In-Database Machine Learning with CorgiPile: Stochastic Gradient Descent without Full Data Shuffle

Anonymous Author(s)

ABSTRACT

Stochastic gradient descent (SGD) is the cornerstone of modern ML systems. Despite its computational efficiency, SGD requires random data access that is inherently inefficient when implemented in systems that rely on *block-addressable secondary storage* such as HDD and SSD, e.g., in-DB ML systems and TensorFlow/PyTorch over large files. To address this impedance mismatch, various data shuffling strategies have been proposed to balance the convergence rate of SGD (which favors randomness) and its I/O performance (which favors sequential access).

In this paper, we first conduct a systematic empirical study on existing data shuffling strategies, which reveals that all existing strategies have room for improvement—many of them can suffer in terms of I/O performance *or* convergence rate. With this in mind, we propose a simple but novel *hierarchical* data shuffling strategy, `CorgiPile`. Compared with existing strategies, `CorgiPile` *avoids* a full data shuffle while maintaining *comparable convergence rate* of SGD as if a full shuffle was performed. We provide a non-trivial theoretical analysis of `CorgiPile` on its convergence behavior. We further integrate `CorgiPile` into PostgreSQL by introducing three new *physical* operators with their implementation details and optimizations. Our experimental results show that `CorgiPile` can achieve comparable convergence rate with the full shuffle based SGD, and 2×-12.8× faster than two state-of-the-art in-DB ML systems, Apache MADlib and Bismarck, on both HDD and SSD.

1 CORGIPILE

As illustrated in the previous section, data shuffling strategies used by existing systems can be suboptimal when dealing with clustered data. Although recent efforts have significantly improved over baseline methods, there is still large room for improvement. Inspired by these previous efforts and their convergence and performance profiles, we present a simple but novel data shuffling strategy, namely `CorgiPile` in this section. The key idea of `CorgiPile` lies the following two-level hierarchical shuffling mechanism:

> *We first randomly select a set of blocks (each block contains a set of tuples) and put them into an in-memory buffer; we then randomly shuffle all tuples in the buffer and use them for SGD.*

Despite its simplicity, `CorgiPile` is highly effective. In terms of hardware efficiency, when the block size is large enough (e.g., 10MB+), a random access on the *block* level can be as efficient as a *sequential* scan. In terms of statistical efficiency, as we will show, *given the same buffer size*, `CorgiPile` converges much better than *Sliding-Window* and *MRS*. Nevertheless, both the convergence analysis and its integration into PostgreSQL are non-trivial. In the following, we first describe the `CorgiPile` algorithm precisely and then present a theoretical analysis on its convergence behavior.

SIGMOD '22, June 12–17, 2022, Philadelphia, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Algorithm 1 `CorgiPile` Algorithm

Input: N blocks with m total tuples, total epochs S ($S \geq 1$), $a \geq 1$, $F(\cdot) = \frac{1}{m}\sum_{i=1}^{m} f_i(\cdot)$.
Initialize $\mathbf{x}_0^0$;
for $s = 0, \cdots, S$ **do**
  Randomly pick n blocks without replacement, each containing b tuples. Load these blocks into the buffer;
  Shuffle tuple indices among all n blocks in the buffer and obtain the permutation $\boldsymbol{\psi}_s$;
  for $k = 1, ..., bn$ **do**
    Update $\mathbf{x}_k^s = \mathbf{x}_{k-1}^s - \eta_s \nabla f_{\psi_s(k)}\left(\mathbf{x}_{k-1}^s\right)$;
  end for
  $\mathbf{x}_0^{s+1} = \mathbf{x}_{bn}^s$;
end for
Return x_{bn}^S;

Notations and Definitions. The following is a list of notations and definitions that we will use:

- $\|\cdot\|$, the ℓ_2-norm for vectors and the spectral norm for matrices;
- $a_s \lesssim g_s$, meaning $a_s \leq C g_s$ with a certain constant C for all s;
- N, the total number of blocks ($N \geq 2$);
- n, the buffer size (i.e., the number of blocks kept in the buffer);
- b, the size (number of tuples) of each data block;
- B_l, the set of tuple indices in the l-th block ($l \in [N]$ and $|B_l| = b$);
- m, the number of tuples for the finite-sum objective ($m = Nb$);
- $f_i(\cdot)$, the function associated with the i-th tuple;
- $\nabla F(\cdot)$ and $\nabla f_i(\cdot)$, the gradients of the functions $F(\cdot)$ and $f_i(\cdot)$;
- $H_i(\cdot) := \nabla^2 f_i(\cdot)$, the Hessian matrix of the function $f_i(\cdot)$;
- $\mathbf{x}^*$, the global minimizer of the function $F(\cdot)$;
- $\mathbf{x}_k^s$, the model $\mathbf{x}$ in the k-th iteration at the s-th epoch;
- μ-strongly convexity: function $F(\mathbf{x})$ is μ-strongly convex if $\forall \mathbf{x}, \mathbf{y}$,

$$F(\mathbf{x}) \geq F(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla F(\mathbf{y}) \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2. \quad (1)$$

1.1 The `CorgiPile` Algorithm

Algorithm 1 illustrates the details of `CorgiPile`. At each epoch (say, the s-th epoch), `CorgiPile` runs the following steps:

(1) (**Sample**) Randomly sample n blocks out of N data blocks *without replacement* and load the n blocks into the buffer.
(2) (**Shuffle**) Shuffle all tuples in the buffer. We use $\boldsymbol{\psi}_s$ to denote an ordered set, whose elements are the indices of the shuffled tuples at the s-th epoch. The size of $\boldsymbol{\psi}_s$ is bn, where b is the number of tuples per block. $\boldsymbol{\psi}_s(k)$ is the k-th element in $\boldsymbol{\psi}_s$.

(3) **(Update)** Perform gradient descent by scanning each tuple with the shuffle indices in $\boldsymbol{\psi}_s$, yielding the updating rule

$$\mathbf{x}_k^s = \mathbf{x}_{k-1}^s - \eta_s \nabla f_{\boldsymbol{\psi}_s(k)}\left(\mathbf{x}_{k-1}^s\right),$$

where $\nabla f_{\boldsymbol{\psi}_s(k)}(\cdot)$ is the gradient function averaging the gradients of all samples in the tuple indexed by $\boldsymbol{\psi}_s(k)$, and η_s is the learning rate for gradient descent at the epoch s. The parameter update is performed for all $k = 1, ..., bn$ in one epoch.

1.2 Convergence Analysis

Despite its simplicity, the convergence analysis of `CorgiPile` is not trivial—even reasoning about the convergence of SGD with *sample without replacement* is a open question for decades [1–4], not to say a hierarchical sampling scheme like ours. Luckily, a recent theoretical advancement [2] provides us with the technical language to reason about `CorgiPile`'s convergence. In the following, we present a novel theoretical analysis for `CorgiPile`.

Note that in our analysis, one epoch represents going through all the tuples in the sampled n blocks.

Assumption 1. *We make the following standard assumptions:*

(1) $F(\cdot)$ and $f_i(\cdot)$ are twice continuously differentiable.
(2) L-Lipschitz gradient: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for all $i \in [m]$.
(3) L_H-Lipschitz Hessian matrix: $\|H_i(\mathbf{x}) - H_i(\mathbf{y})\| \leq L_H\|\mathbf{x}-\mathbf{y}\|$ for all $i \in [m]$.
(4) Bounded gradient: $\|\nabla f_i(\mathbf{x}_k^s)\| \leq G$ for all $i \in [m], k \in [K-1]$, and $s \in \{0, 1 \ldots, S\}$.
(5) Bounded Variance: $\mathbb{E}_\xi[\|\nabla f_\xi(\mathbf{x}) - \nabla F(\mathbf{x})\|^2] = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \sigma^2$ where ξ is the random variable that takes the values in $[m]$ with equal probability $1/m$.

Factor h_D. In our analysis, we use the factor h_D to characterize the upper bound of a block-wise data variance:

$$\frac{1}{N}\sum_{l=1}^{N}\left\|\nabla f_{B_l}(\mathbf{x}) - \nabla F(\mathbf{x})\right\|^2 \leq h_D\frac{\sigma^2}{b},$$

where $b = |B_l|$ is the size of each data block (recall the definition of b). Here, h_D is an essential parameter to measure the "cluster" effect within the original data blocks. Let's consider two extreme cases: 1) ($h_D = 1$) all samples in the data set are fully shuffled, such that the data in each block follows the same distribution; 2) ($h_D = b$) samples are well clustered in each block, for example, all samples in the same block are identical. Therefore, the larger h_D, the more "clustered" the data.

We now present the results for both strongly convex objectives (corresponding to generalized linear models) and non-convex objectives (corresponding the deep learning models) respectively, in order to show the correctness and efficiency of `CorgiPile`.

Strongly convex objective. We first show the result for strongly convex objective that satisfies the strong convexity condition (1).

Theorem 1. *Suppose that $F(\mathbf{x})$ is a smooth and μ-strongly convex function. Let $T = Snb$, that is, the total number of samples used in training and $S \geq 1$ is the number of tuples iterated, and choosing $\eta_s = \frac{6}{bn\mu(s+a)}$ where $a \geq \max\left\{\frac{8LG+24L^2+28L_HG}{\mu^2}, \frac{24L}{\mu}\right\}$, under Assumption 1, `CorgiPile` has the following convergence rate*

$$\mathbb{E}[F(\bar{\mathbf{x}}_S) - F(\mathbf{x}^*)] \lesssim (1-\alpha)h_D\sigma^2\frac{1}{T} + \beta\frac{1}{T^2} + \gamma\frac{m^3}{T^3}, \qquad (2)$$

where $\bar{\mathbf{x}}_S = \frac{\sum_s (s+a)^3\mathbf{x}_s}{\sum_s (s+a)^3}$, and

$$\alpha := \frac{n-1}{N-1}, \beta := \alpha^2 + (1-\alpha)^2(b-1)^2, \gamma := \frac{n^3}{N^3}.$$

Tightness. The convergence rate of `CorgiPile` is tight in the following sense:

- $\alpha = 1$: It means that $n = N$, i.e., all tuples are fetched to the buffer. Then `CorgiPile` reduces to full-shuffle SGD [2]. In this case, the upper bound in Theorem 1 is $O(1/T^2 + m^3/T^3)$, which matches the result of the full shuffle SGD algorithm [2].
- $\alpha = 0$: It means that $n = 1$, i.e., only sampling one block each time. Then `CorgiPile` is very close to *mini-batch* SGD (by viewing a block as a mini-batch), except that the model is updated once per data tuple. Ignoring the higher-order terms in (2), our upper bound $O(h_D\sigma^2/T)$ is consistent with that of mini-batch SGD.

Comparison to vanilla SGD. In vanilla SGD, we only randomly select one tuple from the database to update the model. It admits the convergence rate $O(\sigma^2/T)$. Comparing to the leading term $(1-\alpha)h_D(\sigma^2/T)$ in (2) for our algorithm, if $n \gg (h_D - 1)(N-1)/h_D + 1$ (for $h_D > 0$), $(1-\alpha)h_D$ will be much smaller than 1, indicating that our algorithm outperforms vanilla SGD in terms of sample complexity. It is also worth noting that, even if n is small, `CorgiPile` may still significantly outperform vanilla SGD. Assuming that reading a random single tuple incurs an overhead of $t_{\text{sr}} + t_{\text{t}}$ and reading a block of b tuples incurs an overhead of $t_{\text{sr}} + bt_{\text{t}}$, where t_{sr} is the "seek and rotate" time and t_{t} is the time that one needs to transfer a single tuple. To reach an error of ϵ, vanilla SGD requires, in physical time,

$$O\left(\frac{\sigma^2}{\epsilon}t_{\text{sr}} + \frac{\sigma^2}{\epsilon}t_{\text{t}}\right),$$

whereas `CorgiPile` requires

$$O\left((1-\alpha)\frac{h_D}{b}\cdot\frac{\sigma^2}{\epsilon}t_{\text{sr}} + (1-\alpha)h_D\cdot\frac{\sigma^2}{\epsilon}t_{\text{t}}\right).$$

Because $(1-\alpha)\frac{h_D}{b} < 1$, `CorgiPile` always provides benefit over vanilla SGD in terms of the seek and rotate time t_{sr}. When t_{sr} dominates the transfer time t_{t}, which is often the case in practice, `CorgiPile` can outperform vanilla SGD even for small buffers.

Non-convex objective. We further conduct an analysis on objectives that are non-convex or satisfy the Polyak-Łojasiewicz condition, which leads to similar insights on the behavior of `CorgiPile`. Due to space constraints, we only present the the non-convex case below, and leave the rest to the full version of this paper.

Theorem 2. *Suppose that $F(\mathbf{x})$ is a smooth function. Letting $T = Snb$ be the number of tuples iterated, under Assumption 1, `CorgiPile` has the following convergence rate:*

(1) *When* $\alpha \le \frac{N-2}{N-1}$, *choosing* $\eta_s = \frac{1}{\sqrt{bn(1-\alpha)h_D\sigma^2 S}}$ *and assuming* $S \ge \frac{bn(\frac{104}{3}L+\frac{4}{3}L_H)^2}{\sigma^2(1-\alpha)h_D}$, *we have*

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\|\nabla F(\mathbf{x}_0^s)\|^2 \lesssim (1-\alpha)^{1/2}\frac{\sqrt{h_D}\sigma}{\sqrt{T}} + \beta\frac{1}{T} + \gamma\frac{m^3}{T^{\frac{3}{2}}},$$

where the factors are defined as

$$\alpha := \frac{n-1}{N-1}, \beta := \frac{\alpha^2}{1-\alpha}\frac{1}{h_D\sigma^2} + (1-\alpha)\frac{(b-1)^2}{h_D\sigma^2},$$

$$\gamma := \frac{n^3}{(1-\alpha)N^3};$$

(2) *When* $\alpha = 1$, *choosing* $\eta_s = \frac{1}{(mS)^{\frac{1}{3}}}$ *and assuming* $S \ge (\frac{416}{3}L + \frac{16}{3}L_H)^3 b^2 n^3/N$, *we have*

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\|\nabla F(\mathbf{x}_0^s)\|^2 \lesssim \frac{1}{T^{\frac{2}{3}}} + \gamma'\frac{m^3}{T},$$

where we define $\gamma' := \frac{n^3}{N^3}$.

We can apply a similar analysis as that of Theorem 1 to compare `CorgiPile` with vanilla SGD, in terms of convergence rate, and reach similar insights.

Supplemental Materials

A PRELIMINARIES

Before presenting our theoretical analysis, we first show some preliminary definitions and lemmas which is important to our proofs.

Lemma 1. *Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a μ-strongly convex function. Then $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, there is*

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

Lemma 2. *Suppose $f(\mathbf{x})$ is an L-smooth convex function. Then $\forall \mathbf{x}, \mathbf{x}^* \in \mathbb{R}^d$ where $\mathbf{x}^*$ is one global optimum of $f(\mathbf{x})$, there is*

$$\|\nabla f(\mathbf{x})\|_2^2 \leq 2L\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right)$$

Fact 1. *Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable. $\nabla f(\mathbf{x}) \in \mathbb{R}^d$ and $H(\mathbf{x}) \in \mathbb{R}^{d\times d}$ denote its derivative and Hessian at the point $\mathbf{x} \in \mathbb{R}^d$. Then we have, $\forall \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$,*

$$\nabla f(\mathbf{z}) - \nabla f(\mathbf{y}) = \int_0^{\|\mathbf{z}-\mathbf{y}\|} H\left(\mathbf{y} + \frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}t\right)\frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}\,\mathrm{d}t.$$

For simplification, we further define

$$\int_{\mathbf{y}}^{\mathbf{z}} H(\mathbf{x})\mathrm{d}\mathbf{x} := \int_0^{\|\mathbf{z}-\mathbf{y}\|} H\left(\mathbf{y} + \frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}t\right)\frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}\,\mathrm{d}t.$$

which thus lead to $\nabla f(\mathbf{z}) - \nabla f(\mathbf{y}) = \int_{\mathbf{y}}^{\mathbf{z}} H(\mathbf{x})\mathrm{d}\mathbf{x}$.

Lemma 3. *Suppose there are two non-negative sequences $\{a_s\}_{s=0}^{+\infty}$, $\{b_s\}_{s=0}^{+\infty}$ satisfying*

$$A_0\eta_s a_s \leq (1 - \mu A_1\eta_s)b_s - b_{s+1} + A_2\eta_s^2 + A_3\eta_s^3 + A_4\eta_s^4$$

where $\eta_s = \frac{3}{A_1\mu(s+a)}$ with $a \geq 1, \mu > 0, A_0 > 0, A_1 > 0, A_2 > 0, A_3 > 0, A_4 > 0$ are constants. Then we have, for $S \geq 0$,

$$\frac{\sum_{s=1}^S w_s a_s}{\sum_{s=1}^S w_s} \leq \frac{4A_1\mu a^4 b_1}{3A_0S^4} + \frac{12A_2}{A_0A_1\mu}\frac{\sum_{s=1}^S(s+a)^2}{S^4} + \frac{36A_3}{A_0A_1^2\mu^2}\frac{\sum_{s=1}^S(s+a)}{S^4} + \frac{108A_4}{A_0A_1^3\mu^3}\frac{1}{S^3}$$

where we let $w_s = (s+a)^3$.

Proof.

$$\frac{1-\mu A_1\eta_s}{\eta_s}w_s = \left(\frac{1}{\eta_s} - \mu A_1\right)w_s = \frac{A_1\mu(s+a-3)(t+a)^3}{3} \leq \frac{w_{s-1}}{\eta_{s-1}} = \frac{A_1\mu(t+a-1)^4}{3}$$

where the inequality can be easily verified with the condition $a \geq 1$.

Thus we have

$$\begin{aligned} A_0 w_s a_s \leq& \frac{1-\mu A_1\eta_s}{\eta_s}w_s b_s - \frac{w_s}{\eta_s}b_{s+1} + A_2 w_s\eta_s + A_3 w_s\eta_s^2 + A_4\eta_s^3 w_s \\ \leq& \frac{w_{s-1}}{\eta_{s-1}}b_s - \frac{w_s}{\eta_s}b_{s+1} + A_2 w_s\eta_s + A_3 w_s\eta_s^2 + A_4\eta_s^3 w_s \end{aligned}$$

Taking summation on both sides of the above inequality, we have

$$\begin{aligned} A_0\sum_{s=1}^S w_s a_s \leq& \frac{w_0}{\eta_0}b_1 + A_2\sum_{s=1}^S w_s\eta_s + A_3\sum_{s=1}^S w_s\eta_s^2 + A_4\eta_s^3 w_s \\ \leq& \frac{w_0}{\eta_0}b_1 + \frac{3A_2}{A_1\mu}\sum_{s=1}^S(s+a)^2 + \frac{9A_3}{A_1^2\mu^2}\sum_{s=1}^S(s+a) + \frac{27A_4}{A_1^3\mu^3}S \end{aligned}$$

On the other hand, we also see that $\sum_{s=1}^S w_s = \sum_{s=1}^S(s+a)^3 \geq \sum_{s=1}^S s^3 \geq \frac{S^4}{4}$. Dividing both sides by $\sum_{s=1}^S w_s$, we can obtain

$$\frac{\sum_{s=1}^S w_s a_s}{\sum_{s=1}^S w_s} \leq \frac{4A_1\mu a^4 b_1}{3A_0S^4} + \frac{12A_2}{A_0A_1\mu}\frac{\sum_{s=1}^S(s+a)^2}{S^4} + \frac{36A_3}{A_0A_1^2\mu^2}\frac{\sum_{s=1}^S(s+a)}{S^4} + \frac{108A_4}{A_0A_1^3\mu^3}\frac{1}{S^3}$$

□

The main structure of our proof is based on the work of HaoChen and Sra [2], which tries to theoretically analyze the full shuffle SGD. However, our proofs below are not a trivial extension of the existing work. The key ingredients in our proofs employing some new techniques are the improvement of estimating the upper bounds of $\mathcal{I}_1$ and $\mathcal{I}_4$ in the proof of Theorem 1 and the related parts in the proofs of other theorems.

B PROOFS FOR CORGIPILE

Recall that at the k-th iteration in the s-th epoch, our parameter updating rule can be formulated as follows,

$$\mathbf{x}_k^s = \mathbf{x}_{k-1}^s - \eta_s \nabla f_{\psi_s(k)} \left(\mathbf{x}_{k-1}^s\right), \tag{3}$$

where η_s is the learning rate for the s-th epoch.

If we recursively apply this updating rule (3), we have that, at the k-th iteration in the s-th epoch,

$$\mathbf{x}_k^s = \mathbf{x}_0^s - \eta_s \sum_{k'=1}^{k} \nabla f_{\psi_s(k')} \left(\mathbf{x}_{k'-1}^s\right), \tag{4}$$

After the updates of one epoch, i.e., after bn steps in the s-th epoch, applying (4), we have

$$\mathbf{x}_0^{s+1} = \mathbf{x}_0^s - \eta_s \sum_{k=1}^{bn} \nabla f_{\psi_s(k)} \left(\mathbf{x}_{k-1}^s\right), \tag{5}$$

where we use the fact that

$$\mathbf{x}_0^{s+1} = \mathbf{x}_{bn}^s.$$

B.1 Proof of Theorem 1

Based on the updating rules above, our proof starts from the following formulation,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2 \\
=&\mathbb{E}\|\mathbf{x}_0^s - \eta_s \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \mathbf{x}^*\|^2 \\
=&\mathbb{E}\|\mathbf{x}_0^s - \mathbf{x}^*\|^2 - 2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) \right\rangle + \eta_s^2 \mathbb{E}\left\| \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) \right\|^2 \\
\leq&\mathbb{E}\|\mathbf{x}_0^s - \mathbf{x}^*\|^2 - 2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)] \right\rangle \\
&- 2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) \right\rangle + 2\eta_s^2 \mathbb{E}\left\| \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)] \right\|^2 \\
&+ 2\eta_s^2 \mathbb{E}\left\| \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) \right\|^2 \\
=&\mathbb{E}\|\mathbf{x}_0^s - \mathbf{x}^*\|^2 \underbrace{-2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)] \right\rangle}_{\mathcal{I}_1} \\
&\underbrace{-2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) \right\rangle}_{\mathcal{I}_2} \underbrace{+ 2\eta_s^2 \mathbb{E}\left\| \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)] \right\|^2}_{\mathcal{I}_3} \\
&\underbrace{+ 2\eta_s^2 \mathbb{E}\left\| \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) - \mathbb{E} \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) \right\|^2}_{\mathcal{I}_4} \underbrace{+ 2\eta_s^2 \left\| \mathbb{E} \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) \right\|^2}_{\mathcal{I}_5}
\end{aligned} \tag{6}$$

where the last equality uses the fact that $\mathbb{E}\|X - \mathbb{E}[X]\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}[X]\|^2$.

To prove the upper bound of (6), we need to bound $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{I}_4$ and $\mathcal{I}_5$ respectively.

Bound of $\mathcal{I}_3$

For $\mathcal{I}_3$, we have that

$$\begin{aligned}
&\left\|\sum_{k=1}^{bn}[\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s)-\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\|^2\\
\leq& bn\sum_{k=1}^{bn}\left\|\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s)-\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2\\
\leq& bn\sum_{k=1}^{bn}L^2\left\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\right\|^2\\
\leq& bn\sum_{k=1}^{bn}L^2\left\|\eta_s\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\right\|^2\\
\leq& bn\sum_{k=1}^{bn}\eta_s^2L^2(k-1)\sum_{k'=1}^{k-1}\left\|\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\right\|^2\\
\leq& \eta_s^2L^2G^2bn\sum_{k=1}^{bn}(k-1)^2\leq\frac{1}{3}\eta_s^2L^2G^2(bn)^4
\end{aligned}$$

where the first and the fourth inequalities uses the fact that $\|\sum_{k=1}^{bn}\mathbf{a}_k\|^2\leq bn\sum_{k=1}^{bn}\|\mathbf{a}_k\|^2$, the second inequality holds due to the Lipschitz continuity of the gradient, the third inequality is due to $\mathbf{x}_{k-1}^s=\mathbf{x}_0^s-\eta_s\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)$ and the last inequality is due to $\sum_{i=1}^{n}i^2=\frac{n(n+1)(2n+1)}{6}$ and the boundedness of the gradient.

Therefore, we have

$$\mathcal{I}_3=2\eta_s^2\mathbb{E}\left\|\sum_{k=1}^{bn}[\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s)-\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\|^2\leq\frac{2}{3}\eta_s^4L^2G^2K^4 \tag{7}$$

Bounds of $\mathcal{I}_2$ and $\mathcal{I}_5$

For $\mathcal{I}_2$ and $\mathcal{I}_5$, the key is to know the form of $\mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)$.

$$\mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)=\mathbb{E}\sum_{B_l\in\mathcal{B}_s}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s),$$

where this equality holds since the random shuffling ψ_s does not affect the summation in the LHS formula.

Furthermore, we use indicator random variables to get the value of $\mathbb{E}\sum_{B_l\in\mathcal{B}_s}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)$. Let $\mathbb{I}_{B_l\in\mathcal{B}_s}$ be the random variable to indicate whether the block B_l is in $\mathcal{B}_s$ or not. Therefore, we have

$$\mathbb{I}_{B_l\in\mathcal{B}_s}=\begin{cases}1, & \text{if } B_l\in\mathcal{B}_s\\ 0, & \text{if } B_l\notin\mathcal{B}_s\end{cases},$$

and

$$\mathbb{P}(\mathbb{I}_{B_l\in\mathcal{B}_s}=1)=\mathbb{P}(B_l\in\mathcal{B}_s)=\frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}}=\frac{n}{N}$$

such that $\mathbb{E}[\mathbb{I}_{B_l\in\mathcal{B}_s}]=\frac{n}{N}$.

Thus, we can obtain

$$\mathbb{E}\sum_{B_l\in\mathcal{B}_s}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)=\mathbb{E}\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\left(\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right)=\frac{n}{N}\sum_{l=1}^{N}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)=\frac{n}{N}m\nabla F(\mathbf{x}_0^s).$$

Therefore, we get the values of $\mathcal{I}_2$ and $\mathcal{I}_5$

$$\mathcal{I}_2=-2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s-\mathbf{x}^*,\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\rangle=-2\eta_s\frac{n}{N}m\mathbb{E}\left\langle\mathbf{x}_0^s-\mathbf{x}^*,\nabla F(\mathbf{x}_0^s)\right\rangle \tag{8}$$

$$\mathcal{I}_5=2\eta_s^2\left\|\mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2=2\eta_s^2\frac{n^2}{N^2}m^2\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \tag{9}$$

Bound of $\mathcal{I}_4$

Next, we will show the variance of sampling the n blocks out of N without replacement. We still use the indicator variables defined above.

The upper bound of $\mathcal{I}_4$ determines the $\frac{1}{T}$ term and the $\frac{N-n}{N-1}$ factor existing in the convergence rate, which shows how the leading term $\frac{N-n}{N-1}\frac{1}{T}$ varying with the number of sampled blocks n.

Note that for any $l' \neq l''$, we have

$$\mathbb{P}(\mathbb{I}_{B_{l'}\in\mathcal{B}_s}=1,\mathbb{I}_{B_{l''}\in\mathcal{B}_s}=1)=\mathbb{P}(B_{l'}\in\mathcal{B}_s\wedge B_{l''}\in\mathcal{B}_s)=\frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}}=\frac{n(n-1)}{N(N-1)}$$

Therefore, $\mathbb{E}[\mathbb{I}_{B_{l'}\in\mathcal{B}_s}\cdot\mathbb{I}_{B_{l''}\in\mathcal{B}_s}]=\frac{n(n-1)}{N(N-1)}$.

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)-\mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2\\
=&\mathbb{E}\left\|\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\left(\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right)-\mathbb{E}\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\left(\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right)\right\|^2\\
=&\mathbb{E}\left\|\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\left(\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right)\right\|^2-\left\|\frac{n}{N}\sum_{l=1}^{N}\left(\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right)\right\|^2\\
=&\mathbb{E}\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\cdot\mathbb{I}_{B_l\in\mathcal{B}_s}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2+\mathbb{E}\sum_{l'\neq l''}\mathbb{I}_{B_{l'}\in\mathcal{B}_s}\cdot\mathbb{I}_{B_{l''}\in\mathcal{B}_s}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s),\sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle\\
&-\frac{n^2}{N^2}\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2-\frac{n^2}{N^2}\sum_{l'\neq l''}^{N}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s),\sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle\\
=&\mathbb{E}\sum_{l=1}^{N}\mathbb{I}_{B_l\in\mathcal{B}_s}\cdot\mathbb{I}_{B_l\in\mathcal{B}_s}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2+\mathbb{E}\sum_{l'\neq l''}\mathbb{I}_{B_{l'}\in\mathcal{B}_s}\cdot\mathbb{I}_{B_{l''}\in\mathcal{B}_s}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s),\sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle\\
&-\frac{n^2}{N^2}\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2-\frac{n^2}{N^2}\sum_{l'\neq l''}^{N}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s),\sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle\\
=&\left(\frac{n}{N}-\frac{n^2}{N^2}\right)\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2+\left(\frac{n(n-1)}{N(N-1)}-\frac{n^2}{N^2}\right)\sum_{l'\neq l''}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s),\sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle
\end{aligned}$$

where the last equality is due to $\mathbb{E}[\mathbb{I}_{B_l\in\mathcal{B}_s}\cdot\mathbb{I}_{B_l\in\mathcal{B}_s}]=\mathbb{E}[\mathbb{I}_{B_l\in\mathcal{B}_s}]=\frac{n}{N}$ and $\mathbb{E}[\mathbb{I}_{B_{l'}\in\mathcal{B}_s}\cdot\mathbb{I}_{B_{l''}\in\mathcal{B}_s}]=\frac{n(n-1)}{N(N-1)}$, and the second equality is due to $\mathbb{E}\|X-\mathbb{E}X\|^2=\mathbb{E}\|X\|^2-\|\mathbb{E}X\|^2$.

On the other hand, we have

$$\begin{aligned}
&\mathbb{E}_l\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)-b\nabla F(\mathbf{x}_0^s)\right\|^2\\
=&\mathbb{E}_l\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)-\mathbb{E}_l\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2\\
=&\mathbb{E}_l\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)-\frac{1}{N}\sum_{l=1}^{N}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2\\
=&\mathbb{E}_l\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2-\frac{1}{N^2}\left\|\sum_{l=1}^{N}\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2
\end{aligned}$$

$$= \frac{1}{N}\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2 - \frac{1}{N^2}\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2 - \frac{1}{N^2}\sum_{l'\neq l''}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s), \sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle$$

$$= \left(\frac{1}{N} - \frac{1}{N^2}\right)\sum_{l=1}^{N}\left\|\sum_{i\in B_l}\nabla f_i(\mathbf{x}_0^s)\right\|^2 - \frac{1}{N^2}\sum_{l'\neq l''}\left\langle\sum_{i\in B_{l'}}\nabla f_i(\mathbf{x}_0^s), \sum_{i\in B_{l''}}\nabla f_i(\mathbf{x}_0^s)\right\rangle.$$

By comparing the RHS of the above two equations, we can observe that

$$\mathbb{E}\left\|\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s) - \mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2 = \frac{n(N-n)}{N-1}\mathbb{E}_{\xi'}\left\|\sum_{i\in B_{\xi'}}\nabla f_i(\mathbf{x}_0^s) - b\nabla F(\mathbf{x}_0^s)\right\|^2.$$

If we further apply our assumption that $\mathbb{E}_{\xi'}\left\|\frac{1}{b}\sum_{i\in B_{\xi'}}\nabla f_i(\mathbf{x}_0^s) - \nabla F(\mathbf{x}_0^s)\right\|^2 \leq h_D\frac{\sigma^2}{b}$, then there is

$$\mathcal{I}_4 = 2\eta_s^2\mathbb{E}\left\|\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s) - \mathbb{E}\sum_{k=1}^{bn}\nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2 \leq 2\eta_s^2\frac{nb(N-n)}{N-1}h_D\sigma^2 \tag{10}$$

This result shows the connection between the variance of block-wise sampling without replacement and the variance of sampling single data point independently and uniformly.

Bound of $\mathcal{I}_1$

The upper bound of $\mathcal{I}_1$ is critical to the proof of obtaining a faster rate. Before presenting the upper bound of $\mathcal{I}_1$, recall the Fact 1 that

$$\nabla f(\mathbf{z}) - \nabla f(\mathbf{y}) = \int_{\mathbf{y}}^{\mathbf{z}} H(\mathbf{x})\mathrm{d}\mathbf{x} := \int_0^{\|\mathbf{z}-\mathbf{y}\|} H\left(\mathbf{y} + \frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}t\right)\frac{\mathbf{z}-\mathbf{y}}{\|\mathbf{z}-\mathbf{y}\|}\,\mathrm{d}t.$$

Now we can show the upper bound of $\mathcal{I}_1$ as follows,

$$\begin{aligned}
\mathcal{I}_1 =& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}[\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\rangle \\
=& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} H_{\psi_s(k)}(\mathbf{x})\mathrm{d}\mathbf{x}\right\rangle \\
=& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} H_{\psi_s(k)}(\mathbf{x}^*)\mathrm{d}\mathbf{x}\right\rangle \\
& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s}\left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}^*)\right)\mathrm{d}\mathbf{x}\right\rangle \\
=& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}^*)\left(\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\right)\right\rangle \\
& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s}\left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}^*)\right)\mathrm{d}\mathbf{x}\right\rangle \\
=& 2\eta_s^2\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\right)\right\rangle \\
& -2\eta_s\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s}\left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}^*)\right)\mathrm{d}\mathbf{x}\right\rangle \\
=& \underbrace{2\eta_s^2\mathbb{E}\left\langle\mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\rangle}_{\mathcal{I}_{11}}
\end{aligned}$$

$$\underbrace{2\eta_s^2\mathbb{E}\left\langle \mathbf{x}_0^s-\mathbf{x}^*, \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)-\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\rangle}_{\mathcal{I}_{12}}$$

$$\underbrace{-2\eta_s\mathbb{E}\left\langle \mathbf{x}_0^s-\mathbf{x}^*, \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s}\left(H_{\psi_s(k)}(\mathbf{x})-H_{\psi_s(k)}(\mathbf{x}^*)\right)\mathrm{d}\mathbf{x}\right\rangle}_{\mathcal{I}_{13}}$$

where the second equality holds due to the Fact 1 and the fifth equality holds since $\mathbf{x}_{k-1}^s=\mathbf{x}_0^s-\eta_s\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)$. Note that $H_{\psi_s(k)}(\mathbf{x}^*)$ is the Hessian of the function $f_{\psi_s(k)}(\mathbf{x})$ at the point $\mathbf{x}^*$.

In order to obtain the upper bound of $\mathcal{I}_1$, we need to bound $\mathcal{I}_{11}$, $\mathcal{I}_{12}$ and $\mathcal{I}_{13}$ separately.

Bound of $\mathcal{I}_{11}$

In $\mathcal{I}_{11}$, we need first compute $\mathbb{E}\sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)$, which is the key ingredient for obtaining the $O(\frac{1}{T^2})$ term in the convergence rate.

To make our proof more clear, recall the manipulation in our algorithm for obtain ψ_s:

(1) At the s-th epoch, random sample n out of the total N blocks *without replacement* to get a set of sampled blocks $\mathcal{B}_s$ with size of n. Each block has b data samples.
(2) Then, perform random shuffling of the nb data samples and obtain the shuffled index sequence ψ_s with $|\psi_s|=nb$.

where we define $\mathcal{B}_s$ being the set of blocks that are sampled each epoch.

In order to compute the expectation, we use the indicator random variable for a more clear derivation.

Define $\mathbb{I}_{\psi_s(k)=i}$ be the indicator random variable showing whether one data sample with index i located in the k-th place after the above 2-step manipulation. The event $\psi_s(k)=i$ is equivalent to the event that $i\in B_l\wedge B_l\in\mathcal{B}_s\wedge\psi_s(k)=i$ where B_l is the block that the i-th sample lies in.

Thus, we have

$$\mathbb{I}_{\psi_s(k)=i}=\begin{cases}1, & \text{if } \psi_s(k)=i\\ 0, & \text{if } \psi_s(k)\neq i\end{cases}$$

and

$$\mathbb{P}(\mathbb{I}_{\psi_s(k)=i}=1)=\mathbb{P}(i\in B_l, B_l\in\mathcal{B}_s, \psi_s(k)=i)=\frac{\binom{1}{1}\binom{N-1}{n-1}(bn-1)!}{\binom{N}{n}(bn)!}=\frac{1}{Nb}.$$

Thus, we can observe that

$$H_{\psi_s(k)}(\mathbf{x}^*)=\sum_{i=1}^{m}\mathbb{I}_{\psi_s(k)=i}H_i(\mathbf{x}^*)$$

$$\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)=\sum_{j=1}^{m}\mathbb{I}_{\psi_s(k')=j}\nabla f_j(\mathbf{x}_0^s)$$

Based on the above definition, we are ready to compute the expectation.

$$\begin{aligned}
&\mathbb{E}\sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\\
=&\mathbb{E}\sum_{k=1}^{bn}\sum_{i=1}^{m}\mathbb{I}_{\psi_s(k)=i}H_i(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\sum_{j=1}^{m}\mathbb{I}_{\psi_s(k')=j}\nabla f_j(\mathbf{x}_0^s)\right)\\
=&\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=j}\right]H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\\
=&\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}\sum_{i\neq j}\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=j}\right]H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)
\end{aligned}$$

where the last equality holds since

$$\mathbb{P}(\mathbb{I}_{\psi_s(k)=i}=1,\mathbb{I}_{\psi_s(k')=i}=1)=0\quad\Rightarrow\quad\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=i}\right]=0$$

because $k > k'$ and one data sample cannot appear in different positions at the same time.

Therefore, for any $k > k'$ and $i \neq j$,

(1) If $i \in B_l, j \in B_l$, we have

$$\mathbb{P}(\mathbb{I}_{\psi_s(k)=i} = 1, \mathbb{I}_{\psi_s(k')=j} = 1) = \frac{\binom{1}{1}\binom{N-1}{n-1}(nb-2)!}{\binom{N}{n}(nb)!} = \frac{1}{Nb(nb-1)}$$

$$\Rightarrow$$

$$\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i} \cdot \mathbb{I}_{\psi_s(k')=j}\right] = \frac{1}{Nb(nb-1)}.$$

where $\binom{1}{1}$ means B_l are chosen ahead, $\binom{N-1}{n-1}$ means $(n-1)$ blocks excluding B_l are randomly chosen from $N-1$ blocks excluding B_l, $\binom{N}{n}$ is the total number of ways of choosing n blocks from N blocks, $(nb-2)!$ is the number of ways of shuffling the data in n blocks expect i and j, and $(nb)!$ is the number of ways of shuffling all the data in n blocks.

(2) If $i \in B_l, j \in B_l$ and $l \neq l'$, we have

$$\mathbb{P}(\mathbb{I}_{\psi_s(k)=i} = 1, \mathbb{I}_{\psi_s(k')=j} = 1) = \frac{\binom{2}{2}\binom{N-2}{n-2}(nb-2)!}{\binom{N}{n}(nb)!} = \frac{n-1}{Nb(N-1)(nb-1)}.$$

$$\Rightarrow$$

$$\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i} \cdot \mathbb{I}_{\psi_s(k')=j}\right] = \frac{n-1}{Nb(N-1)(nb-1)}.$$

where $\binom{2}{2}$ means B_l and $B_{l'}$ are chosen ahead, $\binom{N-2}{n-2}$ means $(n-2)$ blocks excluding B_l and $B_{l'}$ are randomly chosen from $N-2$ blocks excluding B_l and $B_{l'}$, $\binom{N}{n}$ is the total number of ways of choosing n blocks from N blocks, $(nb-2)!$ is the number of ways of shuffling the data in n blocks expect i and j, and $(nb)!$ is the number of ways of shuffling all the data in n blocks.

Thus, we have

$$\begin{aligned}
&\mathbb{E}\sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\\
=&\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}\sum_{i\neq j}\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=j}\right]H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\\
=&\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}\left(\sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=j}\right]H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\right.\\
&\left.+\sum_{l\neq l'}\sum_{\substack{i,j\\ i\in B_l, j\in B_{l'}}}\mathbb{E}\left[\mathbb{I}_{\psi_s(k)=i}\cdot\mathbb{I}_{\psi_s(k')=j}\right]H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\right)\\
=&\frac{nb(nb-1)}{2}\left(\sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}\frac{1}{Nb(nb-1)}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)+\sum_{l\neq l'}\sum_{\substack{i,j\\ i\in B_l, j\in B_{l'}}}\frac{n-1}{Nb(N-1)(nb-1)}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\right)\\
=&\sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}\frac{n}{2N}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)+\sum_{l\neq l'}\sum_{\substack{i,j\\ i\in B_l, j\in B_{l'}}}\frac{n(n-1)}{2N(N-1)}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)
\end{aligned}$$

Plugging in the above result into $\mathcal{I}_{11}$, we can get

$$\begin{aligned}
\mathcal{I}_{11} =&2\eta_s^2\mathbb{E}\left\langle \mathbf{x}_0^s-\mathbf{x}^*, \sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}\frac{n}{2N}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)+\sum_{l\neq l'}\sum_{\substack{i,j\\ i\in B_l, j\in B_{l'}}}\frac{n(n-1)}{2N(N-1)}H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}_0^s)\right\rangle\\
=&2\eta_s^2\mathbb{E}\left\langle \mathbf{x}_0^s-\mathbf{x}^*, \sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}\frac{n}{2N}H_i(\mathbf{x}^*)\left(\nabla f_j(\mathbf{x}_0^s)-\nabla f_j(\mathbf{x}^*)\right)\right\rangle
\end{aligned}$$

$$+ 2\eta_s^2 \left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \frac{n(n-1)}{2N(N-1)} H_i(\mathbf{x}^*)\left(\nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\right)\right\rangle$$

$$+ 2\eta_s^2 \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \frac{n}{2N} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*) + \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \frac{n(n-1)}{2N(N-1)} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*)\right\rangle$$

$$= \underbrace{\eta_s^2 \frac{n}{N} \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \left\langle H_i(\mathbf{x}^*)(\mathbf{x}_0^s - \mathbf{x}^*), \nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\right\rangle}_{\mathcal{J}_1}$$

$$\underbrace{+ \eta_s^2 \frac{n(n-1)}{N(N-1)} \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \left\langle H_i(\mathbf{x}^*)(\mathbf{x}_0^s - \mathbf{x}^*), \nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\right\rangle}_{\mathcal{J}_2}$$

$$\underbrace{+ \eta_s^2 \frac{n}{N} \left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*) + \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*)\right\rangle}_{\mathcal{J}_3}$$

To bound $\mathcal{I}_{11}$, we need bound the terms $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ separately.

Bound of $\mathcal{J}_1 + \mathcal{J}_2$

$$\begin{aligned}
\mathcal{J}_1 + \mathcal{J}_2 =& \eta_s^2 \frac{n}{N} \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \left\langle H_i(\mathbf{x}^*)(\mathbf{x}_0^s - \mathbf{x}^*), \nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\right\rangle \\
&+ \eta_s^2 \frac{n(n-1)}{N(N-1)} \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \left\langle H_i(\mathbf{x}^*)(\mathbf{x}_0^s - \mathbf{x}^*), \nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\right\rangle \\
\leq& \eta_s^2 \frac{n}{N} \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \|H_i(\mathbf{x}^*)\| \|\mathbf{x}_0^s - \mathbf{x}^*\| \|\nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\| \\
&+ \eta_s^2 \frac{n(n-1)}{N(N-1)} \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \|H_i(\mathbf{x}^*)\| \|\mathbf{x}_0^s - \mathbf{x}^*\| \|\nabla f_j(\mathbf{x}_0^s) - \nabla f_j(\mathbf{x}^*)\| \\
\leq& \eta_s^2 \frac{n}{N} L^2 \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + \eta_s^2 \frac{n(n-1)}{N(N-1)} L^2 \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 \\
\leq& \eta_s^2 \frac{n}{N} L^2 N b(b-1) \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + \eta_s^2 \frac{n(n-1)}{N(N-1)} L^2 N(N-1) b^2 \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 \\
\leq& \eta_s^2 L^2 nb(nb-1) \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 \\
\leq& \eta_s^2 L^2 n^2 b^2 \|\mathbf{x}_0^s - \mathbf{x}^*\|^2
\end{aligned}$$

Bound of $\mathcal{J}_3$

$$\mathcal{J}_3 = 2\eta_s^2 \left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{l=1}^{N} \sum_{\substack{i\neq j \\ i,j\in B_l}} \frac{n}{2N} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*) + \sum_{l\neq l'} \sum_{\substack{i,j \\ i\in B_l, j\in B_{l'}}} \frac{n(n-1)}{2N(N-1)} H_i(\mathbf{x}^*)\nabla f_j(\mathbf{x}^*)\right\rangle$$

$$\leq \eta_s^2 \frac{n}{N} \|\mathbf{x}_0^s - \mathbf{x}^*\| \left\| \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) \right\|$$

Note the fact that

$$\left(\sum_{l=1}^{N} \sum_{i \in B_l} H_i(\mathbf{x}^*) \right) \left(\sum_{l'=1}^{N} \sum_{j \in B_{l'}} \nabla f_j(\mathbf{x}^*) \right) = 0$$

since $\sum_{l'=1}^{N} \sum_{j \in B_{l'}} \nabla f_j(\mathbf{x}^*) = Nb \nabla f(\mathbf{x}^*) = 0$.

Therefore, letting $\rho \geq 0$, we can obtain a tighter bound for the following term,

$$\begin{aligned}
&\left\| \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) \right\| \\
=&\left\| \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) \right. \\
&\left. - \rho \left(\sum_{l=1}^{N} \sum_{i \in B_l} H_i(\mathbf{x}^*) \right) \left(\sum_{l'=1}^{N} \sum_{j \in B_{l'}} \nabla f_j(\mathbf{x}^*) \right) \right\| \\
=&\left\| (1-\rho) \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \left(\frac{n-1}{N-1} - \rho \right) H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) - \rho \sum_{l=1}^{N} \sum_{i \in B_l} H_i(\mathbf{x}^*) \nabla f_i(\mathbf{x}^*) \right\| \\
\leq& |1-\rho| Nb(b-1)LG + N(N-1)b^2 \left| \frac{n-1}{N-1} - \rho \right| + \rho NbLG
\end{aligned} \tag{11}$$

To find a tight upper bound of (11), we need to discuss its value as follows:

(1) When $\rho \geq 1$, the RHS of the above inequality is increasing with respect to ρ.
(2) When $\rho \leq \frac{n-1}{N-1}$, the RHS of the above inequality is decreasing with respect to ρ.
(3) When $\frac{n-1}{N-1} \leq \rho \leq 1$ and $N \geq 2$, the RHS of the above inequality is increasing with respect to ρ.

Thus $\rho = \frac{n-1}{N-1}$ is the minimizer of (11). Plugging the value of ρ into (11), we can obtain the upper bound as

$$|1-\rho| Nb(b-1)LG + N(N-1)b^2 \left| \frac{n-1}{N-1} - \rho \right| + \rho NbLG \leq \frac{N-n}{N-1} Nb(b-1)LG + \frac{n-1}{N-1} NbLG$$

which leads to

$$\begin{aligned}
\mathcal{J}_3 \leq& \eta_s^2 \frac{n}{N} \|\mathbf{x}_0^s - \mathbf{x}^*\| \left\| \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}^*) \nabla f_j(\mathbf{x}^*) \right\| \\
\leq& \eta_s^2 \|\mathbf{x}_0^s - \mathbf{x}^*\| \frac{N-n}{N-1} nb(b-1)LG + \eta_s^2 \|\mathbf{x}_0^s - \mathbf{x}^*\| \frac{n-1}{N-1} nbLG \\
\leq& \frac{1}{8} \eta_s \mu nb \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + 2\eta_s^3 nb \left(\frac{N-n}{N-1} \right)^2 (b-1)^2 L^2 G^2 \mu^{-1} \\
&+ \frac{1}{8} \eta_s \mu nb \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + 2\eta_s^3 nb \left(\frac{n-1}{N-1} \right)^2 L^2 G^2 \mu^{-1} \\
=& \frac{1}{4} \eta_s \mu nb \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + 2\eta_s^3 nbL^2 G^2 \mu^{-1} \left[\left(\frac{N-n}{N-1} \right)^2 (b-1)^2 + \left(\frac{n-1}{N-1} \right)^2 \right]
\end{aligned}$$

Therefore, we have

$$\mathcal{I}_{11} \leq \frac{1}{4} \eta_s \mu nb \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + \eta_s^2 L^2 n^2 b^2 \|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + 2\eta_s^3 nbL^2 G^2 \mu^{-1} \left[\left(\frac{N-n}{N-1} \right)^2 (b-1)^2 + \left(\frac{n-1}{N-1} \right)^2 \right]$$

Bound of $\mathcal{I}_{12}$

$$
\begin{aligned}
\mathcal{I}_{12} =& 2\eta_s^2 \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) - \nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\rangle \\
\le& 2\eta_s^2 \mathbb{E}\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \left\|\sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}^*)\left(\sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) - \nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\| \\
\le& 2\eta_s^2 \mathbb{E}\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \sum_{k=1}^{bn}\sum_{k'=1}^{k-1} \left\|H_{\psi_s(k)}(\mathbf{x}^*)\left(\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) - \nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\| \\
\le& 2\eta_s^2 \mathbb{E}\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \sum_{k=1}^{bn}\sum_{k'=1}^{k-1} L \left\|\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) - \nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right\| \\
\le& 2\eta_s^2 \mathbb{E}\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \sum_{k=1}^{bn}\sum_{k'=1}^{k-1} L^2 \left\|\mathbf{x}_{k'-1}^s - \mathbf{x}_0^s\right\| \\
=& 2\eta_s^2 \mathbb{E}\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \sum_{k=1}^{bn}\sum_{k'=1}^{k-1} L^2 \left\|\eta_s \sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\right\| \\
\le& 2\eta_s^3 L^2 G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \sum_{k=1}^{bn} (k-1)^2 \\
\le& \frac{2}{3}(bn)^3 \eta_s^3 L^2 G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \\
=& \frac{1}{3}\eta_s^2 b^2 n^2 L G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{1}{3}\eta_s^4 b^4 n^4 L^3 G
\end{aligned}
$$

where the first inequality is due to $\langle \mathbf{a}, \mathbf{b}\rangle \le \|\mathbf{a}\|\|\mathbf{b}\|$, the second inequality is due to $\|\sum_{k=1}^{bn} \mathbf{a}_k\| \le \sum_{k=1}^{bn} \|\mathbf{a}\|$, the third and fourth inequalities are because of Lipschitz gradient assumption, the second equality is because of $\mathbf{x}_{k-1}^s = \mathbf{x}_0^s - \eta_s \sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)$, and the last inequality holds since $ab \le \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$.

Bound of $\mathcal{I}_{13}$

$$
\begin{aligned}
\mathcal{I}_{13} =& -2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} \left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}^*)\right) \mathrm{d}\mathbf{x}\right\rangle \\
=& -2\eta_s \mathbb{E}\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \sum_{k=1}^{bn} \int_0^{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} \left(H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} t\right) - H_{\psi_s(k)}(\mathbf{x}^*)\right) \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} \mathrm{d}t\right\rangle \\
=& -2\eta_s \mathbb{E} \sum_{k=1}^{bn} \int_0^{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} \left\langle \mathbf{x}_0^s - \mathbf{x}^*, \left(H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} t\right) - H_{\psi_s(k)}(\mathbf{x}^*)\right) \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|}\right\rangle \mathrm{d}t \\
\le& 2\eta_s \mathbb{E} \sum_{k=1}^{bn} \int_0^{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \cdot \left\|H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} t\right) - H_{\psi_s(k)}(\mathbf{x}^*)\right\| \frac{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} \mathrm{d}t \\
\le& 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \cdot \mathbb{E} \sum_{k=1}^{bn} \int_0^{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} L_H \left\|\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s - \mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} t - \mathbf{x}^*\right\| \mathrm{d}t \\
\le& 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \cdot \mathbb{E} \sum_{k=1}^{bn} \int_0^{\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|} L_H \left(\|\mathbf{x}_0^s - \mathbf{x}^*\| + t\right) \mathrm{d}t \\
=& 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| L_H \cdot \mathbb{E} \sum_{k=1}^{bn} \left(\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|\|\mathbf{x}_0^s - \mathbf{x}^*\| + \frac{1}{2}\|\mathbf{x}_{k-1}^s - \mathbf{x}_0^s\|^2\right) \\
=& 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| L_H \cdot \mathbb{E} \sum_{k=1}^{bn} \left(\|\eta_s \sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\|\|\mathbf{x}_0^s - \mathbf{x}^*\| + \frac{1}{2}\|\eta_s \sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\|^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| L_H \cdot \sum_{k=1}^{bn} \left(\eta_s(k-1)G\|\mathbf{x}_0^s - \mathbf{x}^*\| + \frac{1}{2}\eta_s^2(k-1)^2G^2\right) \\
&\leq 2\eta_s \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| L_H \left(\frac{(bn)^2}{2}\eta_s G\|\mathbf{x}_0^s - \mathbf{x}^*\| + \frac{(bn)^3}{6}\eta_s^2G^2\right) \\
&= \eta_s^2(bn)^2 L_H G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{(bn)^3}{3}\eta_s^3 L_H G^2 \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\| \\
&\leq \eta_s^2(bn)^2 L_H G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{(bn)^2}{6}\eta_s^2 L_H G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{(bn)^4}{6}\eta_s^4 L_H G^3 \\
&= \frac{7}{6}\eta_s^2 b^2 n^2 L_H G \left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{1}{6}\eta_s^4 b^4 n^4 L_H G^3
\end{aligned}
$$

where the second equality holds since the Fact 1, the third inequality is due to $\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$, the fourth inequality is due to the boundedness of the gradient, the fifth equality is due to $\mathbf{x}_{k-1}^s = \mathbf{x}_0^s - \eta_s \sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)$, and the last inequality is because of $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$.

Based on the upper bounds of $\mathcal{I}_{11}$, $\mathcal{I}_{12}$ and $\mathcal{I}_{13}$, we can obtain the bound of $\mathcal{I}_1$ as

$$
\begin{aligned}
\mathcal{I}_1 =& \mathcal{I}_{11} + \mathcal{I}_{12} + \mathcal{I}_{13} \\
\leq& \frac{1}{4}\eta_s\mu nb\|\mathbf{x}_0^s - \mathbf{x}^*\|^2 + 2\eta_s^3 nbL^2G^2\mu^{-1}\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] \\
&+ \frac{1}{6}\eta_s^2 b^2 n^2 (2LG + 6L^2 + 7L_H G)\left\|\mathbf{x}_0^s - \mathbf{x}^*\right\|^2 + \frac{1}{6}\eta_s^4 b^4 n^4 (2L^3 G + L_H G^3)
\end{aligned} \tag{12}
$$

Now we summarize the above upper bounds of $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{I}_4$, $\mathcal{I}_5$. Plugging (12),(8), (7), (10), (9) into (6), we can eventually obtain that

$$
\begin{aligned}
\mathbb{E}\|x_0^{s+1} - x^*\|^2 \leq& (1 + \frac{1}{4}\eta_s\mu bn + C_1\eta_s^2 b^2 n^2)\|x_0^s - x^*\|^2 - 2\eta_s bn\left\langle \mathbf{x}_0^s - \mathbf{x}^*, \nabla F(\mathbf{x}_0^s)\right\rangle + 2\eta_s^2 b^2 n^2 \left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \\
&+ C_2\eta_s^2 nb\frac{N-n}{N-1}h_D\sigma^2 + C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4\eta_s^4 b^4 n^4
\end{aligned} \tag{13}
$$

where we let

$$
C_1 = \frac{1}{3}LG + L^2 + \frac{7}{6}L_H G, \quad C_2 = 2, \quad C_3 = 2L^2G^2\mu^{-1}, \quad C_4 = \frac{2}{3}L^2G^4 + \frac{1}{3}L^3G + \frac{1}{6}L_H G^3
$$

By the definition of μ-strongly convex function $F(\cdot)$, we have

$$
F(\mathbf{x}^*) - F(\mathbf{x}_0^s) \geq \left\langle \mathbf{x}^* - \mathbf{x}_0^s, \nabla F(\mathbf{x}_0^s)\right\rangle + \frac{\mu}{2}\|\mathbf{x}_0^s - \mathbf{x}^*\|^2. \tag{14}
$$

By the definition of L-smooth convex function $F(\cdot)$, we have

$$
\|\nabla F(\mathbf{x}_0^s)\|^2 \leq 2L(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)). \tag{15}
$$

Plugging (14) and (15) into (13), we have

$$
\begin{aligned}
(2\eta_s bn - 4L\eta_s^2 b^2 n^2)(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) \leq& (1 - \frac{3}{4}\eta_s bn\mu + C_1\eta_s^2 b^2 n^2)\|x_0^s - x^*\|^2 - \mathbb{E}\|x_0^{s+1} - x^*\|^2 \\
&+ C_2\eta_s^2 nb\frac{N-n}{N-1}h_D\sigma^2 + C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4\eta_s^4 b^4 n^4
\end{aligned}
$$

Now assume that $\eta_s \leq \min\{\frac{1}{4C_1\mu^{-1}bn}, \frac{1}{4Lbn}\}$, we eventually obtain

$$
\begin{aligned}
\eta_s bn(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) \leq& (1 - \frac{1}{2}\eta_s bn\mu)\|x_0^s - x^*\|^2 - \mathbb{E}\|x_0^{s+1} - x^*\|^2 \\
&+ C_2\eta_s^2 nb\frac{N-n}{N-1}h_D\sigma^2 + C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4\eta_s^4 b^4 n^4
\end{aligned}
$$

By Lemma 3, letting $S \geq 1$, $\eta_s = \frac{6}{bn\mu(s+a)}$, and $T = bSn$, we can have

$$
F\left(\frac{\sum_{s=1}^S w_s \mathbf{x}_0^s}{\sum_{s=1}^S w_s}\right) - F(\mathbf{x}^*) \leq \frac{\sum_{s=1}^S w_s(F(\mathbf{x}_0^s) - F(\mathbf{x}^*))}{\sum_{s=1}^S w_s} \lesssim (1-\alpha)\frac{h_D\sigma^2}{T} + \beta\frac{1}{T^2} + \gamma\frac{m^3}{T^3}
$$

with

$$
\alpha := \frac{n-1}{N-1}, \beta := \alpha^2 + (1-\alpha)^2(b-1)^2, \gamma := \frac{n^3}{N^3}.
$$

and requiring

$$T \geq bn$$
$$a \geq \max\left\{\frac{8LG + 24L^2 + 28L_H G}{\mu^2}, \frac{24L}{\mu}, 1\right\}.$$

B.2 Proof of Theorem 2

Recall the PL condition for a certain constant μ is in the following form,

$$2\mu(f(\mathbf{x}) - f(\mathbf{x}^*)) \leq \|\nabla f(\mathbf{x})\|^2$$

We begin our proof as follows, by the L-smoothness of the objective function $f(\mathbf{x})$,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_0^{s+1}) - f(\mathbf{x}_0^s)] &\leq \left\langle \mathbb{E}\mathbf{x}_0^{s+1} - \mathbf{x}_0^s, \nabla f(\mathbf{x}_0^s)\right\rangle + \frac{L}{2}\mathbb{E}\|\mathbf{x}_0^{s+1} - \mathbf{x}_0^s\|^2 \\
&\leq -\eta_s \mathbb{E}\left\langle \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s), \nabla f(\mathbf{x}_0^s)\right\rangle + \frac{L}{2}\eta_s^2 \mathbb{E}\left\|\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s)\right\|^2 \\
&\leq \underbrace{-\eta_s \mathbb{E}\left\langle \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)], \nabla f(\mathbf{x}_0^s)\right\rangle}_{\mathcal{G}_1} \\
&\quad \underbrace{-\eta_s \mathbb{E}\left\langle \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s), \nabla f(\mathbf{x}_0^s)\right\rangle}_{\mathcal{G}_2} + \underbrace{L\eta_s^2 \mathbb{E}\left\|\sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\|^2}_{\mathcal{G}_3} \\
&\quad \underbrace{+ L\eta_s^2 \mathbb{E}\left\|\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) - \mathbb{E}\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2}_{\mathcal{G}_4} + \underbrace{L\eta_s^2 \mathbb{E}\left\|\mathbb{E}\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2}_{\mathcal{G}_5}
\end{aligned} \tag{16}$$

Bounds of $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$ and $\mathcal{G}_5$

As shown in the proof of Theorem 1, $\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{I}_4$ and $\mathcal{I}_5$ are the similar terms to $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$ and $\mathcal{G}_5$. We can shown their upper bounds as follows,

$$\mathcal{G}_2 = -\eta_s \mathbb{E}\left\langle \sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s), \nabla f(\mathbf{x}_0^s)\right\rangle = -\eta_s \frac{n}{N} m \left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \tag{17}$$

$$\mathcal{G}_3 = L\eta_s^2 \mathbb{E}\left\|\sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\|^2 \leq \frac{1}{3}\eta_s^4 L^3 G^2 (bn)^4 \tag{18}$$

$$\mathcal{G}_4 = L\eta_s^2 \mathbb{E}\left\|\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s) - \mathbb{E}\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2 \leq L\eta_s^2 \frac{nb(N-n)}{N-1} h_D \sigma^2 \tag{19}$$

$$\mathcal{G}_5 = L\eta_s^2 \left\|\mathbb{E}\sum_{k=1}^{bn} \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)\right\|^2 = \eta_s^2 \frac{n^2}{N^2} m^2 L \left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \tag{20}$$

Bound of $\mathcal{G}_1$

Next, we will show the upper bound of $\mathcal{G}_1$.

$$\begin{aligned}
\mathcal{G}_1 &= -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} [\nabla f_{\psi_s(k)}(\mathbf{x}_{k-1}^s) - \nabla f_{\psi_s(k)}(\mathbf{x}_0^s)]\right\rangle \\
&= -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} H_{\psi_s(k)}(\mathbf{x})\mathrm{d}\mathbf{x}\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&= -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} H_{\psi_s(k)}(\mathbf{x}_0^s) \mathrm{d}\mathbf{x} \right\rangle \\
&\quad -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} \left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}_0^s) \right) \mathrm{d}\mathbf{x} \right\rangle \\
&= -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}_0^s) \left(\mathbf{x}_{k-1}^s - \mathbf{x}_0^s \right) \right\rangle \\
&\quad -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} \left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}_0^s) \right) \mathrm{d}\mathbf{x} \right\rangle \\
&= \eta_s^2 \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}_0^s) \left(\sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) \right) \right\rangle \\
&\quad -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} \left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}_0^s) \right) \mathrm{d}\mathbf{x} \right\rangle \\
&= \underbrace{\eta_s^2 \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}_0^s) \left(\sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_0^s) \right) \right\rangle}_{\mathcal{G}_{11}} \\
&\quad \underbrace{\eta_s^2 \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} H_{\psi_s(k)}(\mathbf{x}_0^s) \left(\sum_{k'=1}^{k-1} \nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s) - \nabla f_{\psi_s(k')}(\mathbf{x}_0^s) \right) \right\rangle}_{\mathcal{G}_{12}} \\
&\quad \underbrace{-\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn} \int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s} \left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}_0^s) \right) \mathrm{d}\mathbf{x} \right\rangle}_{\mathcal{G}_{13}}
\end{aligned}
$$

Bound of $\mathcal{G}_{11}$

As shown in the proof of $\mathcal{J}_3$, we can similarly have

$$
\begin{aligned}
\mathcal{G}_{11} =& \eta_s^2 \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} \frac{n}{2N} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n(n-1)}{2N(N-1)} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle \\
=& \eta_s^2 \frac{n}{2N} \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle \\
=& \eta_s^2 \frac{n}{2N} \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) + \sum_{l \neq l'} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle \\
& - \eta_s^2 \frac{n}{2N} \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{l'=1}^{N} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle \\
& + \eta_s^2 \frac{n}{2N} \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{l'=1}^{N} \sum_{\substack{i,j \\ i \in B_l, j \in B_{l'}}} \frac{n-1}{N-1} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle \\
=& \eta_s^2 \frac{n}{2N} \left\langle \nabla f(\mathbf{x}_0^s), \sum_{l=1}^{N} \sum_{\substack{i \neq j \\ i,j \in B_l}} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) + \sum_{l=1}^{N} \sum_{i,j \in B_l} \frac{n-1}{N-1} H_i(\mathbf{x}_0^s) \nabla f_j(\mathbf{x}_0^s) \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&+\eta_s^2\frac{n}{2N}\left\langle\nabla f(\mathbf{x}_0^s),\frac{n-1}{N-1}N^2b^2H(\mathbf{x}_0^s)\nabla f(\mathbf{x}_0^s)\right\rangle\\
\le&\eta_s^2\frac{n}{2N}\left\|\nabla f(\mathbf{x}_0^s)\right\|\left\|\sum_{l=1}^{N}\sum_{\substack{i\neq j\\ i,j\in B_l}}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)+\sum_{l=1}^{N}\sum_{i,j\in B_l}\frac{n-1}{N-1}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)\right\|\\
&+\eta_s^2\frac{n(n-1)Nb^2}{2(N-1)}L\|\nabla f(\mathbf{x}_0^s)\|^2\\
\le&\eta_s^2\|\nabla f(\mathbf{x}_0^s)\|\frac{N-n}{N-1}nb(b-1)LG+\eta_s^2\|\nabla f(\mathbf{x}_0^s)\|\frac{n-1}{N-1}nbLG+\eta_s^2\frac{n(n-1)Nb^2}{2(N-1)}L\|\nabla f(\mathbf{x}_0^s)\|^2\\
\le&\frac{1}{4}\eta_s nb\|\nabla f(\mathbf{x}_0^s)\|^2+\eta_s^3nb\left(\frac{N-n}{N-1}\right)^2(b-1)^2L^2G^2\\
&+\frac{1}{4}\eta_s nb\|\nabla f(\mathbf{x}_0^s)\|^2+\eta_s^3nb\left(\frac{n-1}{N-1}\right)^2L^2G^2+\eta_s^2\frac{n(n-1)Nb^2}{2(N-1)}L\|\nabla f(\mathbf{x}_0^s)\|^2\\
=&\frac{1}{2}\eta_s nb\|\nabla f(\mathbf{x}_0^s)\|^2+\eta_s^3nbL^2G^2\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2+\left(\frac{n-1}{N-1}\right)^2\right]\\
&+\eta_s^2\frac{n(n-1)Nb^2}{2(N-1)}L\|\nabla f(\mathbf{x}_0^s)\|^2
\end{aligned}
\tag{21}
$$

where the fourth equality is due to

$$
\sum_{l=1}^{N}\sum_{l'=1}^{N}\sum_{\substack{i,j\\ i\in B_l,j\in B_{l'}}}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)=N^2b^2H(\mathbf{x}_0^s)\nabla f(\mathbf{x}_0^s),
$$

$$
\sum_{l\neq l'}\sum_{\substack{i,j\\ i\in B_l,j\in B_{l'}}}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)-\sum_{l=1}^{N}\sum_{l'=1}^{N}\sum_{\substack{i,j\\ i\in B_l,j\in B_{l'}}}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)=\sum_{l=1}^{N}\sum_{i,j\in B_l}H_i(\mathbf{x}_0^s)\nabla f_j(\mathbf{x}_0^s)
$$

with $H(\mathbf{x}_0^s)$ being the Hessian of $f(\mathbf{x})$ at the point $\mathbf{x}_0^s$.

Bound of $\mathcal{G}_{12}$

$$
\begin{aligned}
\mathcal{G}_{12}=&\eta_s^2\mathbb{E}\left\langle\nabla f(\mathbf{x}_0^s),\sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}_0^s)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)-\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\rangle\\
\le&\eta_s^2\mathbb{E}\left\|\nabla f(\mathbf{x}_0^s)\right\|\left\|\sum_{k=1}^{bn}H_{\psi_s(k)}(\mathbf{x}_0^s)\left(\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)-\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\|\\
\le&\eta_s^2\mathbb{E}\left\|\nabla f(\mathbf{x}_0^s)\right\|\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}\left\|H_{\psi_s(k)}(\mathbf{x}_0^s)\left(\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)-\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right)\right\|\\
\le&\eta_s^2\mathbb{E}\left\|\nabla f(\mathbf{x}_0^s)\right\|\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}L\left\|\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)-\nabla f_{\psi_s(k')}(\mathbf{x}_0^s)\right\|\\
\le&\eta_s^2\mathbb{E}\left\|\nabla f(\mathbf{x}_0^s)\right\|\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}L^2\left\|\mathbf{x}_{k'-1}^s-\mathbf{x}_0^s\right\|\\
=&\eta_s^2\mathbb{E}\left\|\nabla f(\mathbf{x}_0^s)\right\|\sum_{k=1}^{bn}\sum_{k'=1}^{k-1}L^2\left\|\eta_s\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\right\|\\
\le&\eta_s^3L^2G\left\|\nabla f(\mathbf{x}_0^s)\right\|\sum_{k=1}^{bn}(k-1)^2\\
\le&\frac{1}{3}(bn)^3\eta_s^3L^2G\left\|\nabla f(\mathbf{x}_0^s)\right\|
\end{aligned}
$$

$$\leq \frac{1}{6}\eta_s^2(bn)^2 L\left\|\nabla f(\mathbf{x}_0^s)\right\|^2 + \frac{1}{6}\eta_s^4(bn)^4 L^3 G^2 \tag{22}$$

Bound of $\mathcal{G}_{13}$

$$\begin{aligned}
\mathcal{G}_{13} =& -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn}\int_{\mathbf{x}_0^s}^{\mathbf{x}_{k-1}^s}\left(H_{\psi_s(k)}(\mathbf{x}) - H_{\psi_s(k)}(\mathbf{x}_0^s)\right)\mathrm{d}\mathbf{x}\right\rangle \\
=& -\eta_s \mathbb{E}\left\langle \nabla f(\mathbf{x}_0^s), \sum_{k=1}^{bn}\int_0^{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\left(H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}t\right) - H_{\psi_s(k)}(\mathbf{x}_0^s)\right)\frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\mathrm{d}t\right\rangle \\
=& -\eta_s \mathbb{E}\sum_{k=1}^{bn}\int_0^{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\left\langle \nabla f(\mathbf{x}_0^s), \left(H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}t\right) - H_{\psi_s(k)}(\mathbf{x}_0^s)\right)\frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\right\rangle \mathrm{d}t \\
\leq& \eta_s \mathbb{E}\sum_{k=1}^{bn}\int_0^{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\left\|\nabla f(\mathbf{x}_0^s)\right\|\cdot\left\|H_{\psi_s(k)}\left(\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}t\right) - H_{\psi_s(k)}(\mathbf{x}_0^s)\right\|\frac{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}\mathrm{d}t \\
\leq& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\|\cdot\mathbb{E}\sum_{k=1}^{bn}\int_0^{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|} L_H\left\|\mathbf{x}_0^s + \frac{\mathbf{x}_{k-1}^s-\mathbf{x}_0^s}{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|}t - \mathbf{x}_0^s\right\|\mathrm{d}t \\
\leq& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\|\cdot\mathbb{E}\sum_{k=1}^{bn}\int_0^{\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|} L_H t\mathrm{d}t \\
=& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\| L_H\cdot\mathbb{E}\sum_{k=1}^{bn}\frac{1}{2}\|\mathbf{x}_{k-1}^s-\mathbf{x}_0^s\|^2 \\
=& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\| L_H\cdot\mathbb{E}\sum_{k=1}^{bn}\frac{1}{2}\eta_s\|\sum_{k'=1}^{k-1}\nabla f_{\psi_s(k')}(\mathbf{x}_{k'-1}^s)\|^2 \\
\leq& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\| L_H\cdot\sum_{k=1}^{bn}\frac{1}{2}\eta_s^2(k-1)^2G^2 \\
\leq& \eta_s\left\|\nabla f(\mathbf{x}_0^s)\right\| L_H\frac{(bn)^3}{6}\eta_s^2G^2 \\
\leq& \frac{(bn)^2}{12}\eta_s^2 L_H\left\|\nabla f(\mathbf{x}_0^s)\right\|^2 + \frac{(bn)^4}{12}\eta_s^4 L_H G^4
\end{aligned} \tag{23}$$

Based on (21), (22), (23), we can have

$$\begin{aligned}
\mathcal{G}_1 =& \mathcal{G}_{11} + \mathcal{G}_{12} + \mathcal{G}_{13} \\
\leq& \frac{1}{2}\eta_s nb\|\nabla f(\mathbf{x}_0^s)\|^2 + \eta_s^3 nbL^2G^2\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] \\
&+ \eta_s^2\frac{n(n-1)Nb^2}{2(N-1)}L\|\nabla f(\mathbf{x}_0^s)\|^2 + \frac{1}{6}\eta_s^2b^2n^2L\left\|\nabla f(\mathbf{x}_0^s)\right\|^2 + \frac{1}{6}\eta_s^4b^4n^4L^3G^2 \\
&+ \frac{(bn)^2}{12}\eta_s^2L_H\left\|\nabla f(\mathbf{x}_0^s)\right\|^2 + \frac{(bn)^4}{12}\eta_s^4L_HG^4
\end{aligned} \tag{24}$$

Plugging (24) (17) (18) (19) into (16), we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_0^{s+1}) - f(\mathbf{x}_0^s) \leq& -\frac{1}{2}\eta_s bn\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_1\eta_s^2b^2n^2\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_2\eta_s^2bn\frac{N-n}{N-1}h_D\sigma^2 \\
&+ C_3\eta_s^3bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4b^4n^4\eta_s^4
\end{aligned} \tag{25}$$

where we let

$$C_1 = \frac{13}{6}L + \frac{1}{12}L_H, \quad C_2 = L, \quad C_3 = L^2G^2, \quad C_4 = \frac{1}{2}L^3G^2 + \frac{1}{12}L_HG^4$$

By the definition of PL condition, we have have

$$2\mu(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) \leq \|\nabla F(\mathbf{x}_0^s)\|^2$$

Plugging this into the above formulation, we have

$$\begin{aligned}
&\left(\frac{1}{4}\eta_s bn - C_1\eta_s^2 b^2 n^2\right)\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \\
\leq& F(\mathbf{x}_0^s) - \mathbb{E}F(\mathbf{x}_0^{s+1}) - \frac{1}{4}\eta_s bn\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_2\eta_s^2 bn\frac{N-n}{N-1}h_D\sigma^2 \\
&+ C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4 b^4 n^4\eta_s^4 \\
\leq& F(\mathbf{x}_0^s) - \mathbb{E}F(\mathbf{x}_0^{s+1}) - \frac{1}{2}\eta_s bn\mu(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) + C_1\eta_s^2 b^2 n^2\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_2\eta_s^2 bn\frac{N-n}{N-1}h_D\sigma^2 \\
&+ C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4 b^4 n^4\eta_s^4 \\
=&(1 - \frac{1}{2}\eta_s bn\mu)(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) - (\mathbb{E}F(\mathbf{x}_0^{s+1}) - F(\mathbf{x}^*)) + C_2\eta_s^3 bn + C_3\eta_s^4 b^4 n^4
\end{aligned}$$

Now assume that $\eta_s \leq \frac{1}{8C_1 bn}$, we eventually obtain

$$\begin{aligned}
\frac{1}{8}\eta_s bn\|\nabla F(\mathbf{x}_0^s)\|^2 \leq&(1 - \frac{1}{2}\eta_s bn\mu)(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) - (\mathbb{E}F(\mathbf{x}_0^{s+1}) - F(\mathbf{x}^*)) + C_2\eta_s^2 bn\frac{N-n}{N-1}h_D\sigma^2 \\
&+ C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4 b^4 n^4\eta_s^4
\end{aligned}$$

By Lemma 3, letting $S \geq 1$, $\eta_s = \frac{6}{bn\mu(s+a)}$, and $T = bSn$, we can have

$$\sum_{s=1}^{S}\frac{w_s\|\nabla F(\mathbf{x}_0^s)\|^2}{\sum_{s=1}^{S} w_s} \lesssim (1-\alpha)\frac{h_D\sigma^2}{T} + \beta\frac{1}{T^2} + \gamma\frac{m^3}{T^3}$$

with

$$\alpha := \frac{n-1}{N-1}, \beta := \alpha^2 + (1-\alpha)^2(b-1)^2, \gamma := \frac{n^3}{N^3}.$$

and requiring

$$\begin{aligned}
&T \geq bn \\
&a \geq \max\left\{\frac{108L + 4L_H}{\mu}, 1\right\}.
\end{aligned}$$

The above result further lead to

$$F\left(\frac{\sum_{s=1}^{S} w_s\mathbf{x}_0^s}{\sum_{s=1}^{S} w_s}\right) - F(\mathbf{x}^*) \lesssim (1-\alpha)\frac{h_D\sigma^2}{T} + \beta\frac{1}{T^2} + \gamma\frac{m^3}{T^3}$$

by PL condition.

For the proof of non-convex objectives without PL condition, we can directly use the formulation (25),

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_0^{s+1}) - f(\mathbf{x}_0^s) \leq& -\frac{1}{2}\eta_s bn\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_1\eta_s^2 b^2 n^2\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 + C_2\eta_s^2 bn\frac{N-n}{N-1}h_D\sigma^2 \\
&+ C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2 + \left(\frac{n-1}{N-1}\right)^2\right] + C_4 b^4 n^4\eta_s^4
\end{aligned} \tag{26}$$

where we let

$$C_1 = \frac{13}{6}L + \frac{1}{12}L_H, \quad C_2 = L, \quad C_3 = L^2G^2, \quad C_4 = \frac{1}{2}L^3G^2 + \frac{1}{12}L_H G^4$$

Now assuming $\eta_s \leq \frac{1}{4C_1 bn}$

$$\frac{1}{4}\eta_s bn\left\|\nabla F(\mathbf{x}_0^s)\right\|^2 \leq f(\mathbf{x}_0^s) - \mathbb{E}f(\mathbf{x}_0^{s+1}) + C_2\eta_s^2 bn\frac{N-n}{N-1}h_D\sigma^2$$

$$+ C_3\eta_s^3 bn\left[\left(\frac{N-n}{N-1}\right)^2(b-1)^2+\left(\frac{n-1}{N-1}\right)^2\right]+C_4b^4n^4\eta_s^4$$

Taking summation from $s=1$ to S and dividing both side by S, and then we set the step size as follows to obtain different convergence rate.

(1) When $\alpha \le \frac{N-2}{N-1}$, choosing $\eta_s = \frac{1}{\sqrt{bn(1-\alpha)h_D\sigma^2 S}}$ and assuming $S \ge \frac{16bn(\frac{13}{6}L+\frac{1}{12}L_H)^2}{\sigma^2(1-\alpha)h_D}$, we have,

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\|\nabla F(\mathbf{x}_0^s)\|^2 \lesssim (1-\alpha)^{1/2}\frac{\sqrt{h_D\sigma^2}}{\sqrt{T}}+\beta\frac{1}{T}+\gamma\frac{m^3}{T^{\frac{3}{2}}},$$

where the factors are defined as follows

$$\alpha := \frac{n-1}{N-1}, \beta := \frac{\alpha^2}{1-\alpha}\frac{1}{h_D\sigma^2}+(1-\alpha)\frac{(b-1)^2}{h_D\sigma^2}, \gamma := \frac{n^3}{(1-\alpha)N^3}.$$

(2) When $\alpha = 1$, choosing $\eta_s = \frac{1}{(mS)^{\frac{1}{3}}}$ and assuming $S \ge 64(\frac{13}{6}L+\frac{1}{12}L_H)^3b^2n^3/N$, we have,

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\|\nabla F(\mathbf{x}_0^s)\|^2 \lesssim \frac{1}{T^{\frac{2}{3}}}+\gamma'\frac{m^3}{T},$$

where we define

$$\gamma' := \frac{n^3}{N^3}.$$

REFERENCES

[1] Mert Gürbüzbalaban, Asuman E. Ozdaglar, and Pablo A. Parrilo. 2021. Why random reshuffling beats stochastic gradient descent. *Math. Program.* 186, 1 (2021), 49–84.
[2] Jeff Z. HaoChen and Suvrit Sra. 2019. Random Shuffling Beats SGD after Finite Epochs. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 2624–2633.
[3] Ohad Shamir. 2016. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*. 46–54.
[4] Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H Sayed. 2019. Stochastic Learning Under Random Reshuffling With Constant Step-Sizes. *IEEE Transactions on Signal Processing* 67, 2 (2019), 474–489.