LOW-RANK OPTIMIZATION FOR DATA SHUFFLING IN WIRELESS DISTRIBUTED COMPUTING

Kai Yang*, Yuanming Shi*, and Zhi Ding†

*School of Information Science and Technology, ShanghaiTech University, Shanghai, China
†Dept. of ECE, University of California, Davis, California 95616, USA

On-Device Intelligence

Deploy artificial intelligence on mobile devices: augmented reality, smart vehicles, drones etc.

Benifits: smart decision, low latency, privacy

Key challenges: performing intensive computation tasks on resource-constrained mobile devices

Solution: Wireless distributed computing – pooling computation & storage resources among the devices

Wireless Distributed Computing: Computation Framework

Target: K users with individual computation task ϕ_k, depending on a common dataset

Dataset: too large for storing in one mobile device (e.g., a feature library of objects)

Solution: stored across devices (separated to N files $\{f_1, \cdots, f_N\}$, but each can only store up to μ files), supported by popular distributed computing framework **MapReduce**

$$\phi_k(d_k; f_1, \cdots, f_N) = h_k(g_{k,1}(d_k; f_1), \cdots, g_{k,N}(d_k; f_N)). \quad (1)$$

Map function: $w_{k,t} = g_{k,t}(d_k; f_t)$; **Reduce** function: $h_k(w_{k,1}, \cdots, w_{k,N})$; **Input**: d_k

Overall procedure for distributed computing:

- **Dataset Placement** Phase: determine file placement strategy index set of files stored at each node $\mathcal{F}_k \subseteq [N]$) and delivery files in advance
- **Map** Phase: compute intermediate values $\{w_{s,t} : s \in [K], t \in \mathcal{F}_k\}$ locally for all $k \in [K]$
- **Shuffle** Phase: exchange intermediate values wirelessly among nodes.
- **Reduce** Phase: construct the output value using the reduce function $h_k(w_{k,1}, \cdots, w_{k,N})$.

Assumptions:
$\mu < N$; the sizes of inputs and intermediate values are small enough to be stored locally; overhead of collecting all inputs is negligible

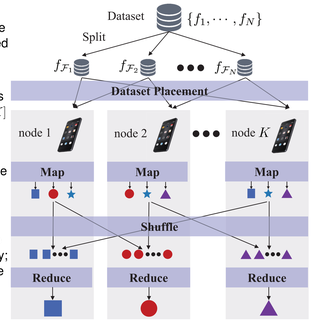

Figure: Distributed computing model.

Challenge and Prior Works

Technical challenge: To enable real-time and low-latency applications, inter-device communications for data shuffling and scheduling become the main bottleneck

Prior works: [Li, et al., TIT 18] [Li, et al., TON 17] proposed coding schemes to reduce communication load (e.g., information bits) for data shuffling in wireline and wireless distributed computing systems; **orthogonal uplink transmission** in [Li, et al., TON 17]

But the communication efficiency (i.e., **achieved data rates**) for data shuffling is also critical in **wireless** networks due to the **limited spectral resources** and **interference**!

Our Work: Low Rank Optimization for Data Shuffling

Communication Model

Network model: K users (each with L antennas) exchange intermediate values via a wireless access point (M antennas); **co-frequency transmission** in both uplink and downlink

Shuffle Phase reformulated as a message delivery problem:

- entire set of messages $\{W_1, \cdots, W_T\}$ (intermediate values$\{w_{k,n}\}$, $T = KN$)
- index set of messages available (can be computed locally) at user k: $\mathcal{T}_k \subseteq [T]$
- index set of messages required by mobile user k: $\mathcal{R}_k \subseteq [T]$

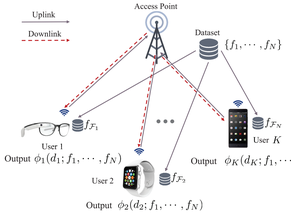

Figure: Communication model.

Uplink multiple access (MAC) stage:

$$\boldsymbol{y} = \sum_{k=1}^{K} (\boldsymbol{H}_k^u \otimes \boldsymbol{I}_r)\boldsymbol{x}_k + \boldsymbol{n}^u, \quad (2)$$

$\boldsymbol{y} \in \mathbb{C}^{Mr}$: received at the AP
$\boldsymbol{x}_k \in \mathbb{C}^{Lr}$: transmitted by mobile user k
r: channel uses (block fading channel)

Downlink broadcasting (BC) stage:

$$\boldsymbol{z}_k = (\boldsymbol{H}_k^d \otimes \boldsymbol{I}_r)\boldsymbol{y} + \boldsymbol{n}_k^d. \quad (3)$$

$\boldsymbol{z}_k \in \mathbb{C}^{Lr}$: received by mobile user k

Overall input-output relationship from mobile user to mobile user:

$$\boldsymbol{z}_k = \sum_{i=1}^{K} (\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{x}_i + \boldsymbol{n}_k, \quad (4)$$

$\boldsymbol{H}_{ki} = \boldsymbol{H}_k^d \boldsymbol{H}_i^u \in \mathbb{C}^{L \times L}$,
$\boldsymbol{n}_k = (\boldsymbol{H}_k^d \otimes \boldsymbol{I}_r)\boldsymbol{n}^u + \boldsymbol{n}_k^d$

Performance metric: For message l at user k, the degree-of-freedom (**DoF**) defined by

$$\mathrm{DoF}_{k,l} \triangleq \limsup_{\mathrm{SNR}_{k,l} \to \infty} \frac{R_{k,l}}{\log(\mathrm{SNR}_{k,l})}. \quad (5)$$

Linear Coding Scheme with Interference Alignment

Precoding matrix $\boldsymbol{V}_{kj} \in \mathbb{C}^{Lr \times d}$: $\boldsymbol{x}_k = \sum_{j \in \mathcal{T}_k} \boldsymbol{V}_{kj}\boldsymbol{s}_j$ **Decoding matrix** $\boldsymbol{U}_{kl} \in \mathbb{C}^{d \times Lr}$

$$\tilde{\boldsymbol{z}}_{kl} = \boldsymbol{U}_{kl}\boldsymbol{z}_k = \boldsymbol{U}_{kl}\sum_{i=1}^{K}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\sum_{j \in \mathcal{T}_i}\boldsymbol{V}_{ij}\boldsymbol{s}_j + \boldsymbol{n}_{kl} \quad (6)$$

$$= \mathcal{I}_1(\underbrace{\boldsymbol{s}_l}_{\text{desired message}}) + \mathcal{I}_2(\underbrace{\{\boldsymbol{s}_j : j \in \mathcal{T}_k\}}_{\text{locally available messages}}) + \mathcal{I}_3(\underbrace{\{\boldsymbol{s}_j : j \notin \mathcal{T}_k \cup \{l\}\}}_{\text{interferences}}) + \tilde{\boldsymbol{n}}_{kl} \quad (7)$$

$\mathcal{I}_1(\boldsymbol{s}_l) = \sum_{i:l \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{il}\boldsymbol{s}_l$ $\quad \mathcal{I}_2(\{\boldsymbol{s}_j : j \in \mathcal{T}_k\}) = \sum_{j \in \mathcal{T}_i}\sum_{i:j \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{ij}\boldsymbol{s}_j$
$\mathcal{I}_3(\{\boldsymbol{s}_j : j \notin \mathcal{T}_k \cup \{l\}\}) = \sum_{j \notin \mathcal{T}_k \cup \{l\}}\sum_{i:j \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{ij}\boldsymbol{s}_j$.

If interference alignment conditions:

$$\det\left(\sum_{i:l \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{il}\right) \neq 0,$$

$$\sum_{i:j \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{ij} = \boldsymbol{0},\ j \notin \mathcal{T}_k \cup \{l\}$$

are satisfied, **symmetric DoF** $\mathrm{DoF}_{\mathrm{sym}} = d/r$ is achievable
Assume $\sum_{i:l \in \mathcal{T}_i} \boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{il} = \boldsymbol{I}$ w.l.o.g.

Low rank formulation: $D = LdKT$

Low-Rank Matrix Representation:

$$\boldsymbol{U}_{kl}(\boldsymbol{H}_{ki} \otimes \boldsymbol{I}_r)\boldsymbol{V}_{ij} = \sum_{m=1}^{L}\sum_{n=1}^{L} H_{ki}[m,n]\boldsymbol{U}_{kl}[m]\boldsymbol{V}_{ij}[n]$$

$$\boldsymbol{X} = [\boldsymbol{U}_{kl}[m]\boldsymbol{V}_{ij}[n]] = \begin{bmatrix} \boldsymbol{U}_{11}[1] \\ \vdots \\ \boldsymbol{U}_{KT}[L] \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{11}[1] & \cdots & \boldsymbol{V}_{KT}[L] \end{bmatrix}$$

$\mathrm{rank}(\boldsymbol{X}) = r$, $\boldsymbol{X} \in \mathbb{C}^{LdKT \times LdKT}$

$$\mathscr{P}: \underset{\boldsymbol{X} \in \mathbb{C}^{D \times D}}{\text{minimize}}\ \mathrm{rank}(\boldsymbol{X}) \quad \text{subject to}\ \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$$

Nuclear norm relaxation yields poor performance due to the poor stucture of $\mathcal{A}$. E.g., in a 2-user case $K = N = 2$, $\mu = d = L = M = 1$, $\mathscr{P}$ becomes

$$\underset{\boldsymbol{X}}{\text{minimize}}\ \mathrm{rank}(\boldsymbol{X}) \quad \text{subject to}\ \boldsymbol{X} = \begin{bmatrix} \star & \star & 1/H_{12} & 0 \\ 0 & 1/H_{21} & \star & \star \end{bmatrix}, \quad (8)$$

DC Algorithmic Approach

Ky Fan k-norm: [Watson, 1993] The Ky Fan k-norm of $\boldsymbol{X}$ = sum of largest-k singular values,

$$\|\boldsymbol{X}\|_k = \sum_{i=1}^{k} \sigma_i(\boldsymbol{X}), \quad (9)$$

For any matrix $\boldsymbol{X} \in \mathbb{C}^{m \times n}$,

$$\mathrm{rank}(\boldsymbol{X}) = \min\{k : \|\boldsymbol{X}\|_* - \|\boldsymbol{X}\|_k = 0, k \leq \min\{m,n\}\}. \quad (10)$$

Then problem $\mathscr{P}$ can be solved by finding the minimum k such that the optimal objective value is zero for the difference-of-convex-functions (DC) problem

$$\underset{\boldsymbol{X} \in \mathbb{C}^{D \times D}}{\text{minimize}}\ \|\boldsymbol{X}\|_* - \|\boldsymbol{X}\|_k \quad \text{subject to}\ \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}. \quad (11)$$

Apply majorization-minimization (MM) algorithm to iteratively solve a convex approximation subproblem

$$\underset{\boldsymbol{X} \in \mathbb{C}^{D \times D}}{\text{minimize}}\ \|\boldsymbol{X}\|_* - \mathrm{Tr}(\partial\|\boldsymbol{X}_{t-1}\|_k, \boldsymbol{X}) \quad \text{subject to}\ \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b} \quad (12)$$

$\partial\|\boldsymbol{X}_t\|_k$ is the subdifferential of $\|\boldsymbol{X}\|_k$ at $\boldsymbol{X}_t$ and given by

$$\{\boldsymbol{U}\mathrm{diag}(\boldsymbol{q})\boldsymbol{V}^{\mathsf{H}}, \boldsymbol{q} = [\underbrace{1, \cdots, 1}_{k}, \underbrace{0, \cdots, 0}_{D-k}]^T\}. \quad (13)$$

Algorithm 1 DC Algorithm (DCA) for $\mathscr{P}$

1: **Inputs:**
   $\mathcal{A}, \boldsymbol{b}$,, accuracy ε
2: **Initialize:**
   $\boldsymbol{Y}_0 \in \mathbb{C}_*^{N \times r}$
3: **for** $k = 1, \cdots, D$ **do**
4: **Initialize:**
    $\boldsymbol{X}_0^{[k]} \in \mathbb{C}^{D \times D}, t = 1$
5: **while** not converge **do**
6: Compute $\partial\|\boldsymbol{X}_{t-1}^{[k]}\|_k$
7: Obtain the optimal solution $\boldsymbol{X}_t^{[k]}$ of 12
8: **end while**
9: **if** $\|\boldsymbol{X}^{[k]}\|_* - \|\boldsymbol{X}^{[k]}\|_k < \varepsilon$ **then return** $\boldsymbol{X}^{[k]}$
10: **end if**
11: **end for**
12: **Outputs:**
   $\boldsymbol{X}^{[k]}$ and rank k.

Simulation

Nuclear norm: nuclear norm relaxation approach. **IRLS-p**: iterative reweighted least square algorithm to further improve performance of nuclear norm relaxation. **DCA**: The present algorithm.

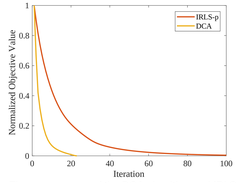

Figure: The convergence of the DCA algorithm and IRLS-p algorithm

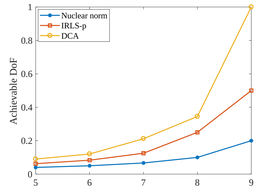

Figure: Maximum achievable symmetric DoF over μ

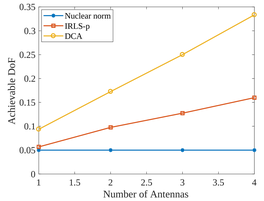

Figure: Maximum achievable symmetric DoF over the number of antennas

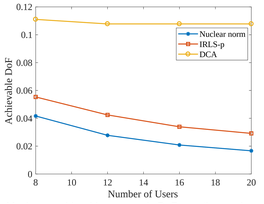

Figure: Maximum achievable symmetric DoF over the number of mobile users with uniform placement strategy

Conclusions

- Co-frequency transmission in both uplink and downlink
- Novel low-rank optimization approach for data shuffling in wireless distributed computing
- Algorithmic advantages of DC approach
- Scalability when the number of mobile users increases

References

[1] Song Han, Huizi Mao, and William J Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *Int. Conf. Learn. Representations (ICLR)*, 2016.

[2] S. Li, Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "A scalable framework for wireless distributed computing," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2643–2654, Oct. 2017.

[3] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A fundamental tradeoff between computation and communication in distributed computing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 109–128, Jan. 2018.

[4] GA Watson, "On matrix approximation problems with Ky Fan k norms," *Numerical Algorithms*, vol. 5, no. 5, pp. 263–272, 1993.

[5] Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Program.*, to appear, 2017.

Contact Information

Authors:

- Kai Yang: yangkai@shanghaitech.edu.cn
- Yuanming Shi: shiym@shanghaitech.edu.cn
- Zhi Ding: zding@ucdavis.edu

ShanghaiTech University UCDAVIS