A Unified Analysis of Stochastic Gradient Descent with Arbitrary Data Permutations and Beyond

Yipeng Li[1], Xinchen Lyu[2] and Zhenyu Liu[1]

[1]Shenzhen International Graduate School, Tsinghua University
[2]National Engineering Research Center for Mobile Network Technologies, Beijing University of Posts and Telecommunications

liyp25@mails.tsinghua.edu.cn, lvxinchen@bupt.edu.cn, zhenyuliu@sz.tsinghua.edu.cn

The Problem Formulation

(1) The finite-sum minimization **problem**

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{N}\sum_{n=0}^{N-1} f_n(\mathbf{x})\right]$$

Here, N denotes the number of local objective functions, f_n denotes the local objective.

(2) **Permutation-based SGD** vs. classic SGD

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \gamma\nabla f_{\pi(n)}(\mathbf{x}^n)$$

Here, γ denotes the step size, $\pi(n)$ is the index of the local objective at iteration n.

- Classic SGD: $\pi(n)$ is chosen uniformly with replacement from $\{0,1,\ldots,N\text{-}1\}$.
- Permutation-based SGD: $\pi(n)$ is the $(n+1)$-th element of a permutation π of $\{0,1,\ldots,N\text{-}1\}$.

The Existing Permutation-based SGD algorithms

Based on the relations among permutations, we classify the existing permutation-based SGD algorithms into the three categories:

1. **Arbitrary Permutations (AP)**: Permutations are generated without any specific structure, allowing for completely arbitrary permutations for all epochs.
2. **Independent Permutations (IP)**: Permutations are independent across epochs.
   – Random Reshuffling (RR): The permutation in each epoch is generated randomly.
   – FlipFlop [Rajput et al., 2022].
   – Greedy Ordering [Lu et al., 2022b; Mohtashami et al., 2022]: The permutation in each epoch is generated by a greedy algorithm
3. **Dependent Permutations (DP)**: Permutations are dependent across epochs, with the permutation in one epoch affected by the permutations in previous epochs (explicitly).
   – One Permutation (OP): The initial (first-epoch) permutation is used repeatedly for all the subsequent epochs. When the initial permutation is arbitrary, it is called Incremental Gradient (IG); when the initial permutation is random, it is called Shuffle Once (SO).
   – GraBs: It includes GraB [Lu et al., 2022a] and PairGraB [Cooper et al., 2023].

The Order Error

(1) The derivation.

- For a small finite step size γ, the cumulative updates in any epoch q are

$$\mathbf{x}_{q+1} - \mathbf{x}_q \approx \underbrace{-\gamma N \nabla f(\mathbf{x}_q) + \gamma^2 \sum_{n=0}^{N-1} \sum_{i<n} \nabla^2 f_{\pi(n)}(\mathbf{x}_q) \nabla f(\mathbf{x}_q)}_{\text{optimization vector}} + \underbrace{\gamma^2 \sum_{n=0}^{N-1} \sum_{i<n} \nabla^2 f_{\pi(n)}(\mathbf{x}_q) \left(\nabla f_{\pi(i)}(\mathbf{x}_q) - \nabla f(\mathbf{x}_q)\right)}_{\text{error vector}},$$

- The goal is to suppress the error vector:

$$\|\text{Error vector}\| \leq \gamma^2 \sum_{n=0}^{N-1} \left\|\nabla^2 f_{\pi(n)}(\mathbf{x}_q)\right\| \left\| \sum_{i<n} \left(\nabla f_{\pi(i)}(\mathbf{x}_q) - \nabla f(\mathbf{x}_q)\right) \right\| \leq \gamma^2 L N \bar{\phi}_q,$$

(2) The definition.

Definition 1 (Order Error, Lu et al. [2022b,a]). *The order error $\bar{\phi}_q$ in any epoch q is defined as*

$$\bar{\phi}_q := \max_{n \in [N]} \left\{ \phi_q^n := \left\| \sum_{i=0}^{n-1} \left(\nabla f_{\pi_q(i)}(\mathbf{x}_q) - \nabla f(\mathbf{x}_q)\right) \right\|_p \right\}.$$

(3) As a measure of the quality of the permutation of examples.

The Order Error

Assumption 1 (Lu et al. [2022b,a]). *There exist nonnegative constants B and D such that for all $\mathbf{x}_q$ (the outputs of Algorithm 1),*

$$\left(\bar{\phi}_q\right)^2 \le B \left\|\nabla f(\mathbf{x}_q)\right\|^2 + D.$$

Existing works implicitly deal with the order error separately across epochs (as in Asm. 1).

Assumption 2. *There exist nonnegative constants $\{A_i\}_{i=1}^{q}$, $\{B_i\}_{i=0}^{q}$ and D such that for all $\mathbf{x}_q$ (the outputs of Algorithm 1),*

$$\left(\bar{\phi}_q\right)^2 \le \sum_{i=1}^{q} A_i \left(\bar{\phi}_{q-i}\right)^2 + \sum_{i=0}^{q} B_i \left\|\nabla f(\mathbf{x}_{q-i})\right\|^2 + D.$$

Asm. 2 explicitly characterizes the dependence between permutations across different epochs.

In particular, when $A_i=0$, $B_i=0$ for all $i=1,2,\ldots,q$, Asm. 2 reduces to Asm. 1.

Result 1: Main Theorem

Theorem 1. *Let the global objective function f be L-smooth and each local objective functions f_n be $L_{2,p}$-smooth and L_p-smooth ($p \geq 2$). Let $\nu \geq 0$ be a numerical constant. Suppose that there exist $\tilde{B}$ and $\tilde{D}$ such that for $0 \leq q \leq \nu - 1$,*

$$(\bar{\phi}_q)^2 \leq \tilde{B} \left\|\nabla f(\mathbf{x}_q)\right\|^2 + \tilde{D},$$

and there exist $\{A_i\}$, $\{B_i\}$ and D such that for $q \geq \nu$,

$$(\phi_q)^2 \leq \sum_{i=1}^{\nu} A_i (\bar{\phi}_{q-i})^2 + \sum_{i=0}^{\nu} B_i \left\|\nabla f(\mathbf{x}_{q-i})\right\|^2 + D.$$

The first ν epochs rely on Asm. 1, and the subsequent epochs rely on Asm. 2.

If $\gamma \leq \min\left\{ \frac{1}{LN}, \frac{1}{32 L_{2,p} N}, \frac{\sqrt{1-\sum_{i=1}^{\nu} A_i}}{4 L_{2,p} \sqrt{\sum_{i=0}^{\nu} B_i}}, \frac{\sqrt{1-\sum_{i=1}^{\nu} A_i}}{4 L_{2,p} \sqrt{\tilde{B}}}, \frac{1}{32 L_p N} \right\}$, *then*

It does not impose stronger constraints on γ.

Optimization term

Error terms

$$\frac{1}{Q} \sum_{q=0}^{Q-1} \left\|\nabla f(\mathbf{x}_q)\right\|^2 \leq \frac{5 F_0}{\gamma N Q} + c\gamma^2 L_{2,p}^2 \frac{1}{Q} \nu \tilde{D} + c\gamma^2 L_{2,p}^2 D,$$

where $c = {}^{10}\!/\!\left(1-\sum_{i=1}^{\nu} A_i\right)$ is a numerical constant.

Result 2: Case Studies

Table 7: Specific choices of A_i, B_i and D for different algorithms. The coefficients not explicitly specified equal 0. The numerical constants and polylogarithmic factors of B_i and D are omitted.

Algorithm	B_0	A_1	B_1	A_2	B_2	D	γ
AP	$N^2\alpha^2$	0	0	0	0	$N^2\varsigma^2$	$\gamma \lesssim \frac{1}{LN(1+\alpha)}$
RR/FlipFlop	$N^2\alpha^2$	0	0	0	0	$N\varsigma^2$	$\gamma \lesssim \frac{1}{LN(1+\alpha/\sqrt{N})}$
GraB-proto	0	$\frac{3}{4}^{(1)}$	$N^2+\alpha^2$	0	0	ς^2	$\gamma \lesssim \min\{\frac{1}{LN}, \frac{1}{L_{2,\infty}N(1+\alpha)}, \frac{1}{L_\infty N}\}$
GraB	0	$\frac{3}{5}^{(1)}$	$N^2+\alpha^2$	$\frac{1}{50}^{(1)}$	N^2	ς^2	$\gamma \lesssim \min\{\frac{1}{LN}, \frac{1}{L_{2,\infty}N(1+\alpha)}, \frac{1}{L_\infty N}\}$
PairGraB	0	$\frac{4}{5}^{(1)}$	$N^2+\alpha^2$	0	0	ς^2	$\gamma \lesssim \min\{\frac{1}{LN}, \frac{1}{L_{2,\infty}N(1+\alpha)}, \frac{1}{L_\infty N}\}$

[1] A_i may take other values as long as $\sum_{i=1}^{\nu} A_i < 1$ for GraBs.

It determines the error terms.

It does not impose stronger constraints on γ. It determines the optimization term.

$$(\phi_q)^2 \leq \sum_{i=1}^{\nu} A_i(\bar{\phi}_{q-i})^2 + \sum_{i=0}^{\nu} B_i \left\|\nabla f(\mathbf{x}_{q-i})\right\|^2 + D$$

Result 2: Case Studies

	Alg.	Lu et al. [2022b]	Koloskova et al. [2024][(1)]	This work
AP[(2)]	AP	$\frac{LF_0}{Q} + \left(\frac{LF_0 N\varsigma}{NQ}\right)^{\frac{2}{3}}$	$\frac{LF_0}{NQ} + \left(\frac{LF_0 N\varsigma}{NQ}\right)^{\frac{2}{3}} \wedge \left(\frac{LF_0 N\varsigma^2}{NQ}\right)^{\frac{1}{2}}$ (3)	$\frac{LF_0}{Q} + \left(\frac{LF_0 N\varsigma}{NQ}\right)^{\frac{2}{3}}$
IP	RR	$\frac{LF_0}{Q} + \left(\frac{LF_0 \sqrt{N}\varsigma}{NQ}\right)^{\frac{2}{3}}$	AP[(4)]	$\frac{LF_0}{Q} + \left(\frac{LF_0 \sqrt{N}\varsigma}{NQ}\right)^{\frac{2}{3}}$
	FlipFlop	–	–	$\frac{LF_0}{Q} + \left(\frac{LF_0 \sqrt{N}\varsigma}{NQ}\right)^{\frac{2}{3}}$
DP	GraBs	–	–	$\frac{\tilde{L}F_0 + (L_{2,\infty} F_0 \varsigma)^{\frac{2}{3}}}{Q} + \left(\frac{L_{2,\infty} F_0 \varsigma}{NQ}\right)^{\frac{2}{3}}$ (5)

This work is the first unified framework that includes DP.

Federated Learning (Beyond Permutation-based SGD)

Algorithm 1: Permutation-based SGD

Input: $\pi_0, \mathbf{x}_0$; **Output:** $\{\mathbf{x}_q\}$

for $q = 0, 1, \ldots, Q-1$ **do**
- $\mathbf{x}_q^0 \leftarrow \mathbf{x}_q$
- **for** $n = 0, 1, \ldots, N-1$ **do**
  - $\mathbf{x}_q^{n+1} \leftarrow \mathbf{x}_q^n - \gamma \nabla f_{\pi_q(n)}(\mathbf{x}_q^n)$
- $\mathbf{x}_{q+1} \leftarrow \mathbf{x}_q^N$
- $\pi_{q+1} \leftarrow \texttt{Permute}(\cdots)$

Algorithm 2: Regularized-participation FL

Input: $\pi_0, \mathbf{x}_0$; **Output:** $\{\mathbf{x}_q\}$

for $q = 0, 1, \ldots, Q-1$ **do**
- $\mathbf{w} \leftarrow \mathbf{x}_q$
- **for** $n = 0, 1, \ldots, N-1$ **do**
  - Initialize $\mathbf{x}_{q,0}^n \leftarrow \mathbf{w}$
  - **for** $k = 0, 1, \ldots, K-1$ **do**
    - $\mathbf{x}_{q,k+1}^n \leftarrow \mathbf{x}_{q,k}^n - \gamma \nabla f_{\pi_q(n)}(\mathbf{x}_{q,k}^n)$
  - $\mathbf{p}_q^n \leftarrow \mathbf{x}_{q,0}^n - \mathbf{x}_{q,K}^n$
  - **if** $(n+1) \mod S = 0$ **then**
    - $\mathbf{w} \leftarrow \mathbf{w} - \frac{1}{S}\sum_{s=0}^{S-1} \mathbf{p}_q^{n-s}$
- $\mathbf{x}_{q+1} \leftarrow \mathbf{x}_q - \eta(\mathbf{x}_q - \mathbf{w})$
- $\pi_{q+1} \leftarrow \texttt{Permute}(\cdots)$

We also develop a unified framework for regularized-participation FL with arbitrary permutations of clients

Thank You!

liyp25@mails.tsinghua.edu.cn, lvxinchen@bupt.edu.cn, zhenyuliu@sz.tsinghua.edu.cn