Tighter Lower Bounds for Shuffling SGD: Random Permutations and Beyond

Jaeyoung Cha Jaewook Lee Chulhee Yun

Graduate School of AI, KAIST

ICML 2023
Oral: A3 ML Theory, July 25th
Poster: Session 3, July 26th

KAIST AI Kim Jaechul Graduate School

OptiML Optimization & Machine Learning Laboratory

ICML International Conference On Machine Learning

Problem Setting

Objective: Finite-sum minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x}).$$

Ex. Supervised Learning

$f_i \leftarrow$ training loss of the i-th sample, $\boldsymbol{x} \leftarrow$ neural network parameters

Algorithm: Stochastic Gradient Descent (constant step size η)

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta \nabla f_{i(t)} \left(\boldsymbol{x}_{t-1} \right)$$

Question: Which choice of $i(t)$ achieves faster convergence?

With vs Without-replacement SGD

With-replacement SGD: Sample $i(t) \sim \text{Unif}(\{1, \dots, n\})$ i.i.d.

T

iteration t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
sample $f_{i(t)}$	f_1	f_2	f_2	f_4	f_1	f_4	f_5	f_3	f_5	f_1	f_4	f_2	f_4	f_4	f_1

With vs Without-replacement SGD

With-replacement SGD: Sample $i(t) \sim \mathrm{Unif}\left(\{1, \ldots, n\}\right)$ i.i.d.

T

iteration t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
sample $f_{i(t)}$	f_1	f_2	f_2	f_4	f_1	f_4	f_5	f_3	f_5	f_1	f_4	f_2	f_4	f_4	f_1

Most theoretical results focus on **with-replacement SGD**.

However, in real-world applications, **without-replacement SGD** is commonly used due to its simplicity and is believed to converge faster.

With vs Without-replacement SGD

Without-replacement SGD (Shuffling SGD)

1. In the k-th epoch, choose a permutation $\sigma_k : \{1, \ldots, n\} \to \{1, \ldots, n\}$
2. Use $f_{\sigma_k(j)}$ at the j-th iteration of k-th epoch, total $T = nK$ iterations

			$k=1$					$k=2$					$k=3$		
epoch k	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
iteration j	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
sample $f_{\sigma_k(j)}$	f_4	f_3	f_1	f_2	f_5	f_3	f_5	f_4	f_2	f_1	f_4	f_1	f_5	f_3	f_2
			n					n					n		

- Random Reshuffling (**SGD-RR**): choose σ_k **randomly**

With vs Without-replacement SGD

Without-replacement SGD (Shuffling SGD)

1. In the k-th epoch, choose a permutation $\sigma_k : \{1, \ldots, n\} \to \{1, \ldots, n\}$
2. Use $f_{\sigma_k(j)}$ at the j-th iteration of k-th epoch, total $T = nK$ iterations

	$k=1$					$k=2$					$k=3$				
epoch k	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
iteration j	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
sample $f_{\sigma_k(j)}$	f_4	f_3	f_1	f_2	f_5	f_3	f_5	f_4	f_2	f_1	f_4	f_1	f_5	f_3	f_2
	n					n					n				

- Random Reshuffling (**SGD-RR**): choose σ_k **randomly**
- Permutation-based SGD: can choose σ_k **arbitrarily** *Ex.* **GraB** [LGS22]

Our Contributions

We present the convergence lower bounds for both **SGD-RR** and **permutation-based SGD** on smooth f_i's with strongly-convex F.

Our lower bound results are...

1. the first to *completely* match the upper bounds for all factors
2. the first to generalize to *weighted average (end-of-epoch) iterates*

Especially, our lower bounds for arbitrary permutation-based SGD imply that GraB [LGS22] achieves the optimal rate!

Function Class

In this work, we mainly consider the function class $\mathcal{F}(L, \mu, \tau, \nu)$, which satisfies properties P1, P2, and P3.

P1. Strong convexity. F is μ*-strongly convex:* for $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$F(\boldsymbol{y}) \geq F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \left\| \boldsymbol{y} - \boldsymbol{x} \right\|^2 .$$

P2. Smoothness & Component Convexity. Each component function f_i is L*-smooth and convex:* for $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\left\| \nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y}) \right\| \leq L \left\| \boldsymbol{x} - \boldsymbol{y} \right\| ,$$
$$f_i(\boldsymbol{y}) \geq f_i(\boldsymbol{x}) + \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle .$$

Function Class

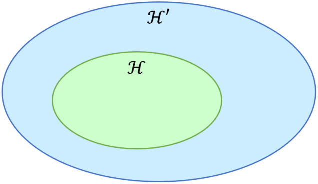
P3. Bounded Gradient Error. There exists $\tau, \nu \geq 0$ s.t. every component function f_i satisfies the following: for $\forall \boldsymbol{x} \in \mathbb{R}^d$,

$$\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq \tau \|\nabla F(\boldsymbol{x})\| + \nu.$$

Function Class

P3. Bounded Gradient Error. There exists $\tau, \nu \geq 0$ s.t. every component function f_i satisfies the following: for $\forall \boldsymbol{x} \in \mathbb{R}^d$,

$$\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq \tau \|\nabla F(\boldsymbol{x})\| + \nu.$$



Note that if $\mathcal{H} \subset \mathcal{H}'$,
then (LB for $\mathcal{H}$) $\leq$ (LB for $\mathcal{H}'$).

Showing the same lower bound for a **narrower** function class makes the result **stronger**.

Previous Results

Known Facts: (1) Without-replacement is faster than with-replacement, (2) Permutation-based SGD is faster than SGD-RR, in terms of **upper bounds**

With-replacement: $\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F^* = \mathcal{O}\left(\frac{\nu^2}{\mu n K}\right)$ [RSS12]

SGD-RR: $\mathbb{E}[F(\boldsymbol{x}_n^K)] - F^* = \tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 n K^2}\right)$ [AYS20] [MKR20]

Permutation-based: $F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$ by **GraB** [LGS22]

Previous Results

Known Facts: (1) Without-replacement is faster than with-replacement, (2) Permutation-based SGD is faster than SGD-RR, in terms of **upper bounds**

With-replacement: $\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F^* = \mathcal{O}\left(\frac{\nu^2}{\mu n K}\right)$ [RSS12]

SGD-RR: $\mathbb{E}[F(\boldsymbol{x}_n^K)] - F^* = \tilde{\mathcal{O}}\left(\frac{L^2 \nu^2}{\mu^3 n K^2}\right)$ [AYS20] [MKR20]

Permutation-based: $F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$ by **GraB** [LGS22]

Our Work: We provide matching **lower bounds** for SGD-RR and permutation-based SGD to guarantee that the upper bounds are tight.

Our Lower Bound Results on Random Reshuffling

- κ: condition number L/μ
- c_1, c_2: universal constant
- $\boldsymbol{x}_n^K$: final iterate
- Gray cell: lower bound result

Random Reshuffling				
Function Class	**Output**	**Ref**	**Rate**	**Assumptions**
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[MKR20]	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$	$K \gtrsim \kappa$
		Thm 3.1	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\kappa \geq c_1$, $K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}_{\text{tail}}$	Prop 3.4	$\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 3.3	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\eta \leq \frac{1}{c_2 nL}$, $\kappa \geq c_1$, $K \gtrsim \kappa$

Our Lower Bound Results on Random Reshuffling

- κ: condition number L/μ
- c_1, c_2: universal constant
- $\boldsymbol{x}_n^K$: final iterate
- Gray cell: lower bound result

Random Reshuffling				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[MKR20]	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 n K^2}\right)$	$K \gtrsim \kappa$
		Thm 3.1	$\Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$\kappa \geq c_1$, $K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}_{\text{tail}}$	Prop 3.4	$\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 3.3	$\Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$\eta \leq \frac{1}{c_2 n L}$, $\kappa \geq c_1$, $K \gtrsim \kappa$

Previous Lower Bound: $\Omega(\frac{\nu^2}{\mu n K^2})$ [YRS22]

Our Lower Bound Results on Random Reshuffling

- κ: condition number L/μ
- c_1, c_2: universal constant
- $\hat{\boldsymbol{x}} = \sum_{k=0}^{K} \alpha_k \boldsymbol{x}_n^k / \sum_{k=0}^{K} \alpha_k$
- $\hat{\boldsymbol{x}}_{\text{tail}} = \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \boldsymbol{x}_n^k / \left(K - \lceil \frac{K}{2} \rceil + 1\right)$

Random Reshuffling				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[MKR20]	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 n K^2}\right)$	$K \gtrsim \kappa$
		Thm 3.1	$\Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$\kappa \geq c_1$, $K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}_{\text{tail}}$	Prop 3.4	$\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 3.3	$\Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right)$	$\eta \leq \frac{1}{c_2 n L}$, $\kappa \geq c_1$, $K \gtrsim \kappa$

First lower bound results considering average end-of-epoch iterates!

Our Lower Bound Results on Permutation-based SGD

- κ: condition number L/μ
- Gray cell: lower bound result
- H: Herding bound $\mathcal{O}\left(\sqrt{d \log n}\right)$
- $\boldsymbol{x}_n^K$: final iterate
- $\hat{\boldsymbol{x}} = \sum_{k=0}^{K} \alpha_k \boldsymbol{x}_n^k / \sum_{k=0}^{K} \alpha_k$

Permutation-based SGD				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[LGS22]	$\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 4.1	$\Omega\left(\frac{L \nu^2}{\mu^2 n^2 K^2}\right)$	-
$\mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$	$\boldsymbol{x}_n^K$	Prop 4.6	$\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$n \geq H,\ K \gtrsim \kappa(\tau + 1)$
	$\hat{\boldsymbol{x}}$	Thm 4.5	$\Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$\tau = \kappa \geq 8n,\ K \gtrsim \kappa^2$

Our Lower Bound Results on Permutation-based SGD

- κ: condition number L/μ
- Gray cell: lower bound result
- H: Herding bound $\mathcal{O}\left(\sqrt{d \log n}\right)$
- $\boldsymbol{x}_n^K$: final iterate
- $\hat{\boldsymbol{x}} = \sum_{k=0}^{K} \alpha_k \boldsymbol{x}_n^k / \sum_{k=0}^{K} \alpha_k$

Permutation-based SGD				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[LGS22]	$\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 4.1	$\Omega\left(\frac{L \nu^2}{\mu^2 n^2 K^2}\right)$	-
$\mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$	$\boldsymbol{x}_n^K$	Prop 4.6	$\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$n \geq H$, $K \gtrsim \kappa(\tau + 1)$
	$\hat{\boldsymbol{x}}$	Thm 4.5	$\Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$\tau = \kappa \geq 8n$, $K \gtrsim \kappa^2$

Previous Lower Bound: $\Omega(\frac{\nu^2}{L n^3 K^2})$ [RLP22]

Our Lower Bound Results on Permutation-based SGD

- κ: condition number L/μ
- Gray cell: lower bound result
- H: Herding bound $\mathcal{O}\left(\sqrt{d \log n}\right)$
- $\boldsymbol{x}_n^K$: final iterate
- $\hat{\boldsymbol{x}} = \sum_{k=0}^{K} \alpha_k \boldsymbol{x}_n^k / \sum_{k=0}^{K} \alpha_k$

Permutation-based SGD				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	[LGS22]	$\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 4.1	$\Omega\left(\frac{L \nu^2}{\mu^2 n^2 K^2}\right)$	-

Remark

The lower bound in Thm 4.1 holds for ***arbitrary*** **sampling methods.**

Previous Lower Bound: $\Omega(\frac{\nu^2}{L n^3 K^2})$ [RLP22]

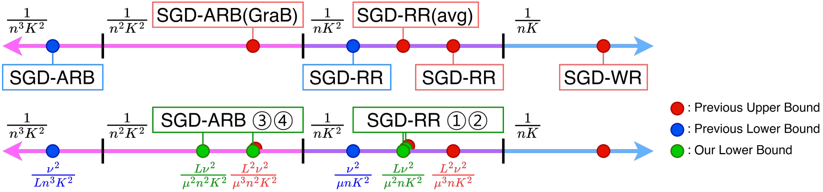
Our Lower Bound Results on Permutation-based SGD

- κ: condition number L/μ
- Gray cell: lower bound result
- H: Herding bound $\mathcal{O}\left(\sqrt{d \log n}\right)$
- $\boldsymbol{x}_n^K$: final iterate
- $\hat{\boldsymbol{x}} = \sum_{k=0}^{K} \alpha_k \boldsymbol{x}_n^k / \sum_{k=0}^{K} \alpha_k$

Permutation-based SGD				
Function Class	Output	Ref	Rate	Assumptions
$\mathcal{F}(L,\mu,0,\nu)$	$\boldsymbol{x}_n^K$	[LGS22]	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3n^2K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Thm 4.1	$\Omega\left(\frac{L\nu^2}{\mu^2n^2K^2}\right)$	-
$\mathcal{F}_{\text{PŁ}}(L,\mu,\tau,\nu)$	$\boldsymbol{x}_n^K$	Prop 4.6	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3n^2K^2}\right)$	$n \geq H,\ K \gtrsim \kappa(\tau+1)$
	$\hat{\boldsymbol{x}}$	Thm 4.5	$\Omega\left(\frac{L^2\nu^2}{\mu^3n^2K^2}\right)$	$\tau = \kappa \geq 8n,\ K \gtrsim \kappa^2$

$\mathcal{F}_{\text{PŁ}}$: No component convexity & relaxes strong convexity to PŁ condition

Summary



$\frac{1}{n^3K^2}$
$\frac{1}{n^2K^2}$
SGD-ARB(GraB)
$\frac{1}{nK^2}$
SGD-RR(avg)
$\frac{1}{nK}$
SGD-ARB
SGD-RR
SGD-RR
SGD-WR
$\frac{1}{n^3K^2}$
$\frac{1}{n^2K^2}$
SGD-ARB ③④
$\frac{1}{nK^2}$
SGD-RR ①②
$\frac{1}{nK}$
$\frac{\nu^2}{Ln^3K^2}$
$\frac{L\nu^2}{\mu^2n^2K^2}$
$\frac{L^2\nu^2}{\mu^3n^2K^2}$
$\frac{\nu^2}{\mu nK^2}$
$\frac{L\nu^2}{\mu^2nK^2}$
$\frac{L^2\nu^2}{\mu^3nK^2}$
: Previous Upper Bound
: Previous Lower Bound
: Our Lower Bound

Summary

We also have...

- The first lower bound that applies to *convex functions* and perfectly matches the previously known upper bound by [MKR20]
- Some novel *upper bound* results, such as Propositions 3.4 and 4.6

For more details, please check the QR link to our paper below...
or even better, come and visit our poster tomorrow!

LINK TO OUR PAPER

Poster Session 3

Date: July 26th (Wed)
Time: 11 a.m. - 12:30 p.m.
Place: Exhibit Hall 1 #713

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra.
SGD with shuffling: Optimal rates without component convexity and large epoch requirements.
In *Advances in Neural Information Processing Systems*, 2020.

Yucheng Lu, Wentao Guo, and Christopher De Sa.
GraB: Finding provably better data permutations than random reshuffling.
In *Advances in Neural Information Processing Systems*, 2022.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik.
Random reshuffling: Simple analysis with vast improvements.
In *Advances in Neural Information Processing Systems*, 2020.

Shashank Rajput, Kangwook Lee, and Dimitris Papailiopoulos.
Permutation-based SGD: Is random optimal?
In *International Conference on Learning Representations*, 2022.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan.
Making gradient descent optimal for strongly convex stochastic optimization.
In *International Conference on Machine Learning*, 2012.

Chulhee Yun, Shashank Rajput, and Suvrit Sra.
Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond.
In *International Conference on Learning Representations*, 2022.