Incremental Gradient Descent with Small Epoch Counts is Surprisingly Slow on Ill-Conditioned Problems

Yujun Kim*, Jaeyoung Cha*, Chulhee Yun
Graduate School of AI, KAIST
{kyujun02, chajaeyoung, chulhee.yun}@kaist.ac.kr

KAIST AI Kim Jaechul Graduate School

ICML International Conference On Machine Learning

Problem Setup

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

Method: $x_t = x_{t-1} - \eta \nabla f_{i(t)}(x_{t-1})$

- With-replacement SGD: choose $i(t) \sim Unif(\{1,\dots,n\})$
- Permutation-based SGD: choose $i(t)$ sequentially from the permutation σ_k

  $\sigma_1(1),\sigma_1(2),\cdots,\sigma_1(n), \quad \sigma_2(1),\sigma_2(2),\cdots,\sigma_2(n), \quad \cdots \quad ,\sigma_k(1),\sigma_k(2),\cdots,\sigma_k(n)$
  - Incremental Gradient Descent (IGD): $\sigma_k = id_n$ (identity permutation)
  - Random Reshuffling (RR): $\sigma_k \sim Unif(S_n)$ (random permutation)
  - Gradient Balancing (GraB): σ_k is manually selected by previous observations

***Question**. What is the convergence rate of permutation-based SGD in the small epoch regime?*

As an initial step toward understanding permutation-based SGD in the small epoch regime, we focus on IGD—the simplest deterministic variant.

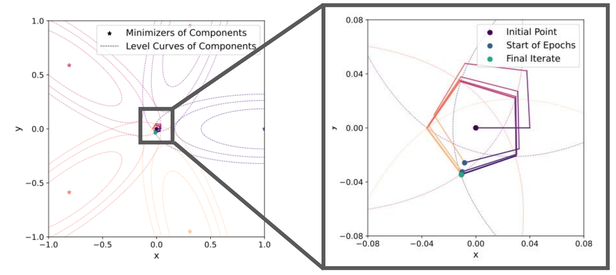
Notation

- number of components: n
- number of epochs: K
- condition number: $\kappa = L/\mu$
- asymptotic notation: $O(\cdot), \Omega(\cdot), \tilde{O}(\cdot)$ (hide polylog terms)
- step size: η
- permutation at k-th epoch: σ_k
- initial point: $x_0 = x_0^1$
- i-th iterate of k-th epoch: x_i^k

Assumption

1. (**Strong Convexity**) F is μ-strongly convex.
2. (**Smoothness**) Each component function f_i is L-smooth.
3. (**Bounded Gradient Errors**, LB) For all $x \in \mathbb{R}^d$ and $i \in [n]$, $\|\nabla f_i(x) - \nabla F(x)\| \le G + P\|\nabla F(x)\|$.
4. (**Bounded Gradients at Optimum**, UB) For all $i \in [n]$, $\|\nabla f_i(x^*)\| \le G_*$.

Key Idea



Strategy. Place the minimizer of each component function at a vertex of regular n-gon.

Result. By rotational symmetry, the iterates trace a regular n-gon.

Gaps in Existing Theory

- When K is sufficiently large ($K \gtrsim \kappa$, **large epoch regime**), the convergence rates of permutation-based SGD methods are well-studied:

  (fast) GraB $\tilde{O}\left(\frac{1}{n^2K^2}\right)$ < RR $\tilde{O}\left(\frac{1}{nK^2}\right)$ < with-replacement SGD $\tilde{O}\left(\frac{1}{T}\right)$, IGD $\tilde{O}\left(\frac{1}{K^2}\right)$ (slow)

- When K is small ($K \le \kappa$, **small epoch regime**), little is known:

  For quadratics, RR $\tilde{O}\left(\frac{1}{nK}\right)$ = with-replacement SGD $\tilde{O}\left(\frac{1}{T}\right)$ [Safran & Shamir., 2021]

- Existing analyses either require large K, or become loose when K is small.

Prior Works	Assump.	Alg.	Conv. Rate	Note
Theorem 1 [Mishchenko et al., 2020]	f_i: str. convex	RR	$\tilde{O}(\exp(-nK/\kappa) + \kappa^3/nK^2)$	gap of $\Omega(\kappa \sim \kappa n)$ to LB when $K \le \kappa$
Theorem 2 [Mishchenko et al., 2020]	f_i: convex	RR	$\tilde{O}(\exp(-K/\kappa) + \kappa^3/nK^2)$	exp. term remain large when $K \le \kappa$
Theorem 4.6 [Liu & Zhou, 2024]	f_i: convex	RR	$\tilde{O}(\exp(-K/\kappa)/K + \kappa^2/nK^2)$	exp. term remain large when $K \le \kappa$
Theorem 1 [Nguyen et al., 2021]	-	Any	$\tilde{O}(\kappa^3/K^2)$	require $K \gtrsim \kappa^2$
Theorem 4.6 [Liu & Zhou, 2024]	f_i: convex	Any	$\tilde{O}(\exp(-K/\kappa)/K + \kappa^2/K^2)$	exp. term remain large when $K \le \kappa$
Theorem 1 [Lu et al., 2023]	-	GraB	$\tilde{O}(\kappa^3/n^2K^2)$	require $K \gtrsim \kappa$

(Assuming F is strongly-convex and each f_i is smooth; set $L = 1$)

Main Results in Small Epoch Regime ($\kappa/n \le K \le \kappa$)

f_i shares the **same Hessian**: $\nabla^2 f_i(x) = \nabla^2 F(x)$

Theorem 3.1 (LB). There exists a function F satisfying assumption 1, 2, 3, such that for any constant step size η, IGD satisfies

$$F(x_n^K) - F(x^*) = \Omega\left(\frac{G^2}{\mu K}\right).$$

Theorem 3.2 (UB). Suppose F is 1-dimensional and satisfy assumption 1, 2, 4. Then, there exists η such that any permutation-based SGD satisfies

$$F(x_n^K) - F(x^*) = \tilde{O}\left(\frac{G_*^2}{\mu K}\right).$$

f_i are **strongly convex**

Theorem 3.3 (LB). There exists a function F satisfying assumption 1, 2, 3, such that for any constant step size η, IGD satisfies

$$F(x_n^K) - F(x^*) = \Omega\left(\frac{LG^2}{\mu^2}\min\left\{1, \frac{\kappa^2}{K^4}\right\}\right).$$

Match when $K = \Theta(\sqrt{\kappa})$

Proposition 3.4 (UB). Suppose F satisfies assumption 1, 2, 4. Then, there exists η such that any permutation-based SGD satisfies

$$F(x_n^K) - F(x^*) = \tilde{O}\left(\frac{L^2G_*^2}{\mu^3K^2}\right).$$

f_i can be **nonconvex**

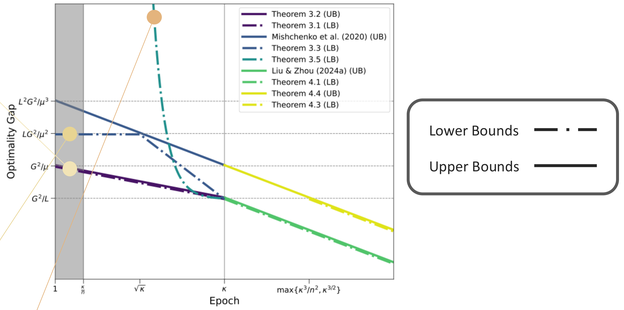
Theorem 3.5 (LB). There exists a function F satisfying assumption 1, 2, 3, such that for any constant step size η, IGD satisfies

$$F(x_n^K) - F(x^*) = \Omega\left(\frac{G^2}{L}\left(1 + \frac{L}{2\mu nK}\right)^n\right).$$

When $K = \Theta(\kappa/n)$, the rate becomes $\frac{G^2}{L}(1+c)^n$

⇒ **No polynomial upper bound exists** in this small epoch regime!

Summary



- Present convergence rates in the **small epoch regime** under
  ① shared Hessian ② strongly convex ③ nonconvex components.
- Present tight rates in the **large epoch regime** under
  ① strongly convex ② nonconvex components.
- In the **small epoch regime**:
  - when **strongly convex**, IGD converges slower than expected.
  - when **nonconvex**, IGD can suffer exponential slowdown.