Local Updates in Distributed Optimization

Xiang Li
Peking University, China

Federated Learning (FL)

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;**
  - **unable to control** over users' devices;
  - the training data are **non-iid.**

Communication efficiency.

Partial device participation.

Data Heterogeneity.

Problem Setup

- Consider the distributed optimization: $\min_{w} F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.
- The k-th device holds n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathscr{D}_k$.
- The local objective is defined by $F_k(w) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(w; x_{k,j})$ where $\ell(\cdot\,;\cdot)$ is a loss function.
- Note that (i) N could be very large; (ii) $\mathscr{D}_i \neq \mathscr{D}_j$ with $i \neq j$ due to heterogeneity; (iii) $p_k = \frac{n_k}{n}$.

FedAvg

- First, the central server **activates** a random small set (say $\mathcal{S}_t$) of devices and then **broadcasts** the latest model w_t to the **activated** devices;

- Second, every activated device (say the k-th and $k \in \mathcal{S}_t$) performs $E(\geq 1)$ **local updates**:$w_{t+i+1}^k \longleftarrow w_{t+i}^k - \eta_{t+i} \nabla F_k(w_{t+i}^k, \xi_{t+i}^k), i = 0,1,\cdots, E-1$ where η_{t+i} is the learning rate and ξ_{t+i}^k is a sample uniformly chosen from the k-th local dataset.

- Last, the server **aggregates** the local models, $\{w_{t+E}^k\}_{k \in \mathcal{S}_t}$ to produce the new global model, $w_{t+E} \longleftarrow$ Aggregate$(\{w_{t+E}^k\}_{k \in \mathcal{S}_t})$.

- Local Updates = multiple local training steps before synchronization

Theoretical Analysis For FedAvg

Under more realistic setting: namely partial device participation and non-iid data.

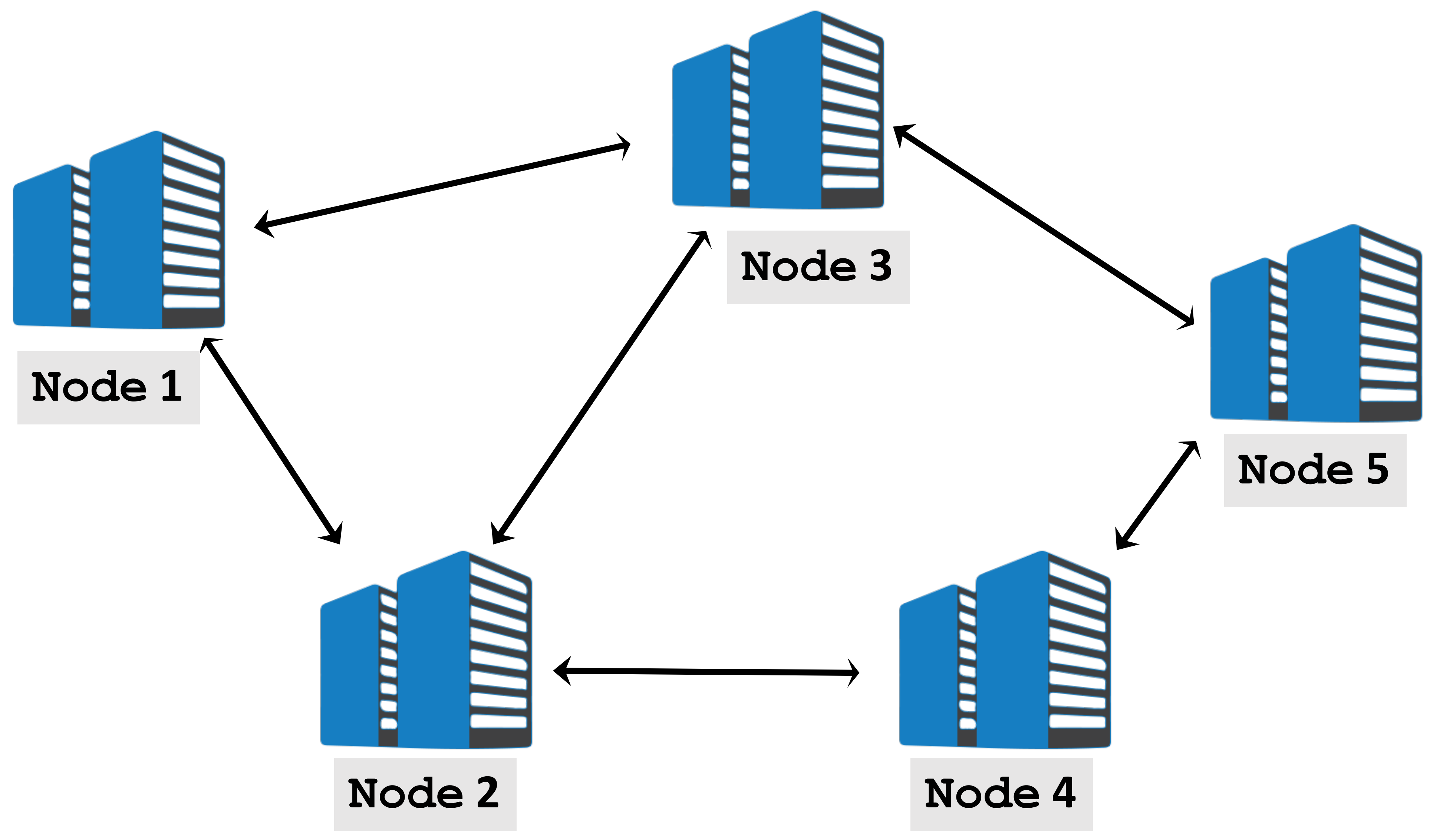
- Under some regularity conditions and decaying the learning rate, we have $\mathbb{E}\left[F(w_T) - F^*\right] \leq \mathcal{O}\left((\text{degree of non-iid} + (\text{local updates})^2 + \text{variance})/T\right)$.

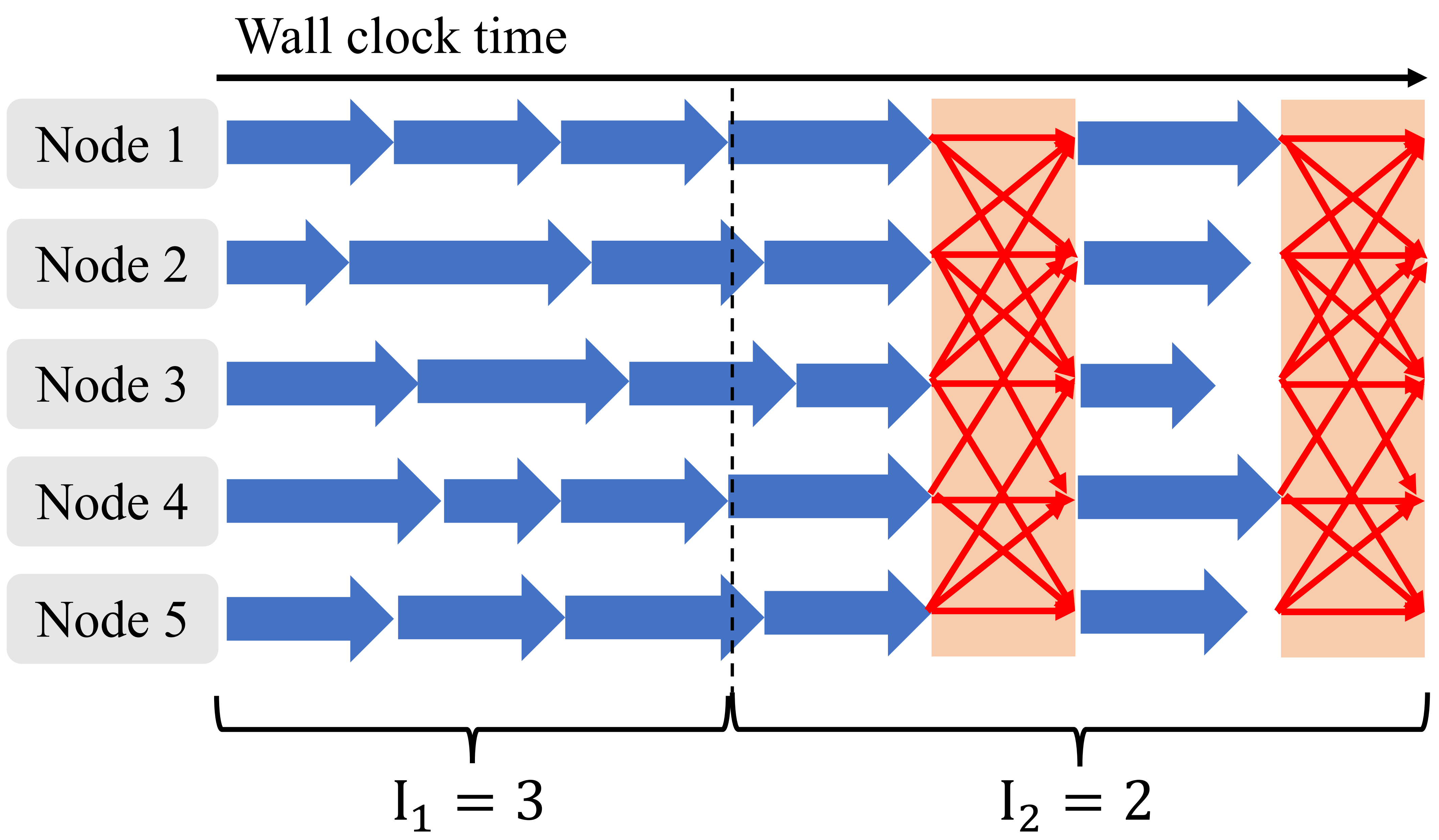
- If the learning rate doesn't decay, then $\tilde{w}^*$ (produced by FedAvg) is away from the optimal w^* (the optimal point): $\|\tilde{w}^* - w^*\|_2 = \Omega((E-1)\eta) \cdot \|w^*\|_2$.

- FedAvg converges when data are non-iid and devices participate in partially.

- The decay of learning rate is necessary.

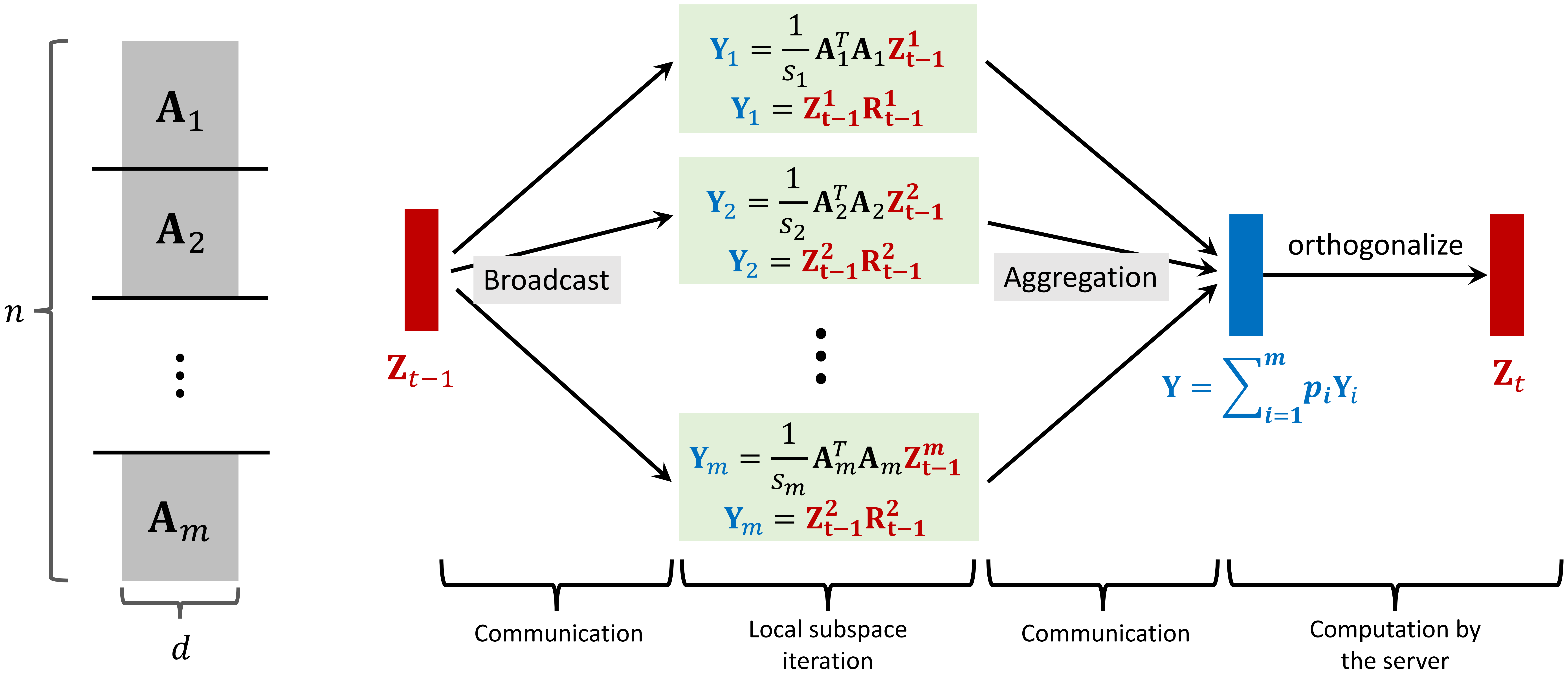
Local Updates for Decentralized Optimization





- For general smoothed non-convex decentralized optimization, local updates can be used to improve communication efficiency even the data is non-iid.

Local Updates for Distributed PCA



- For distributed top-k PCA, local updates can be combined with subspace iteration to improve communication efficiency.
- If p local updates are performed, communication complexity is reduced by a factor of p.

Thank You !