Distributed Optimization for Machine Learning

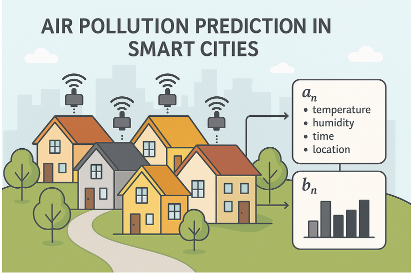
Lecture 14 - Communication-efficient Distributed Training

Tianyi Chen

School of Electrical and Computer Engineering
Cornell Tech, Cornell University

October 15, 2025

Example: air pollution prediction in smart cities



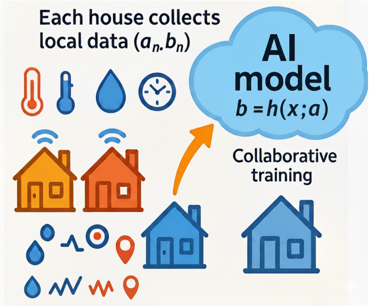
There is a community of multiple houses, where each house has a smart sensor that records environmental information.

Each house collects data pairs $\boldsymbol{\xi}_n = \{\boldsymbol{a}_n, \boldsymbol{b}_n\}$ over time, where:

$$\boldsymbol{a}_n = (\text{temperature, humidity, time, location, etc.})$$

$$\boldsymbol{b}_n = \text{concentration of a particular pollutant.}$$

Motivation of distributed training



Goal: All houses want to collaboratively train a machine learning model to predict future $\boldsymbol{b}$ given $\boldsymbol{a}$:

$$\boldsymbol{b} = h(\boldsymbol{x}; \boldsymbol{a})$$

where $\boldsymbol{x}$ denotes the model parameters to be learned.

Example: next-word prediction on smart keyboards

Each smartphone collects sequences of words typed by the user.

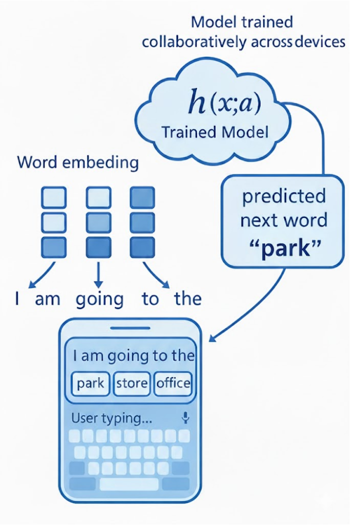
$$\boldsymbol{a}_n = \mathrm{Vec}\begin{pmatrix} \text{current word} \\ \vdots \\ \text{past word} \end{pmatrix},$$

$$\boldsymbol{b}_n = \mathrm{Vec}(\text{next word}).$$

Here, $\mathrm{Vec}(\cdot)$ denotes the Word-to-vector embedding operation.

Goal: Learn a model $h(\boldsymbol{x};\boldsymbol{a})$ to predict the next word embedding:

$$\boldsymbol{b} = h(\boldsymbol{x};\boldsymbol{a})$$



Why perform distributed training?

Key question: Why *distributed* data?

Main reason: Privacy!

- Each house may not want to share its raw sensor data with others or with a central server.
- Instead, they exchange only model updates or gradients to preserve local data confidentiality.

Secondary reason: Bandwidth and latency!

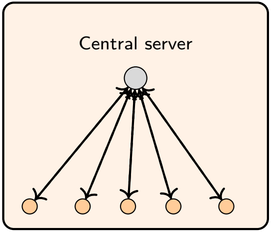
- Reduce communication overhead of transferring large datasets.
- Enable real-time, edge-level learning across smart devices.

Optimization formulation of data parallelism

A network of n nodes (such as mobile devices) collaborate to solve:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \quad \text{where } \boxed{f_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}_i\sim D_i}[F(\boldsymbol{x};\boldsymbol{\xi}_i)]}$$

- Each component $f_i : \mathbb{R}^d \to \mathbb{R}$ is local and private to node i.
- Random variable $\boldsymbol{\xi}_i$ denotes local data following distribution D_i.
- D_i may be different $\Rightarrow$ **data heterogeneity**.



Local data on nodes

Parallel SGD: compute locally, communicate globally

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \qquad \text{where } f_i(\mathbf{x}) = \mathbb{E}_{\xi_i\sim D_i}[F(\mathbf{x};\xi_i)].$$

PSGD

$$g_i^k = \nabla F\big(\mathbf{x}^k; \xi_i^k\big) \qquad \text{(Local compt.)}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\eta}{n}\sum_{i=1}^{n} g_i^k \qquad \text{(Global comm.)}$$

- Each node i samples mini-batch ξ_i^k and computes $\nabla F\big(\mathbf{x}^k; \xi_i^k\big)$.
- All nodes synchronize (i.e., *globally average*) to update $\mathbf{x}$.

Communication overhead of distributed training

Communication overhead:

- Each entry of a d-dimensional vector (model or gradient) requires 32 bits by default float32 (IEEE 754 single-precision floating-point).
- Each upload or download of the vector incurs:

$$\text{Communication cost} = 32 \times d \times n$$

where

- 32: bits per entry,
- d: number of dimensions ($10^6 \sim 10^{11}$),
- n: number of workers ($10^3 \sim 10^4$).

$\Rightarrow$ Total communication per round $= \mathcal{O}(10^{10}$ to $10^{16})$ bits.

Solutions to overcome communication overhead

Goal: Reduce the total communication cost per iteration:

$$32 \times d \times n$$

Possible solutions:

S1: **Reduce the communication rounds:**
e.g., Local SGD: perform τ local updates before synchronization to reduce communication frequency while maintaining accuracy.

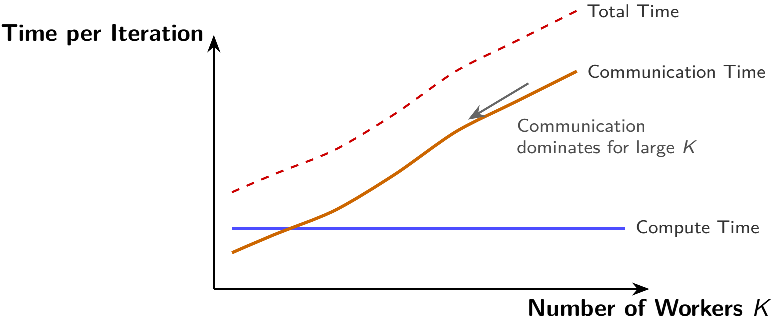
S2: **Reduce the number of bits via quantization/sparsification:**
e.g., Stochastic or deterministic quantization, threshold-based or Top-k sparsification.

S3: **Reduce the number of workers:**
e.g., Randomized / cyclic and adaptive worker selection (LAG).

Table of Contents

Communication bottleneck in Parallel SGD

- In Parallel SGD, workers **synchronize after every step**.
- Comm. dominates runtime when n is large or network is slow.



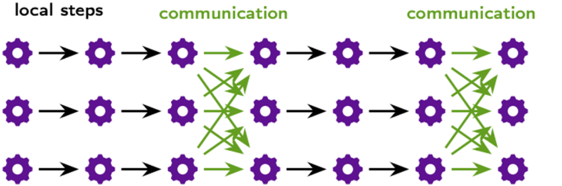
Idea: Local updates before synchronization

Key idea: Each node i performs several SGD steps before averaging.

$$\boldsymbol{x}_i^{(s+1)} = \boldsymbol{x}_i^{(s)} - \eta \nabla F(\boldsymbol{x}_i^{(s)}; \xi_i^{(s)}), \quad s = 0, \ldots, \tau - 1$$

where $\boldsymbol{x}_i^{(0)} = \boldsymbol{x}^k$. After every τ steps:

$$\boldsymbol{x}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^{(\tau)}$$



Benefit: Reduces communication by a factor of τ.

Mini-batch SGD vs. Local SGD

Mini-batch or Parallel SGD:

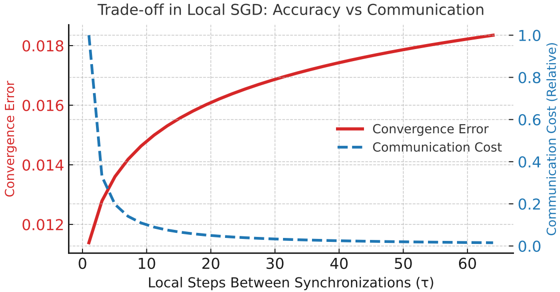
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta_t \frac{1}{n\tau} \sum_{i=1}^{n} \sum_{s=1}^{\tau} \nabla F(\boldsymbol{x}^{t}; \xi_i^{t,s})$$

Local SGD:

$$\boldsymbol{x}_i^{t+1} = \begin{cases} \boldsymbol{x}_i^{t} - \eta_t \nabla F(\boldsymbol{x}_i^{t}; \boldsymbol{\xi}_i^{t}), & t \bmod \tau \neq 0 \\ \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^{t} - \eta_t \nabla F(\boldsymbol{x}_i^{t}; \boldsymbol{\xi}_i^{t}) \right), & t \bmod \tau = 0 \end{cases}$$

Method	Mini-batch SGD	Local SGD
# Comm. rounds	K	K
Batch size	$n\tau$	n
# Model updates	K	τK
# Gradient calcs	$n\tau K$	$n\tau K$

Communication vs. computation trade-off



$$\text{Runtime per iteration} = \text{Compute} + \frac{1}{\tau}\text{Comm.}$$

Insight: Increasing τ improves efficiency but risks model drift.

Why Local SGD works under homogeneous data?

If $f(\boldsymbol{x})$ is convex and all workers start synchronized ($\boldsymbol{x}_i^t = \bar{\boldsymbol{x}}^t$):

$$f_i(\boldsymbol{x}_i^{t+\tau}) \le f_i(\bar{\boldsymbol{x}}^t) - (\text{descent term for worker } i).$$

Thus, synchronization preserves global descent.

$$f(\bar{\boldsymbol{x}}^{t+\tau}) \le \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}_i^{t+\tau}) \le f(\bar{\boldsymbol{x}}^t) - (\text{averaged progress}).$$

Not true in general if f_i differ across workers (non-i.i.d.).

Quadratic objectives: analytical insight

For local quadratic objectives $f_i(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x} - \boldsymbol{b}_i^\top \boldsymbol{x}$:

$$\boldsymbol{x}_i^{k+1} = \boldsymbol{x}_i^k - \eta_k \nabla f_i(\boldsymbol{x}_i^k).$$

Averaging yields:

$$\mathbb{E}[\bar{\boldsymbol{x}}^{k+1} | \mathcal{F}^k] = \bar{\boldsymbol{x}}^k - \eta_k \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_i^k) \approx \bar{\boldsymbol{x}}^k - \eta_k \nabla f(\bar{\boldsymbol{x}}^k).$$

Hence, Local SGD mimics global descent dynamics.

Local SGD improves efficiency in quadratic setting

Theorem 1 (Local SGD under smooth and convex loss)

The error bound for Local SGD with τ local updates equals the bound for Mini-batch SGD with batch size n and $K\tau$ rounds:

$$\epsilon_{\text{L-SGD}} := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[||\nabla f(\boldsymbol{x}^k)||_2^2] = \boldsymbol{\Theta}\left(\frac{1}{K\tau} + \frac{\sigma}{\sqrt{nK\tau}}\right)$$

- More local updates τ always help convergence.
- Mini-batch SGD: $\epsilon_{\text{MB-SGD}} = \boldsymbol{\Theta}\left(\frac{1}{K} + \frac{\sigma}{\sqrt{nK\tau}}\right)$
- Local SGD *can be better* given the same computation budget.

Performance for general convex objectives*

Upper and lower bounds for Local SGD:

Upper: $$\epsilon_{L\text{-}SGD} = \mathcal{O}\left(\frac{\sigma^{2/3}}{K^{2/3}\tau^{1/3}} + \frac{\sigma}{\sqrt{nK\tau}}\right)$$

Lower: $$\Omega\left(\frac{\sigma^{2/3}}{K^{2/3}\tau^{2/3}} + \frac{\sigma}{\sqrt{nK\tau}}\right)$$

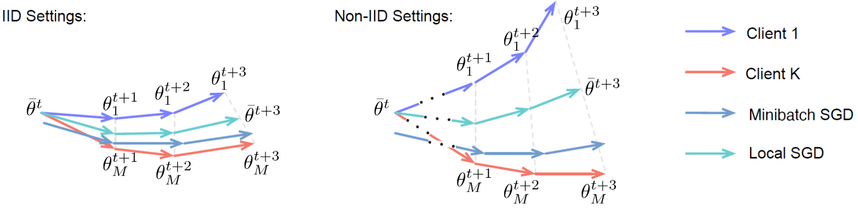
Mini-batch SGD: $\boldsymbol{\Theta}\left(\frac{1}{K} + \frac{\sigma}{\sqrt{nK\tau}}\right)$

Local SGD better when $K \lesssim \tau$, worse when $K \gtrsim \tau$.

Woodworth et al. "Is Local SGD Better than Mini-batch SGD?", *ICML 2020*

Why local SGD may fail under heterogeneous data?

In heterogeneous (non-i.i.d.) data settings, local gradients are misaligned

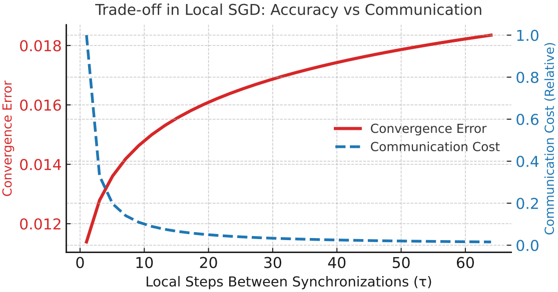


- Local updates diverge due to heterogeneous data (Γ^2).
- Need additional assumptions to control gradient dissimilarity.
- Larger $\tau \Rightarrow$ greater deviation from the global model.

Summary: Parallel SGD vs local SGD

Aspect	**Parallel SGD**	**Local SGD**
Communication	Every iteration	Every τ iterations
Local computation	1 gradient step	τ local steps
Speed	Communication-limited	Compute-efficient
Convergence rate	Stable	Slower ($(\eta^2\tau\Gamma^2)$ bias)
Best for	Data centers, i.i.d. data	Federated settings

Takeaway: Communication - accuracy trade-off



- $\tau = 1$: fully synchronized (Parallel SGD)
- $\tau > 1$: fewer syncs $\Rightarrow$ faster but drift grows
- Choose τ based on network bandwidth and data heterogeneity

Rule of thumb: $\tau^* \propto \sqrt{\frac{c_{\text{comm}}}{c_{\text{comp}}}}$

When to use local SGD?

Recommended if:

- Communication cost $c_{\mathrm{comm}} \gg c_{\mathrm{comp}}$
- Data across workers are relatively homogeneous
- Occasional synchronization suffices for convergence

Avoid if:

- Highly non-i.i.d. data (strong gradient heterogeneity)
- Models are unstable to small parameter changes

Table of Contents

Deterministic quantization

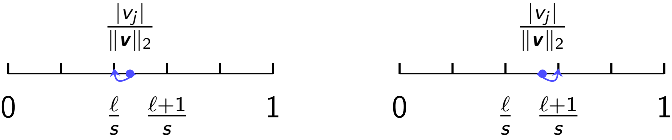
Goal: Compress model updates to fewer bits.

For any vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_d]^\top \in \mathbb{R}^d$, the j-th entry of the s-level quantized vector $Q_s(\boldsymbol{v})$ is defined as:

$$[Q_s(\boldsymbol{v})]_j := \|\boldsymbol{v}\|_2 \cdot \mathrm{sign}(v_j) \cdot \zeta_j(\boldsymbol{v}, s),$$

Let $0 \leq \ell < s$ be an integer such that $\frac{|v_j|}{\|\boldsymbol{v}\|_2} \in \left[\frac{\ell}{s}, \frac{\ell+1}{s}\right]$. Then:

$$\zeta_j(\boldsymbol{v}, s) = \begin{cases} \frac{\ell}{s}, & \text{if } \frac{|v_j|}{\|\boldsymbol{v}\|_2} - \frac{\ell}{s} \leq \frac{1}{2s}, \\ \frac{\ell+1}{s}, & \text{otherwise} \end{cases}$$

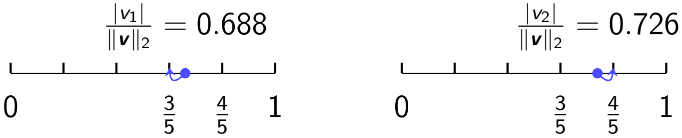


Example: deterministic quantization ($s = 5$)

Example: Consider a 2-D vector $\boldsymbol{v} = [0.36,\ 0.38]$. Its ℓ_2-norm is $\|\boldsymbol{v}\|_2 = \sqrt{0.36^2 + 0.38^2} \approx 0.523$. Thus,

$$\frac{|v_1|}{\|\boldsymbol{v}\|_2} = 0.688, \qquad \frac{|v_2|}{\|\boldsymbol{v}\|_2} = 0.726.$$

Both values fall into the same quantization interval $\left[\frac{3}{5}, \frac{4}{5}\right] = [0.6, 0.8]$.



According to the rule: $[Q_s(\boldsymbol{v})]_j = \|\boldsymbol{v}\|_2 \cdot \mathrm{sign}(v_j) \cdot \zeta_j(\boldsymbol{v}, s)$,
we obtain: $Q_5(\boldsymbol{v}) = 0.523\,[0.6,\ 0.8] = [0.314,\ 0.418]$.

Loss of deterministic quantization

Problem with this strategy: Higher quantization error for values that are further away from the center of the interval.

Lemma. For any vector $\boldsymbol{v} \in \mathbb{R}^d$, we have:

(i) $\|Q_s(\boldsymbol{v}) - \boldsymbol{v}\|_\infty \leq \frac{1}{s}\|\boldsymbol{v}\|_2$ (bias)

(ii) $\|Q_s(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \leq \frac{d^2}{s}\|\boldsymbol{v}\|_2^2$

Stochastic quantization

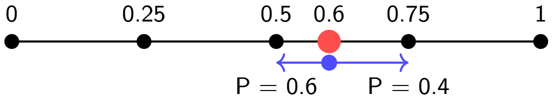
For any vector $\boldsymbol{v} = [v_1, \ldots, v_d]^\top \in \mathbb{R}^d$, the j-th entry of the s-level quantized vector $Q_s(\boldsymbol{v})$ is:

$$[Q_s(\boldsymbol{v})]_j := \|\boldsymbol{v}\|_2 \, \mathrm{sign}(v_j) \, \zeta_j(\boldsymbol{v}, s),$$

where the random variable $\zeta_j(\boldsymbol{v}, s)$ is:

$$\zeta_j(\boldsymbol{v}, s) = \begin{cases} \dfrac{\ell + 1}{s}, & \text{with probability } s\left(\dfrac{|v_j|}{\|\boldsymbol{v}\|_2} - \dfrac{\ell}{s}\right), \\ \dfrac{\ell}{s}, & \text{otherwise} \end{cases}$$

See example for $s = 4$ levels below:



Lemma: properties of stochastic quantization

Lemma: For any vector $\boldsymbol{v} \in \mathbb{R}^d$, if we apply stochastic quantization $Q_s(\boldsymbol{v})$, then we have:

(i) **Unbiasedness:**

$$\mathbb{E}[Q_s(\boldsymbol{v})] = \boldsymbol{v}$$

(ii) **Bounded variance:**

$$\mathbb{E}\left[\|Q_s(\boldsymbol{v}) - \boldsymbol{v}\|_2^2\right] \leq \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right) \|\boldsymbol{v}\|_2^2$$

The proof of the second property is given in Appendix 1 of the QSGD paper `https://arxiv.org/pdf/1610.02132`.

Convergence guarantees for QSGD: error bound

If $Q(\boldsymbol{v}_i^{(t)})$ is an unbiased stochastic estimator of $\nabla F_i(\boldsymbol{x}^t)$, then the quantized update is equivalent to a stochastic gradient update, and the standard SGD analysis can be applied.

Theorem 2 (Convergence of QSGD)

Let f be L-smooth and $\eta_k \equiv \eta = 1/\sqrt{K}$. Then the following holds:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}^k)\|_2^2\right] \leq \mathcal{O}\left(\frac{\sigma}{\sqrt{nK}}\sqrt{1 + \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)}\right).$$

- The error versus iterations convergence becomes worse if we use fewer quantization levels s.

Proof of QSGD error convergence bound

By combining the variance upper bound with the bounded estimation error property of the stochastic quantizer, we have:

$$\mathbb{E}_Q\left[\|Q(g(\boldsymbol{x};\boldsymbol{\xi})) - g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2\right] \leq \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2$$

$$\Rightarrow \quad \mathbb{E}_Q\left[\|Q(g(\boldsymbol{x};\boldsymbol{\xi}))\|_2^2\right] \leq \|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2 + \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2$$

$$\mathbb{E}_\xi\left[\mathbb{E}_Q\left[\|Q(g(\boldsymbol{x};\boldsymbol{\xi}))\|_2^2\right]\right] \leq \mathbb{E}_\xi\left[\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2\right] + \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)\mathbb{E}_\xi\left[\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2\right]$$

$$\mathbb{E}\left[\|Q(g(\boldsymbol{x};\boldsymbol{\xi}))\|_2^2\right] \leq \left(1 + \min\left(\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right)\right)\left(\|\nabla F(\boldsymbol{x})\|_2^2 + \sigma^2\right)$$

Implementation of stochastic quantization

After quantization, we transmit $Q_s(\boldsymbol{v})$ instead of the full vector $\boldsymbol{v}$.

The quantized vector $Q_s(\boldsymbol{v})$ can be represented by the tuple:

$$Q_s(\boldsymbol{v}) = \Big(\underbrace{\|\boldsymbol{v}\|_2}_{32 \text{ bits}}, \underbrace{\text{sign}(v_j)_{j=1}^d}_{d \text{ bits}}, \underbrace{\zeta_j(\boldsymbol{v}, s)_{j=1}^d}_{d \log_2 s \text{ bits}} \Big)$$

Total:

$$32 + d(1 + \log_2 s) \quad \text{vs.} \quad 32d \text{ (full precision)}$$

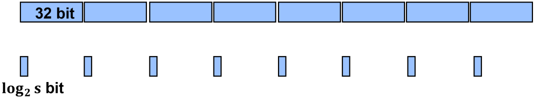
Conclusion: stochastic quantization effectively reduces communication cost while introducing a moderate increase in the error versus iterations convergence.

Sparsification

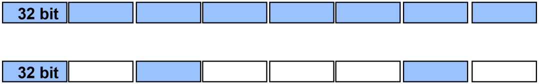
Goal: Reduce num of communicated entries by making vectors *sparse*.

Q: What is sparse?

- **Quantization:**



- **Sparsification:**



Idea: Communicate only a few coordinates and set the rest to zero.

Stochastic sparsification*

For any $\boldsymbol{v} \in \mathbb{R}^d$, define a sparsified vector $Q(\boldsymbol{v})$ coordinate-wise by:

$$[Q(\boldsymbol{v})]_j = \begin{cases} \dfrac{v_j}{p_j}, & \text{with probability } p_j, \\ 0, & \text{with probability } 1 - p_j, \end{cases} \qquad j = 1, \ldots, d.$$

Let $\boldsymbol{p} = (p_1, \ldots, p_d)$ be a predetermined probability vector belonging to a simplex ($p_j \in (0, 1], \sum_{j=1}^{d} p_j = 1$).

Lemma.

(i) **Unbiasedness:** $\mathbb{E}[Q(\boldsymbol{v})] = \boldsymbol{v}$ since $\mathbb{E}\big[[Q(\boldsymbol{v})]_j\big] = \frac{v_j}{p_j} p_j = v_j$

(ii) **Variance bound:** $\mathbb{E}\big[\|Q(\boldsymbol{v}) - \boldsymbol{v}\|_2^2\big] \leq \max_j \frac{1-p_j}{p_j} \|\boldsymbol{v}\|_2^2$

Wang, H., Sievert, S., Liu, S., Charles, Z., Papailiopoulos, D., and Wright, S. Atomo: Communication-efficient Learning via Atomic Sparsification, *NeurIPS 2018*

Deterministic sparsification

D1) Threshold-based rule

For any $\boldsymbol{v} \in \mathbb{R}^d$, denote the sparsified vector $Q(\boldsymbol{v})$.

$$[Q(\boldsymbol{v})]_j = \begin{cases} v_j, & \text{if } |v_j| \geq \gamma, \\ 0, & \text{otherwise,} \end{cases} \qquad \gamma : \text{predefined threshold.}$$

Idea: Only transmit coordinates whose magnitudes exceed τ.

Deterministic sparsification

D2) Memory-based threshold rule

If the algorithm transmits:

$$\begin{aligned} \text{Original:} \quad & \boldsymbol{v}^{(0)},\ \boldsymbol{v}^{(1)},\ \ldots,\ \boldsymbol{v}^{(K)} \\ \text{Sparsified:} \quad & Q(\boldsymbol{v}^{(0)}),\ Q(\boldsymbol{v}^{(1)}),\ \ldots,\ Q(\boldsymbol{v}^{(K)}) \end{aligned}$$

Initialize:

$$\tilde{\boldsymbol{v}}^{(0)} = \boldsymbol{v}^{(0)}.$$

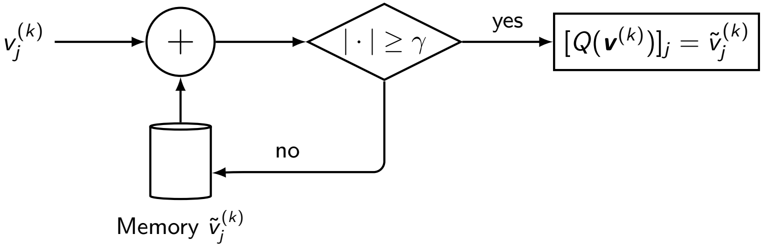
For $k = 0, 1, \ldots, K-1$:

$$[Q(\boldsymbol{v}^{(k)})]_j = \begin{cases} \tilde{v}_j^{(k)}, & \text{if } \left|\tilde{v}_j^{(k)}\right| \geq \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

$$\tilde{\boldsymbol{v}}^{(k+1)} = \boldsymbol{v}^{(k+1)} + \left(\tilde{\boldsymbol{v}}^{(k)} - Q(\boldsymbol{v}^{(k)})\right).$$

EndFor

Deterministic sparsification



For $k = 0, 1, \ldots, K-1$:

$$[Q(\boldsymbol{v}^{(k)})]_j = \begin{cases} \tilde{v}_j^{(k)}, & \text{if } \left|\tilde{v}_j^{(k)}\right| \geq \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

$$\tilde{\boldsymbol{v}}^{(k+1)} = \boldsymbol{v}^{(k+1)} + \left(\tilde{\boldsymbol{v}}^{(k)} - Q(\boldsymbol{v}^{(k)})\right).$$

EndFor

Deterministic sparsification

D3) Top-k sparsification rule*

Consider $\boldsymbol{\pi} \in \mathbb{R}^d$ as a permutation of $\{1, 2, \ldots, d\}$ such that for $\boldsymbol{v} \in \mathbb{R}^d$,

$$|v_{\pi(1)}| \geq |v_{\pi(2)}| \geq \cdots \geq |v_{\pi(d)}|.$$

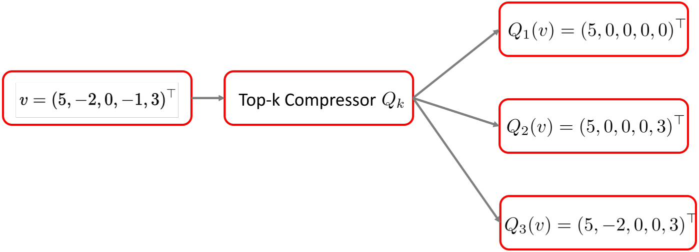
Then the j-th entry of the sparsified vector is:

$$[Q_k(\boldsymbol{v})]_j = \begin{cases} v_j, & \text{if } j = \pi(j) \text{ and } j \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

Stich, S.U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with Memory, *NeurIPS 2018*

Deterministic sparsification

D3) Top-k sparsification rule*



The error due to sparsification:

$$\|Q_k(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \leq \left(1 - \frac{k}{d}\right)\|\boldsymbol{v}\|_2^2.$$

Stich, S.U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with Memory, *NeurIPS 2018*

Implementation of quantized / sparsified gradient descent

For iteration $k = 1, 2, \ldots, K$:

1. **Server broadcasts** the current model parameter $\boldsymbol{x}^k$ to all workers.
2. **For each worker** $i = 1, 2, \ldots, n$ (in parallel):
   - Worker i calculates $\boldsymbol{v}_i^{(k)} = \nabla F_i(\boldsymbol{x}^k)$.
   - Worker i computes sparsified/quantized gradient $Q(\boldsymbol{v}_i^{(k)})$.
   - Worker i uploads $Q(\boldsymbol{v}_i^{(k)})$ to the server.
3. **Server updates** the global model:

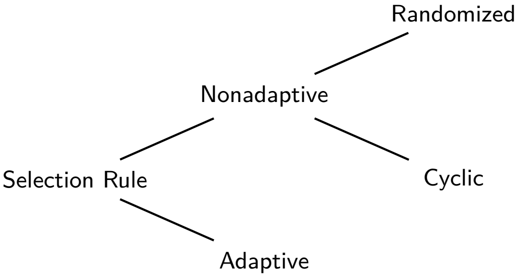
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{\alpha}{n} \sum_{i=1}^{n} Q\left(\boldsymbol{v}_i^{(k)}\right).$$

Remark: Quantization / sparsification can be performed either at the server side or at the worker side or both.

Table of Contents

Reduce the number of workers

Goal: Reduce the number of workers participating in communication.



Idea: Only a subset of workers upload/download gradients at each round, based on either fixed (nonadaptive) or dynamic (adaptive) rules.

Non-adaptive randomized rule

For iteration $k = 1, 2, \ldots, K$:

1. Server randomly selects a worker $i_k \in \{1, \ldots, n\}$ (or a set $\mathcal{I}_k \subseteq \{1, \ldots, n\}$).
2. Server sends $\boldsymbol{x}^k$ to worker i_k (or all $i \in \mathcal{I}_k$).
3. Worker i_k computes and uploads $\nabla F_{i_k}(\boldsymbol{x}^k)$.
4. **Server updates $\boldsymbol{x}^k$ via:**

   Option I (SGD):

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha \nabla F_{i_k}(\boldsymbol{x}^k).$$

   Option II (Randomized Incremental Aggregated Gradient (RIAG)):

$$\boldsymbol{x}_i^{k+1} = \begin{cases} \boldsymbol{x}^k, & i = i_k, \\ \boldsymbol{x}_i^k, & i \neq i_k, \end{cases} \qquad \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha \nabla F_{i_k}(\boldsymbol{x}^k) - \alpha \sum_{i \neq i_k} \nabla F_i(\boldsymbol{x}_i^k).$$

Memory for RIAG

If the server pursues Option II, it stores a table $\in \mathbb{R}^{d\times n}$.

∇F_1	∇F_2	∇F_3	$\cdots$	∇F_n

Overcome Memory Overhead:

Store the summation $\nabla_k = \sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^k)$.
Worker uploads only the change of gradients:

$$\nabla_k^i = \nabla F_i(\boldsymbol{x}^k) - \nabla F_i(\boldsymbol{x}_i^k).$$

Server updates the summation via:

$$\nabla_{k+1} = \nabla_k + \nabla_k^i.$$

Non-adaptive cyclic rule

For $k = 1, 2, \ldots, K$:

1. Server selects worker $i_k = k \bmod n$.
2. Server sends $\boldsymbol{x}^k$ to worker i_k.
3. Worker i_k computes and uploads $\nabla F_{i_k}(\boldsymbol{x}^k)$.
4. Server updates $\boldsymbol{x}^k$ via Option I or II.

CIAG: Cyclic Incremental Aggregated Gradient

Theoretical guarantees of CIAG

Theorem 3 (Convergence of CIAG)

Under the L-smooth and μ-strongly convex assumption, if the stepsize α in CIAG satisfies:

$$0 < \alpha \leq \frac{1}{n(\mu + L)},$$

then CIAG achieves an **R-linear convergence rate:**

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 \leq \rho^k \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2, \quad \text{for some } 0 < \rho < 1.$$

Plan: adaptive worker selection

Compare: Gradient Descent vs. RIAG/CIAG

Tradeoff Factors:

(c1) Amount of communication per iteration

(c2) Number of iterations required for convergence

Observation:

$$\text{RIAG/CIAG} \approx \frac{1}{n} \text{ communications as GD (fewer uploads per iteration)},$$
$$\text{GD} \approx \frac{1}{n} \text{ iterations as RIAG/CIAG (faster convergence per round)}.$$

Total Communication Cost:

$$\text{Total communication rounds} = (c1) \times (c2).$$

Adaptive worker selection best tradeoff

A slight generalization of Incremental Aggregated Gradient (IAG):

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha \sum_{i \in \mathcal{I}^k} \nabla F_i(\boldsymbol{x}^k) - \alpha \sum_{i \notin \mathcal{I}^k} \nabla F_i(\boldsymbol{x}_i^k),$$

$$\text{where } \mathcal{I}^k \subset \{1, 2, \ldots, n\}.$$

Special cases:

- **RIAG (Randomized IAG):** $\mathcal{I}^k = \{i_k\}$, with i_k randomly generated.
- **CIAG (Cyclic IAG):** $\mathcal{I}^k = \{k \text{ mod } n\}$.
- **GD (Full Gradient Descent):** $\mathcal{I}^k = \{1, 2, \ldots, n\}$.

Incremental aggregated gradient

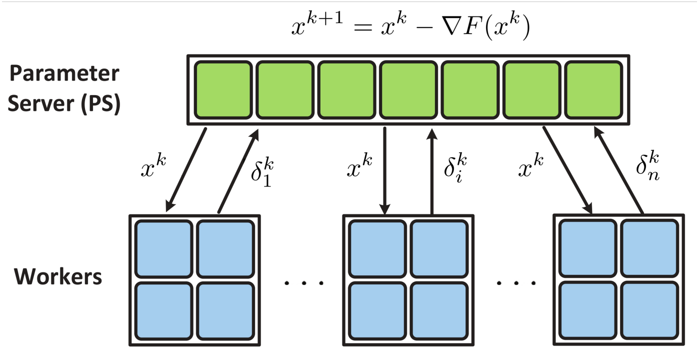
$$\begin{aligned}\boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - \alpha \sum_{i\in\mathcal{I}^k} \nabla F_i(\boldsymbol{x}^k) - \alpha \sum_{i\notin\mathcal{I}^k} \nabla F_i(\boldsymbol{x}_i^k) \\ &= \underbrace{\boldsymbol{x}^k - \alpha \sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}^k)}_{\text{GD update}} + \underbrace{\alpha \sum_{i\notin\mathcal{I}^k} \Big(\nabla F_i(\boldsymbol{x}^k) - \nabla F_i(\boldsymbol{x}_i^k)\Big)}_{\boldsymbol{\delta}_i^k}\end{aligned}$$

Error of using old gradients: $\boldsymbol{\delta}_i^k$

Intuition: If $\|\boldsymbol{\delta}_i^k\|$ are small relative to $\sum_{i=1}^n \|\nabla F_i(\boldsymbol{x}^k)\|$, then the price paid for saving uploads/downloads is small.

Incremental aggregated gradient

Question: The intuition is good but *how to quantify small?*



Toward adaptive worker selection

Design an *adaptive selection rule* by analyzing the IAG iteration.

Lemma (IAG)

Under the L-smooth assumption of $F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{x})$, $\boldsymbol{x}^{k+1}$ *is generated by performing one-step generic IAG update given* $\boldsymbol{x}^k$ *and* $\{\boldsymbol{x}_i^k\}_{i=1}^n$, *then:*

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{\alpha}{2}\|\nabla F(\boldsymbol{x}^k)\|_2^2 + \frac{\alpha}{2}\Big\|\sum_{i\notin\mathcal{I}^k}\boldsymbol{\delta}_i^k\Big\|_2^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2$$

$$\overset{\alpha=\frac{1}{L}}{\Longrightarrow} \leq -\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|_2^2 + \frac{1}{2L}\Big\|\sum_{i\notin\mathcal{I}^k}\boldsymbol{\delta}_i^k\Big\|_2^2 \triangleq \Delta_{CIAG}^k.$$

Communication principle

Lemma (GD)

Under the same L-smooth assumption, the one-step GD update satisfies:

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|_2^2 \triangleq \Delta_{GD}^k.$$

Principle: Larger progress per communication:

$$\frac{\Delta_{\text{IAG}}^k}{|\mathcal{I}^k|} \leq \frac{\Delta_{\text{GD}}^k}{n}$$

Plugging Δ_{GD}^k and Δ^k_{IAG} leads to:

$$\frac{-\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{1}{2L}\sum_{i\notin\mathcal{I}^k}\|\boldsymbol{\delta}_i^k\|^2}{|\mathcal{I}^k|} \leq \frac{-\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|^2}{n}$$

Deriving the sufficient condition for the principle

$$\frac{-\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{1}{2L}\sum_{i\notin\mathcal{I}^k}\|\boldsymbol{\delta}_i^k\|^2}{|\mathcal{I}^k|} \leq \frac{-\frac{1}{2L}\|\nabla F(\boldsymbol{x}^k)\|^2}{n}$$

$$\iff \Big\|\sum_{i\notin\mathcal{I}^k}\boldsymbol{\delta}_i^k\Big\|^2 \leq \Big(1 - \frac{|\mathcal{I}^k|}{n}\Big)\|\nabla F(\boldsymbol{x}^k)\|^2.$$

By Cauchy–Schwarz inequality, $\|\boldsymbol{a}_1 + \boldsymbol{a}_2 + \cdots + \boldsymbol{a}_n\|^2 \leq n\sum_{i=1}^n \|\boldsymbol{a}_i\|^2$, it holds that:

$$\Big\|\sum_{i\notin\mathcal{I}^k}\boldsymbol{\delta}_i^k\Big\|^2 \leq \Big(n - |\mathcal{I}^k|\Big)\sum_{i\notin\mathcal{I}^k}\|\boldsymbol{\delta}_i^k\|^2 \leq \Big(n - |\mathcal{I}^k|\Big) n \max_{i\notin\mathcal{I}^k}\|\boldsymbol{\delta}_i^k\|^2.$$

Deriving the sufficient condition for progress principle

Sufficient Condition for the Principle:

$$\left(n - |\mathcal{I}^k|\right) n \max_{i \notin \mathcal{I}^k} \|\boldsymbol{\delta}_i^k\|^2 \leq \frac{n - |\mathcal{I}^k|}{n} \|\nabla F(\boldsymbol{x}^k)\|^2.$$

$$\iff \|\boldsymbol{\delta}_i^k\|^2 \leq \frac{1}{\alpha^2 n^2} \|\nabla F(\boldsymbol{x}^k)\|^2, \quad \text{for all } i \in \{1, \ldots, n\}.$$

Q: How can we check this condition either at the server or at worker?

$$\|\nabla F(\boldsymbol{x}^k)\|^2 = \left\| \sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}^k) \right\|^2$$

This cannot be computed locally.

Checking the sufficient condition

Approximation:

$$\|\nabla F(\boldsymbol{x}^k)\|^2 \approx \frac{1}{\alpha^2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2$$

so that each worker can check condition locally by:

$$\boxed{\|\boldsymbol{\delta}_i^k\|^2 \leq \frac{1}{\alpha^2 n^2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 \quad (\textit{Worker side})}$$

Q: What if we find an upper bound on the left-hand side?

$$\|\nabla F_i(\boldsymbol{x}^k) - \nabla F_i(\boldsymbol{x}_i^k)\| \leq L_i\|\boldsymbol{x}^k - \boldsymbol{x}_i^k\|$$

A sufficient condition rule is:

$$\boxed{L_i^2\|\boldsymbol{x}^k - \boldsymbol{x}_i^k\|^2 \leq \frac{1}{\alpha^2 n^2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 \quad (\textit{Server side})}$$

Implementation of adaptive selection rule (LAG)*

Worker side:
For iteration $k = 1, 2, \ldots, K$:

1. **Server broadcasts** the current model parameter $\boldsymbol{x}^k$ to all workers.
2. **For each worker** $i = 1, 2, \ldots, n$ **(in parallel):**
   - Worker i computes the local gradient $\nabla F_i(\boldsymbol{x}^k)$.
   - Worker i checks the upload condition:

$$\|\boldsymbol{\delta}_i^k\|^2 \leq \frac{1}{\alpha^2 \boldsymbol{n}^2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2.$$

   - If the condition is satisfied $\Rightarrow$ Do not upload.
   - Otherwise $\Rightarrow$ Upload.
3. **Server updates** the global model via the generic IAG update rule.

Chen, T., Giannakis, G., Sun, T., and Yin, W. LAG: Lazily Aggregated Gradient for Communication-efficient Distributed Learning, *NeurIPS 2018*

Implementation of adaptive selection rule (LAG)*

Server side:
For iteration $k = 1, 2, \ldots, K$:

1. **Server checks** the condition for each worker $i = 1, 2, \ldots, n$:

$$L_i^2 \|\boldsymbol{x}^k - \boldsymbol{x}_i^k\|^2 \leq \frac{1}{\alpha^2 n^2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2.$$

2. Collect all violating workers into the set $\mathcal{I}^k$.
3. **Server sends** the current model $\boldsymbol{x}^k$ to all $i \in \mathcal{I}^k$.
4. **For each worker** $i \in \mathcal{I}^k$:
   - Worker i computes and uploads $\nabla F_i(\boldsymbol{x}^k)$ to the server.
5. **Server updates** the global parameter $\boldsymbol{x}^{k+1}$ via the generic IAG update rule.

Chen, T., Giannakis, G., Sun, T., and Yin, W. LAG: Lazily Aggregated Gradient for Communication-efficient Distributed Learning, *NeurIPS 2018*

Theoretical guarantee of LAG

Theorem 4 (Convergence of LAG)

1. Under the L-smooth assumption of $F_i(\boldsymbol{x})$, we have:

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\boldsymbol{x}^k)\|^2 = \mathcal{O}\left(\frac{1}{K}\right) \qquad \text{(Same as GD)}$$

2. Under the additional convex assumption, we have:

$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) = \mathcal{O}\left(\frac{1}{K}\right) \qquad \text{(Same as GD)}$$

3. Under the additional μ-strong convexity assumption, we have:

$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) = \mathcal{O}\big((1-\tfrac{\mu}{L})^k\big) \qquad \text{(Same as GD)}$$

Empirical performance of LAG

- **Faster convergence per iteration:** LAG achieves similar or faster convergence compared with IAG and GD in terms of iteration complexity.
- **Significantly reduced communication cost:** LAG requires fewer communication rounds while maintaining accuracy.

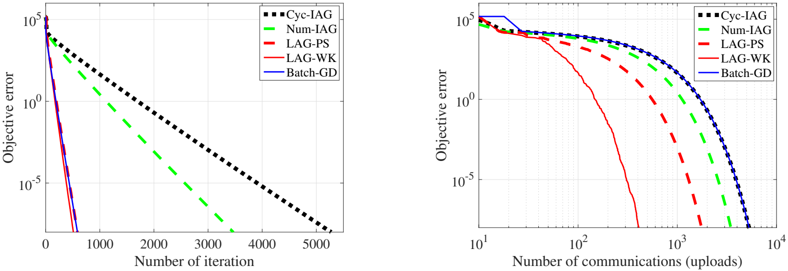


Figure: Iteration and communication complexity for linear regression.

Proof sketch of worker-side condition of LAG

$$\|\boldsymbol{\delta}_i^k\|^2 \ \leq \ \frac{\zeta}{\alpha^2 n^2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2$$

ζ is a hyperparameter controlling the magnitude of the condition.
Larger $\zeta \Rightarrow$ condition becomes easier to satisfy.

Recall: To prove GD under smooth and nonconvex settings, we first establish the *one-step progress* (descent lemma).

Next: Derive the one-step progress of LAGWorker.

Under the L-smoothness assumption, we have:

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq \underbrace{\left\langle \nabla F(\boldsymbol{x}^k), \boldsymbol{x}^{k+1} - \boldsymbol{x}^k \right\rangle}_{\text{(I)}} + \underbrace{\frac{L}{2} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2}_{\text{(II)}}.$$

Bounding the inner-product term (I)

$$
\begin{aligned}
&\left\langle \nabla F(\boldsymbol{x}^k),\, \boldsymbol{x}^{k+1} - \boldsymbol{x}^k \right\rangle \\
&= \left\langle \nabla F(\boldsymbol{x}^k),\, -\alpha \sum_{i \in n} \nabla F_i(\boldsymbol{x}^k) - \alpha \sum_{i \in n \backslash \mathcal{I}^k} \nabla F_i(\boldsymbol{x}_i^k) \right\rangle \\
&= \left\langle \nabla F(\boldsymbol{x}^k),\, -\alpha \nabla F(\boldsymbol{x}^k) - \alpha \sum_{i \in n \backslash \mathcal{I}^k} \boldsymbol{\delta}_i^k \right\rangle \\
&= -\alpha \|\nabla F(\boldsymbol{x}^k)\|^2 + \alpha \left\langle -\nabla F(\boldsymbol{x}^k),\, \sum_{i \in n \backslash \mathcal{I}^k} \left(\nabla F_i(\boldsymbol{x}_i^k) - \nabla F_i(\boldsymbol{x}^k) \right) \right\rangle \\
&\qquad (\text{using } 2\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2) \\
&= -\alpha \|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{\alpha}{2} \|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{\alpha}{2} \left\| \sum_{i \in n \backslash \mathcal{I}^k} \left(\nabla F_i(\boldsymbol{x}_i^k) - \nabla F_i(\boldsymbol{x}^k) \right) \right\|^2 \\
&\quad - \frac{\alpha}{2} \left\| \sum_{i \in n} \nabla F_i(\boldsymbol{x}^k) + \sum_{i \in n \backslash \mathcal{I}^k} \left(\nabla F_i(\boldsymbol{x}_i^k) - \nabla F_i(\boldsymbol{x}^k) \right) \right\|^2
\end{aligned}
$$

Bounding the inner-product term (I)

$$= -\frac{\alpha}{2}\,\|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{\alpha}{2}\,\Big\|\sum_{i\in n\setminus\mathcal{I}^k} \boldsymbol{\delta}_i^k\Big\|^2$$

$$- \frac{\alpha}{2}\,\underbrace{\Big\|\sum_{i\in n}\nabla F_i(\boldsymbol{x}^k) + \sum_{i\notin\mathcal{I}^k}\nabla F_i(\boldsymbol{x}_i^k)\Big\|^2}_{\frac{1}{\alpha}(\boldsymbol{x}^k-\boldsymbol{x}^{k+1})}.$$

Plugging into the L-smoothness inequality, we have:

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \le -\frac{\alpha}{2}\|\nabla F(\boldsymbol{x}^k)\|^2 + \frac{\alpha}{2}\sum_{i\notin\mathcal{I}^k}\|\boldsymbol{\delta}_i^k\|^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2$$

Bounding the inner-product term (I)

(Use worker condition)

$$\leq -\frac{\alpha}{2}\|\nabla F(\boldsymbol{x}^k)\|^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 + \frac{\alpha n}{2}\sum_{i \notin \mathcal{I}^k} \frac{\zeta}{\alpha^2 n^2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2$$

$$\leq -\frac{\alpha}{2}\|\nabla F(\boldsymbol{x}^k)\|^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 + \frac{\zeta}{2\alpha n}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2.$$

Rearranging terms, we have:

$$\frac{\alpha}{2}\|\nabla F(\boldsymbol{x}^k)\|^2 \leq F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) + \frac{\zeta}{2\alpha}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 - \left(\frac{1-\alpha L}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2$$

(Choose $\zeta = 1 - \alpha L$)

$$= F(\boldsymbol{x}^k) - F(\boldsymbol{x}^{k+1}) + \frac{1-\alpha L}{2\alpha}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 - \left(\frac{1-\alpha L}{2\alpha}\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2$$

Telescoping and final convergence rate

Telescoping $k = 1, 2, \ldots, K$, we have:

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\boldsymbol{x}^k)\|^2 \leq \frac{1}{\alpha K}\big(F(\boldsymbol{x}^1) - F(\boldsymbol{x}^{K+1})\big) + \frac{1-\alpha L}{2\alpha^2 K}\|\boldsymbol{x}^1\|^2 - \frac{1-\alpha L}{2\alpha^2 K}\|\boldsymbol{x}^{K+1} - \boldsymbol{x}^K\|^2$$

Since $-F(\boldsymbol{x}^{K+1}) \leq -F(\boldsymbol{x}^*)$, we get:

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\boldsymbol{x}^k)\|^2 \leq \frac{1}{\alpha K}\big(F(\boldsymbol{x}^1) - F(\boldsymbol{x}^*)\big) + \frac{1-\alpha L}{2\alpha^2 K}\|\boldsymbol{x}^1\|^2 = \mathcal{O}\big(\tfrac{1}{K}\big).$$

Recap and fine-tuning

- What we have talked about today?

  ⇒ **Local SGD** reduces communication rounds by allowing each worker to perform multiple local updates before synchronization.

  ⇒ **Quantization / sparsification** reduces communication cost by transmitting gradients with fewer bits or fewer entries.

  ⇒ **Worker selection** reduces communication cost by letting only a subset of workers upload gradients adaptively.

Welcome anonymous survey!