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

Yujun Kim

Febuary 7th Fri, 2025

Optimization & Machine Learning Laboratory

IGD with Small Epoch Counts is Surprisingly Slow

What is Permutation-Based SGD?

Why we Divide Small and Large Epoch Regime?

What Happens for the Worst Permutation-Based SGD (IGD)?

What is Permutation-Based SGD?

Finite Sum Minimization

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

Stochastic Gradient Descent(SGD)

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

Stochastic Gradient Descent(SGD)

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

With-Replacement SGD

Permutation-Based SGD

With-Replacement SGD

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

$$i_t \overset{\text{i.i.d.}}{\sim} \text{Uniform}([n])$$

$$i_1, i_2, \cdots, i_T$$

Permutation-Based SGD

$$\boldsymbol{x}_i^k = \boldsymbol{x}_{i-1}^k - \eta \nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)$$

$$\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$$

$i \in [n]$ counts for index of component

$k \in [K]$ counts for epoch

$\sigma_k : [n] \to [n]$ is a permutation

$$\underbrace{\sigma_1(1), \sigma_1(2), \cdots, \sigma_1(n)}_{1^{st}\text{ Epoch}}, \cdots, \underbrace{\sigma_K(1), \sigma_K(2), \cdots, \sigma_K(n)}_{K^{th}\text{ Epoch}}$$

Permutation-Based SGD

Incremental Gradient Descent(IGD)

Random Reshuffling(RR)

Gradient Balancing(GraB)

$\vdots$

Permutation-Based SGD

Incremental Gradient Descent(IGD): $\sigma_k = \mathrm{id}_n$

Random Reshuffling(RR)

Gradient Balancing(GraB)

Permutation-Based SGD

Incremental Gradient Descent(IGD): $\sigma_k = \mathrm{id}_n$

Random Reshuffling(RR): $\sigma_k \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(S_n)$

Gradient Balancing(GraB)

Permutation-Based SGD

Incremental Gradient Descent(IGD): $\sigma_k = \mathrm{id}_n$

Random Reshuffling(RR): $\sigma_k \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(S_n)$

Gradient Balancing(GraB): Choose σ_k based on previous observations

are known to be faster than with-replacement SGD when K is sufficiently large

What happens when K is small?

Problem Setting

F is L-smooth and μ-strongly convex

f_i is L-smooth

Different convexity assumptions for f_i

Why we Divide Small and Large Epoch Regime?

Iteration-wise analysis of With-Replacement SGD

Allow large η (Similar to GD)

v.s.

Epoch-wise analysis of Permutation-Based SGD

Require small η (Relative to GD)

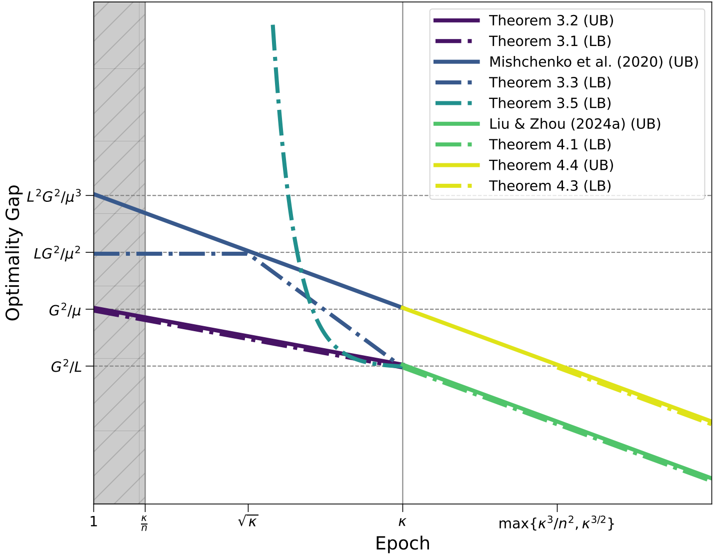
$\eta \gtrsim 1/K$ for sufficient contraction

η should be small for epoch-wise analysis

Small Epoch $K \lesssim \kappa$ v.s. Large Epoch $K \gtrsim \kappa$

What Happens for the Worst Permutation-Based SGD (IGD)?

Overview(IGD)



Small Epoch - Identical Hessian

In the small epoch regime,

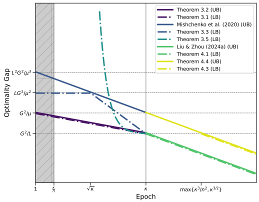
There exist F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\nabla^2 f_i \equiv \nabla^2 F$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{G^2}{\mu K}.$$

For any 1-dimensional F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\nabla^2 f_i \equiv \nabla^2 F$ and for any x_0, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(x_n^K) - F(x^*) \lesssim \frac{G^2}{\mu K}.$$

Small Epoch - Identical Hessian



In the small epoch regime,

There exist F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\nabla^2 f_i \equiv \nabla^2 F$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{G^2}{\mu K}.$$

For any 1-dimensional F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\nabla^2 f_i \equiv \nabla^2 F$ and for any x_0, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(x_n^K) - F(x^*) \lesssim \frac{G^2}{\mu K}.$$

What happens if we allow distinct Hessians while maintaining the component strong convexity?

Small Epoch - Strongly Convex

In the small epoch regime,

There exist F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

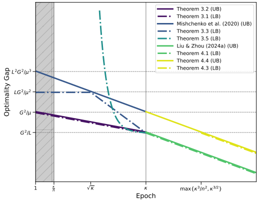
$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{LG^2}{\mu^2} \min\left\{1, \frac{\kappa^2}{K^4}\right\}.$$

Mishchenko et al., 2020

For any F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and for any $\boldsymbol{x}_0$, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \lesssim \frac{L^2G^2}{\mu^3K^2}.$$

Small Epoch - Strongly Convex



In the small epoch regime,

There exist F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

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

Mishchenko et al., 2020

For any F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and for any $\boldsymbol{x}_0$, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \lesssim \frac{L^2G^2}{\mu^3K^2}.$$

What if we allow concave components?

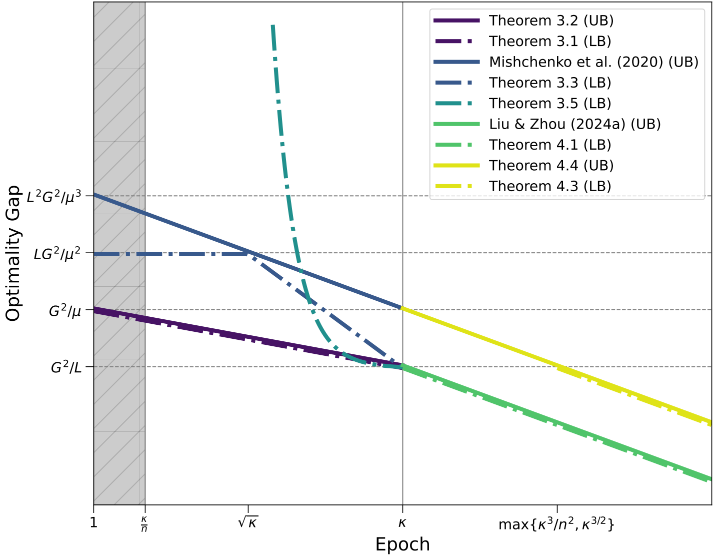
Small Epoch - Concave

In the small epoch regime,

There exist F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq G + 3\|\nabla F(\boldsymbol{x})\|$ such that for any $\eta > 0$, IGD starting at $\boldsymbol{x}_0 = (D, 0)$ results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \min\left\{\mu D^2, \frac{G^2}{L}\left(1 + \frac{L}{2\mu n K}\right)^{\frac{n}{2}}\right\}.$$

Small Epoch



Theorem 3.2 (UB)
Theorem 3.1 (LB)
Mishchenko et al. (2020) (UB)
Theorem 3.3 (LB)
Theorem 3.5 (LB)
Liu & Zhou (2024a) (UB)
Theorem 4.1 (LB)
Theorem 4.4 (UB)
Theorem 4.3 (LB)
Optimality Gap
L^2G^2/μ^3
LG^2/μ^2
G^2/μ
G^2/L
1
$\frac{\kappa}{n}$
$\sqrt{\kappa}$
κ
$\max\{\kappa^3/n^2, \kappa^{3/2}\}$
Epoch

Large Epoch - Convex

In the large epoch regime,

There exist F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

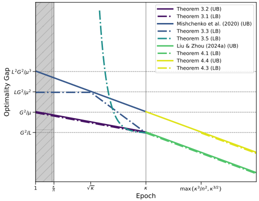
$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{LG^2}{\mu^2 K^2}.$$

Liu and Zhou, 2024

For any F and convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and for any $\boldsymbol{x}_0$, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \lesssim \frac{LG^2}{\mu^2 K^2}.$$

Large Epoch - Convex



In the large epoch regime,

There exist F and μ-strongly convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{LG^2}{\mu^2 K^2}.$$

Liu and Zhou, 2024

For any F and convex f_i satisfying $\|\nabla f_i(\boldsymbol{x}^*)\| \leq G$ and for any $\boldsymbol{x}_0$, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \lesssim \frac{LG^2}{\mu^2 K^2}.$$

What if we allow concave components?

Large Epoch - Concave

In the large epoch regime,

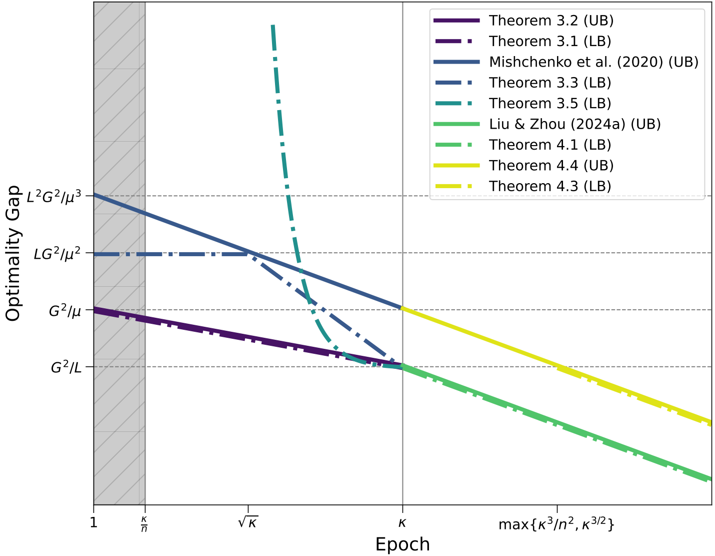
Under extra condition on κ, n, and K, there exists F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq G + \kappa \|\nabla F(\boldsymbol{x})\|$ and $\boldsymbol{x}_0$, such that for any $\eta > 0$, IGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \gtrsim \frac{L^2 G^2}{\mu^3 K^2}.$$

Suppose $K \gtrsim (1+P)\kappa$. For any F and f_i satisfying $\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq G + P \|\nabla F(\boldsymbol{x})\|$, there exists $\eta > 0$ such that **any** permutation-based SGD results

$$F(\boldsymbol{x}_n^K) - F(\boldsymbol{x}^*) \lesssim \frac{L^2 G^2}{\mu^3 K^2}.$$

Small v.s. Large Epoch



Convergence of IGD in small epoch is significantly slow,
even under component strong convexity

Nonconvex components slowdown convergence even more

What are the Convergence Rate of Other Permutation-Based Methods in Small Epoch Regime?

Can we Design Better Permutation in Small Epoch Regime?

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In Forty-first International Conference on Machine Learning, 2024. URL `https://openreview.net/forum?id=Xdy9bjwHDu`.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. Advances in Neural Information Processing Systems, 33:17309–17320, 2020.