples of client p and $\sum_{p=1}^{P} N_p = N$.

$$\min_{\substack{\boldsymbol{W},\, \boldsymbol{H}_p,\\ p=1,\ldots,P}} \; F(\boldsymbol{W}, \boldsymbol{H}) \triangleq \sum_{p=1}^{P} \omega_p F_p(\boldsymbol{W}, \boldsymbol{H}_p) \tag{1a}$$

$$\text{s.t. } \boldsymbol{W} \in \mathcal{W}, \boldsymbol{H}_p \in \mathcal{H}_p, \forall p \in \mathcal{P}, \tag{1b}$$

where $F_p(\boldsymbol{W}, \boldsymbol{H}_p) = \frac{1}{N_p} \Phi_p(\boldsymbol{X}_p, \boldsymbol{W}\boldsymbol{H}_p)$, $p \in \mathcal{P}$.

- $\Phi_p(\boldsymbol{X}_p, \boldsymbol{W}\boldsymbol{H}_p)$ measures the quality of the approximation $\boldsymbol{X}_p \approx \boldsymbol{W}\boldsymbol{H}_p$, e.g. $\frac{1}{N_p}\|\boldsymbol{X}_p - \boldsymbol{W}\boldsymbol{H}_p\|_F^2$.
- P could be large, N_p, $p = 1, \ldots, P$, could be unbalanced, and $\boldsymbol{X}_p$, $p \in \mathcal{P}$ could be non-i.i.d.
- Problem (1) is challenging to solve since it is non-convex and non-smooth, and involves two blocks of variables $\boldsymbol{W}$ and $\boldsymbol{H}$.



Algorithm Development

- **Alternating Minimization:**
  Given $\boldsymbol{W}^{s-1}$, each client p performs

$$\boldsymbol{H}_p^s = \arg\min_{\boldsymbol{H}_p \in \mathcal{H}_p} \; F_p(\boldsymbol{W}^{s-1}, \boldsymbol{H}_p), \tag{2a}$$

$$\boldsymbol{W}_p^s = \arg\min_{\boldsymbol{W}} \; F_p(\boldsymbol{W}, \boldsymbol{H}_p^s). \tag{2b}$$

The server does $\boldsymbol{W}^s = \mathcal{P}_{\mathcal{W}}(\sum_{p=1}^{P} \omega_p \boldsymbol{W}_p^s)$.

- **Local GD with Diminishing Q_2:**
  - (2a) via $Q_1 \geq 1$ consecutive steps of PGD with respect to $\boldsymbol{H}_p$.
  - (2b) via $Q_2^s \geq 1$ ($Q_2^s = \lfloor \frac{\hat{Q}}{s} \rfloor + 1$) consecutive steps of GD with respect to $\boldsymbol{W}_p$.
- **Partial Client Communication (PCC):**
  - For each round, m clients in $\mathcal{A}^s$ are selected by the server.
  - All clients perform updating but only the clients in $\mathcal{A}^s$ upload their models to the server for averaging.

Proposed FedMAvg Method

Algorithm 1 Proposed FedMAvg algorithm

Input: initial values of $\mathbf{W}_1^0 = \cdots = \mathbf{W}_P^0$ at the server side, initial values of $\{\mathbf{H}_p^0\}_{p=1}^P$ at the clients, $\mathcal{A}^0 = \{1, \ldots, P\}$ and $\hat{Q}$.
for round $s = 1$ **to** S **do**
  Server side: Compute

$$\mathbf{W}^s = \mathcal{P}_{\mathcal{W}}\left(\frac{1}{m} \sum_{p \in \mathcal{A}^{s-1}} \mathbf{W}_p^{s-1}\right),$$

  and select a set of clients $\mathcal{A}^s$ (with size $|\mathcal{A}^s| = m$) by sampling with replacement according to probabilities $\{\omega_1, \ldots, \omega_P\}$, and broadcast $\mathbf{W}^s$ to all clients.
  Client side:
  for client $p = 1$ **to** P in parallel **do**
    Set $\mathbf{H}_p^{s,0} = \mathbf{H}_p^{s-1}$ and $\mathbf{W}_p^{s,0} = \mathbf{W}^s$.
    for epoch $t = 1$ **to** Q_1 **do**

$$\mathbf{H}_p^{s,t} = \mathcal{P}_{\mathcal{H}_p}\left(\mathbf{H}_p^{s,t-1} - \frac{\nabla_{H_p} F_p(\mathbf{W}_p^{s,t-1}, \mathbf{H}_p^{s,t-1})}{c_p^s}\right)$$

$$\mathbf{W}_p^{s,t} = \mathbf{W}_p^{s,t-1}.$$

    end for
    for epoch $t = Q_1 + 1$ **to** $Q^s = Q_1 + Q_2^s$ **do**

$$\mathbf{W}_p^{s,t} = \mathbf{W}_p^{s,t-1} - \frac{\nabla_W F_p(\mathbf{W}_p^{s,t-1}, \mathbf{H}_p^{s,t-1})}{d^s},$$

$$\mathbf{H}_p^{s,t} = \mathbf{H}_p^{s,t-1}.$$

    end for
    Denote $\mathbf{W}_p^s = \mathbf{W}_p^{s,Q^s}$ and $\mathbf{H}_p^s = \mathbf{H}_p^{s,Q^s}$.
    if client $p \in \mathcal{A}^s$ **then**
      Upload $\mathbf{W}_p^s$ to the server.
    end if
  end for
end for

Convergence Analysis

Bounds:

$$\|\nabla_W F_p(\boldsymbol{W}, \boldsymbol{H}_p) - \nabla_W F(\boldsymbol{W}, \boldsymbol{H})\|_F^2 \leq \zeta^2, \tag{3}$$

$$\|\nabla_W F(\boldsymbol{W}, \boldsymbol{H})\|_F^2 \leq \phi^2, \tag{4}$$

Virtual Sequences: $\forall t = 1, \ldots, Q$,

$$\widetilde{\boldsymbol{W}}^{s,t} = \mathcal{P}_{\mathcal{W}}\left(\frac{1}{m} \sum_{p \in \mathcal{A}^s} \boldsymbol{W}_p^{s,t}\right), \; \widetilde{\boldsymbol{W}}^{s,0} = \boldsymbol{W}^s, \tag{5}$$

Proximal Gradient:

$$G_H^{s,t} \triangleq \sum_{p=1}^{P} \omega_p (c_p^s)^2 \Big\| \boldsymbol{H}_p^{s,t} - \mathcal{P}_{\mathcal{H}_p}\big(\boldsymbol{H}_p^{s,t} - (c_p^s)^{-1} \nabla_{H_p} F_p(\widetilde{\boldsymbol{W}}^{s,t}, \boldsymbol{H}_p^{s,t})\big)\Big\|_F^2, \; \forall t \in \mathcal{Q}_1, \tag{6}$$

$$G_W^{s,t} \triangleq (d^s)^2 \|\widetilde{\boldsymbol{W}}^{s,t} - \mathcal{P}_{\mathcal{W}}(\widetilde{\boldsymbol{W}}^{s,t} - (d^s)^{-1} \nabla_W F(\widetilde{\boldsymbol{W}}^{s,t}, \boldsymbol{H}^{s,t}))\|_F^2, \; \forall t \in \mathcal{Q}_2^s, \tag{7}$$

Theorem 1 *Let $Q_2^s = \lfloor \frac{\hat{Q}}{s} \rfloor + 1$, and let T be the total number of iterations. Moreover, let $c_p^s = \frac{\gamma_1}{2} L_{H_p}^s$, $d^s = \gamma_2 L_W^s$, where $\gamma_1 > 1$ and $\gamma_2 \geq Q_2^1 \sqrt{2(7 + 4\overline{L}_W^2 / \underline{L}_W^2)}$. Then, under Assumptions, the sequence $\{(\widetilde{\boldsymbol{W}}^{s,t}, \boldsymbol{H}^{s,t})\}$ satisfies*

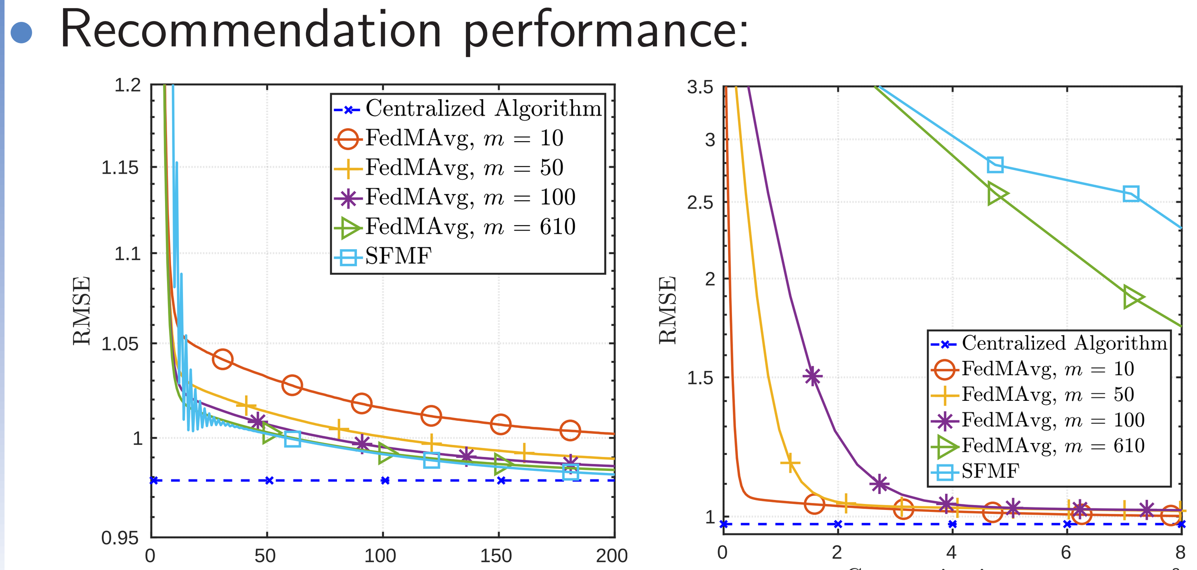
$$\begin{aligned}
&\frac{1}{T}\left[\sum_{s=1}^{S}\sum_{t=1}^{Q_1} \mathbb{E}[G_H^{s,t-1}] + \sum_{s=1}^{S}\sum_{t=Q_1+1}^{Q^s} \mathbb{E}[G_W^{s,t-1}]\right] \\
\leq &\frac{D}{T}\left(F(\widetilde{\boldsymbol{W}}^{1,0}, \boldsymbol{H}^{1,0}) - \underline{F}\right) + \left(\frac{8D\zeta^2}{m\gamma_2 \underline{L}_W} + \frac{96\zeta^2}{m}\right) \\
&+ \frac{2D(1 + 8/m)(\frac{11}{3}\zeta^2 + \phi^2)\sum_{s=1}^{S} C_1^s}{T\gamma_2^3 \underline{L}_W} \\
&+ \frac{(\frac{11}{3}\zeta^2 + \phi^2)\sum_{s=1}^{S} C_2^s}{T\gamma_2^2} + \frac{3(\zeta^2 + \phi^2)\sum_{s=1}^{S} C_1^s}{2T},
\end{aligned} \tag{8}$$

where $D \triangleq \frac{\gamma_1^2 \overline{L}_H}{2(\gamma_1 - 1)} + \frac{6(\gamma_2^2 + 1)\overline{L}_W^2}{(\gamma_2 - 1)\underline{L}_W}$, $C_1^s \triangleq Q_2^s(Q_2^s - 1)(2Q_2^s - 1)$, *and* $C_2^s \triangleq 6(3Q_2^s(Q_2^s - 1)/2 + 4 + 32/m)C_1^s$.

Numerical Results II

Application to Item Recommendation:

- Recommendation performance:

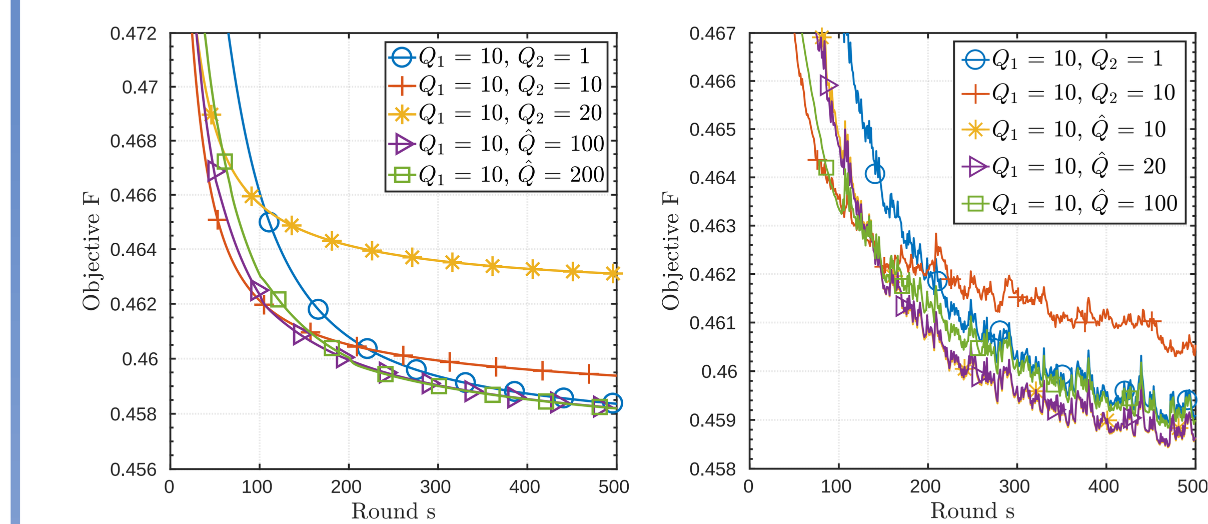


(a) (b)

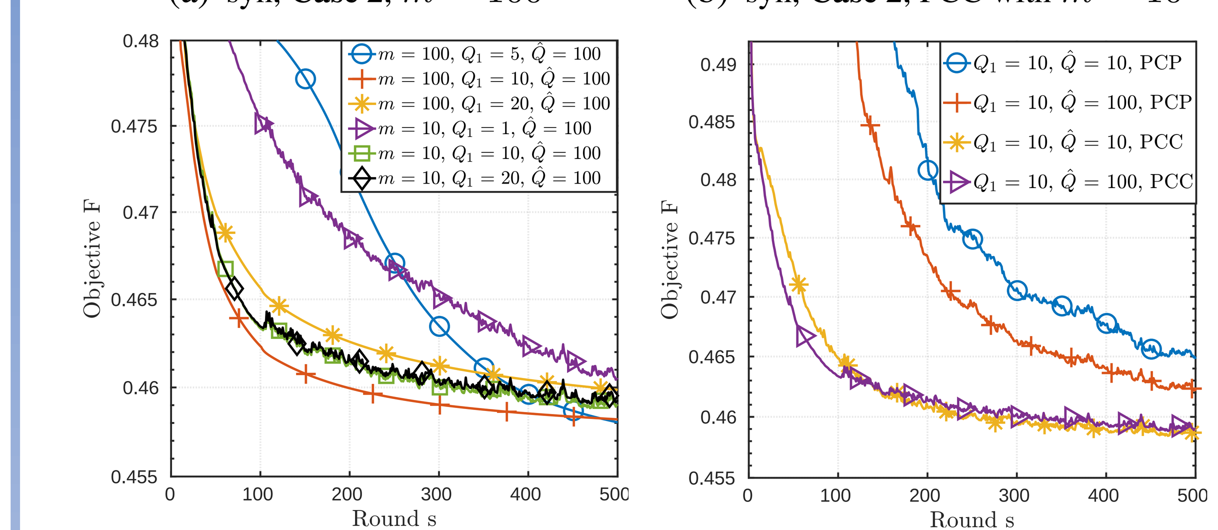
Numerical Results I

Application to Data Clustering:

- Convergence behavior:

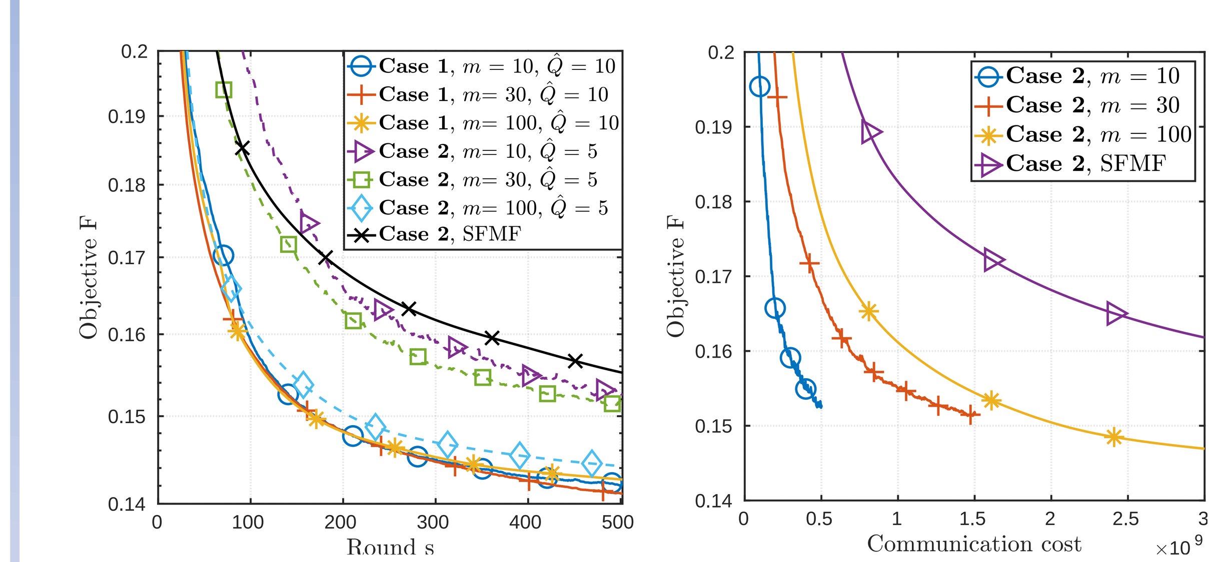


(a) syn, **Case 2**, $m = 100$ (b) syn, **Case 2**, PCC with $m = 10$



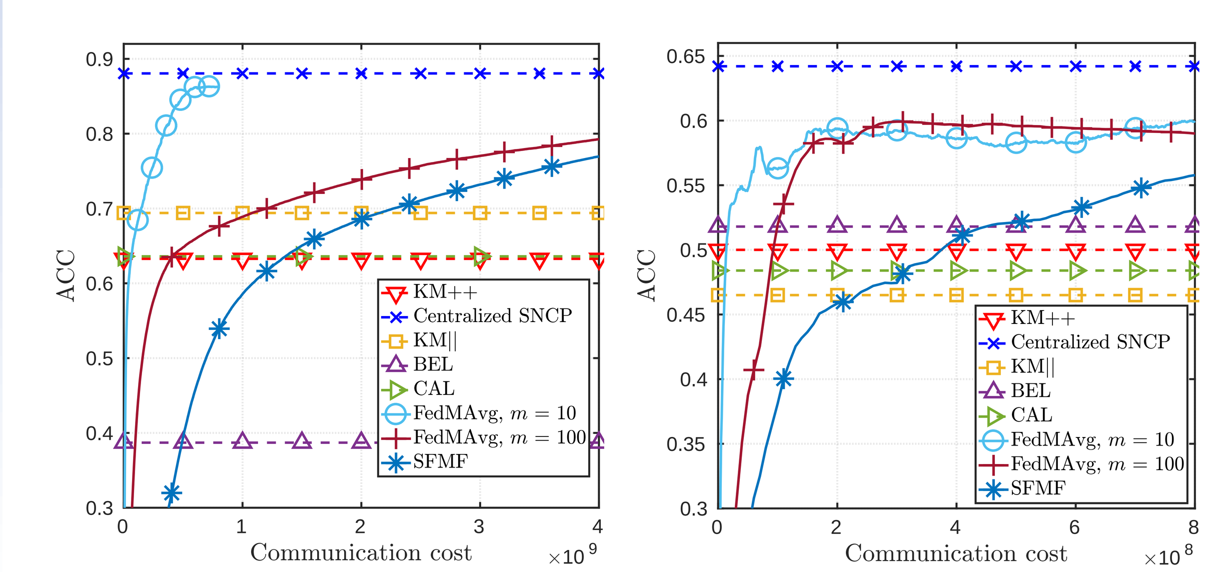
(c) syn, **Case 2** (d) syn, **Case 2**, $m = 10$

- Effect of non-i.i.d and PCC (Case 2: non-i.i.d):



(a) TCGA, F v.s. round s (b) TCGA, F v.s. com. cost

- Clustering performance versus distributed clustering (KM||[1], BEL[2], CAL[3]) and MF (SFMF[4]):



(a) syn, **Case 2** (b) TCGA, **Case 2**

References

[1] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," in Proc. VLDB 2012, Istanbul, Turkey, Aug. 27-31 2012, pp. 622–633.

[2] M.-F. Balcan, S. Ehrlich, and Y. Liang, "Distributed k-means and kmedian clustering on general topologies," in Proc. NeuIPS 2013, Lake Tahoe, USA, Dec. 5-10 2013, pp. 1995–2003.

[3] J. Chen, E. S. Azer, and Q. Zhang, "A practical algorithm for distributed clustering and outlier detection," in Proc. NeuIPS 2018, Montreal, Quebec, Canada, Dec. 2-8 2018, pp. 2248–2256.

[4] D. Chai, L. Wang, K. Chen, and Q. Yang, "Secure federated matrix factorization," IEEE Intelligent Systems, vol. 1, No. 1, pp. 1–8, Aug. 2020.