

---

# A Computation and Communication Efficient Method for Distributed Nonconvex Problems in the Partial Participation Setting

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We present a new method that includes three key components of distributed opti-  
2 mization and federated learning: variance reduction of stochastic gradients, partial  
3 participation, and compressed communication. We prove that the new method has  
4 optimal oracle complexity and state-of-the-art communication complexity in the  
5 partial participation setting. Regardless of the communication compression feature,  
6 our method successfully combines variance reduction and partial participation: we  
7 get the optimal oracle complexity, never need the participation of all nodes, and do  
8 not require the bounded gradients (dissimilarity) assumption.

## 9 1 Introduction

10 Federated and distributed learning have become very popular in recent years (Konečný et al., 2016;  
11 McMahan et al., 2017). The current optimization tasks require much computational resources and  
12 machines. Such requirements emerge in machine learning, where massive datasets and computations  
13 are distributed between cluster nodes (Lin et al., 2017; Ramesh et al., 2021). In federated learning,  
14 nodes, represented by mobile phones, laptops, and desktops, do not send their data to a server due to  
15 privacy and their huge number (Ramaswamy et al., 2019), and the server remotely orchestrates the  
16 nodes and communicates with them to solve an optimization problem.

17 As in classical optimization tasks, one of the main current challenges is to find **computationally**  
18 **efficient** optimization algorithms. However, the nature of distributed problems induces many other  
19 (Kairouz et al., 2021), including i) **partial participation** of nodes in algorithm steps: due to stragglers  
20 (Li et al., 2020) or communication delays (Vogels et al., 2021), ii) **communication bottleneck**: even  
21 if a node participates, it can be costly to transmit information to a server or other nodes (Alistarh  
22 et al., 2017; Ramesh et al., 2021; Kairouz et al., 2021; Sapio et al., 2019; Narayanan et al., 2019). It  
23 is necessary to develop a method that considers these problems.

## 24 2 Optimization Problem

25 Let us consider the nonconvex distributed optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}, \quad (1)$$

26 where  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$  is a smooth nonconvex function for all  $i \in [n] := \{1, \dots, n\}$ . The full  
27 information about function  $f_i$  is stored on  $i^{\text{th}}$  node. The communication between nodes is maintained  
28 in the parameters server fashion (Kairouz et al., 2021): we have a server that receives compressed

information from nodes, updates a state, and broadcasts an updated model.<sup>1</sup> Since we work in the nonconvex world, our goal is to find an  $\varepsilon$ -solution ( $\varepsilon$ -stationary point) of (1): a (possibly random) point  $\hat{x} \in \mathbb{R}^d$ , such that  $\mathbb{E}[\|\nabla f(\hat{x})\|^2] \leq \varepsilon$ .

We consider three settings:

1. **Gradient Setting.** The  $i^{\text{th}}$  node has only access to the gradient  $\nabla f_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$  of function  $f_i$ . Moreover, the following assumptions for the functions  $f_i$  hold.

**Assumption 1.** *There exists  $f^* \in \mathbb{R}$  such that  $f(x) \geq f^*$  for all  $x \in \mathbb{R}$ .*

**Assumption 2.** *The function  $f$  is  $L$ -smooth, i.e.,  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$  for all  $x, y \in \mathbb{R}^d$ .*

**Assumption 3.** *The functions  $f_i$  are  $L_i$ -smooth for all  $i \in [n]$ . Let us define  $\hat{L}^2 := \frac{1}{n} \sum_{i=1}^n L_i^2$ .<sup>2</sup>*

2. **Finite-Sum Setting.** The functions  $\{f_i\}_{i=1}^n$  have the finite-sum form

$$f_i(x) = \frac{1}{m} \sum_{j=1}^m f_{ij}(x), \quad \forall i \in [n], \quad (2)$$

where  $f_{ij} : \mathbb{R}^d \rightarrow \mathbb{R}$  is a smooth nonconvex function for all  $j \in [m]$ . We assume that Assumptions 1, 2 and 3 hold and the following assumption.

**Assumption 4.** *The function  $f_{ij}$  is  $L_{ij}$ -smooth for all  $i \in [n], j \in [m]$ . Let  $L_{\max} := \max_{i \in [n], j \in [m]} L_{ij}$ .*

3. **Stochastic Setting.** The function  $f_i$  is an expectation of a stochastic function,

$$f_i(x) = \mathbb{E}_{\xi} [f_i(x; \xi)], \quad \forall i \in [n], \quad (3)$$

where  $f_i : \mathbb{R}^d \times \Omega_{\xi} \rightarrow \mathbb{R}$ . For a fixed  $x \in \mathbb{R}$ ,  $f_i(x; \xi)$  is a random variable over some distribution  $\mathcal{D}_i$ , and, for a fixed  $\xi \in \Omega_{\xi}$ ,  $f_i(x; \xi)$  is a smooth nonconvex function. The  $i^{\text{th}}$  node has only access to a stochastic gradients  $\nabla f_i(\cdot; \xi_{ij})$  of the function  $f_i$  through the distribution  $\mathcal{D}_i$ , where  $\xi_{ij}$  is a sample from  $\mathcal{D}_i$ . We assume that Assumptions 1, 2 and 3 hold and the following assumptions.

**Assumption 5.** *For all  $i \in [n]$  and for all  $x \in \mathbb{R}^d$ , the stochastic gradient  $\nabla f_i(x; \xi)$  is unbiased and has bounded variance, i.e.,  $\mathbb{E}_{\xi} [\nabla f_i(x; \xi)] = \nabla f_i(x)$ , and  $\mathbb{E}_{\xi} [\|\nabla f_i(x; \xi) - \nabla f_i(x)\|^2] \leq \sigma^2$ , where  $\sigma^2 \geq 0$ .*

**Assumption 6.** *For all  $i \in [n]$  and for all  $x, y \in \mathbb{R}$ , the stochastic gradient  $\nabla f_i(x; \xi)$  satisfies the mean-squared smoothness property, i.e.,  $\mathbb{E}_{\xi} [\|\nabla f_i(x; \xi) - \nabla f_i(y; \xi)\|^2] \leq L_{\sigma}^2 \|x - y\|^2$ .*

We compare algorithms using the *oracle complexity*, i.e., the number of (stochastic) gradients that each node has to calculate to get  $\varepsilon$ -solution, and the *communication complexity*, i.e., the number of bits that each node has to send to the server to get  $\varepsilon$ -solution.

## 2.1 Unbiased Compressors

We use the concept of unbiased compressors to alleviate the communication bottleneck. The unbiased compressors quantize and/or sparsify vectors that the nodes send to the server.

**Definition 1.** A stochastic mapping  $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is an *unbiased compressor* if there exists  $\omega \in \mathbb{R}$  such that

$$\mathbb{E}[\mathcal{C}(x)] = x, \quad \mathbb{E}[\|\mathcal{C}(x) - x\|^2] \leq \omega \|x\|^2, \quad (4)$$

for all  $x \in \mathbb{R}^d$ .

<sup>1</sup>Note that this strategy can be used in peer-to-peer communication, assuming that the server is an abstraction and all its algorithmic steps are performed on each node.

<sup>2</sup>Note that  $L \leq \hat{L}$ ,  $\hat{L} \leq L_{\max}$ , and  $\hat{L} \leq L_{\sigma}$ .

Table 1: Summary of methods that solve the problem (1) in the stochastic setting (3). Abbr.: *VR* (Variance Reduction) = Does a method have the optimal oracle complexity  $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon} + \frac{\sigma}{\varepsilon^{3/2}}\right)$ ? *PP* (Partial Participation) = Does a method support partial participation from Section 2.2? *CC* = Does a method have the communication complexity equals to  $\mathcal{O}\left(\frac{d}{\sqrt{n\varepsilon}}\right)$ ?

Method	VR	PP	CC	Limitations
<b>SPIDER, SARAH, PAGE, STORM</b> (Fang et al., 2018; Nguyen et al., 2017) (Li et al., 2021a; Cutkosky and Orabona, 2019)	✓	✗	✗	—
<b>MARINA</b> (Gorbunov et al., 2021)	✓	✗ <sup>(a)</sup>	✓ <sup>(b)</sup>	Suboptimal convergence rate (see (Tyurin and Richtárik, 2023)).
<b>FedPAGE</b> (Zhao et al., 2021b)	✗	✗ <sup>(a)</sup>	✗	Suboptimal oracle complexity $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right)$ .
<b>FRECON</b> (Zhao et al., 2021a)	✗	✓	✓	—
<b>FedAvg</b> (McMahan et al., 2017; Karimireddy et al., 2020b)	✗	✓	✗	Bounded gradients (dissimilarity) assumption of $f_i$ .
<b>SCAFFOLD</b> (Karimireddy et al., 2020b)	✗	✓	✗	Suboptimal convergence rate <sup>(e)</sup> .
<b>MIME</b> <sup>(c)</sup> (Karimireddy et al., 2020a)	✗ <sup>(d)</sup>	✓	✗	Calculates full gradient. Bounded gradients (dissimilarity) assumption of $f_i$ . Suboptimal oracle complexity $\mathcal{O}(1/\varepsilon^{3/2})$ in the setting (2).
<b>CE-LSGD</b> (for Partial Participation) <sup>(c)</sup> (Patel et al., 2022) (concurrent work)	✓	✓	✗	Bounded gradients (dissimilarity) assumption of $f_i$ . Suboptimal oracle complexity $\mathcal{O}(1/\varepsilon^{3/2})$ in the setting (2).
<b>DASHA</b> (Tyurin and Richtárik, 2023)	✓ ✗	✗ or ✓	✓ ✓	—
<b>DASHA-PP</b> (new)	✓	✓	✓	—

<sup>(a)</sup> **MARINA** and **FedPAGE**, with a small probability, require the participation of all nodes so that they can not support partial participation from Section 2.2. Moreover, these methods provide suboptimal oracle complexities.

<sup>(b)</sup> On average, **MARINA** provides the compressed communication mechanism with complexity  $\mathcal{O}\left(\frac{d}{\sqrt{n\varepsilon}}\right)$ . However, with a small probability, this method sends non-compressed vectors.

<sup>(c)</sup> Note that **MIME** and **CE-LSGD** can not be directly compared with **DASHA-PP** because **MIME** and **CE-LSGD** consider the online version of the problem (1), and require more strict assumptions.

<sup>(d)</sup> Although **MIME** obtains the convergence rate  $\mathcal{O}\left(\frac{1}{\varepsilon^{3/2}}\right)$  of a variance reduced method, it requires the calculation of the full (exact) gradients.

<sup>(e)</sup> It can be seen when  $\sigma^2 = 0$ . Let us consider the  $s$ -nice sampling of the nodes, then **SCAFFOLD** requires  $\mathcal{O}\left(\frac{n^{3/2}}{\varepsilon s^{3/2}}\right)$  communication rounds to get  $\varepsilon$ -solution, while **DASHA-PP** requires  $\mathcal{O}\left(\frac{\sqrt{n}}{\varepsilon s}\right)$  communication rounds (see Theorem 4 with  $\omega = 0$ ,  $b = \frac{p_a}{2-p_a}$ , and  $p_a = \frac{s}{n}$ ).

We denote a set of stochastic mappings that satisfy Definition 1 as  $\mathbb{U}(\omega)$ . In our methods, the nodes make use of unbiased compressors  $\{\mathcal{C}_i\}_{i=1}^n$ . The community developed a large number of unbiased compressors, including *RandK* (see Definition 5) (Beznosikov et al., 2020; Stich et al., 2018), Adaptive sparsification (Wangni et al., 2018) and Natural compression and dithering (Horváth et al., 2019a). We are aware of correlated compressors by Szlendak et al. (2021) and quantizers by Suresh et al. (2022) that help in the homogeneous regimes, but in this work, we are mainly concentrated on generic heterogeneous regimes, though, for simplicity, assume the independence of the compressors.

**Assumption 7.**  $\mathcal{C}_i \in \mathbb{U}(\omega)$  for all  $i \in [n]$ , and the compressors are statistically independent.

Table 2: Summary of methods that solve the problem (1) in the finite-sum setting (2). Abbr.: *VR* (Variance Reduction) = Does a method have the optimal oracle complexity  $\mathcal{O}\left(m + \frac{\sqrt{m}}{\varepsilon}\right)$ ? *PP* and *CC* are defined in Table 1.

Method	VR	PP	CC	Limitations
<b>SPIDER, PAGE</b> (Fang et al., 2018; Li et al., 2021a)	✓	✗	✗	—
<b>MARINA</b> (Gorbunov et al., 2021)	✓	✗ <sup>(a)</sup>	✓ <sup>(b)</sup>	Suboptimal convergence rate (see (Tyurin and Richtárik, 2023)).
<b>ZeroSARAH</b> (Li et al., 2021b)	✓	✓	✗	Only homogeneous regime, i.e., the functions $f_i$ are equal.
<b>FedPAGE</b> (Zhao et al., 2021b)	✗	✗ <sup>(a)</sup>	✗	Suboptimal oracle complexity $\mathcal{O}\left(\frac{m}{\varepsilon}\right)$ .
<b>DASHA</b> (Tyurin and Richtárik, 2023)	✓	✗	✓	—
<b>DASHA-PP</b> (new)	✓	✓	✓	—

(a), (b) : see Table 1.

## 2.2 Nodes Partial Participation Assumptions

We now try to formalize the notion of partial participation. Let us assume that we have  $n$  events  $\{i^{\text{th}} \text{ node is participating}\}$  with the following properties.

**Assumption 8.** *The partial participation of nodes has the following distribution: exists constants  $p_a \in (0, 1]$  and  $p_{aa} \in [0, 1]$ , such that*

1.  $\text{Prob}(i^{\text{th}} \text{ node is participating}) = p_a, \quad \forall i \in [n],$
2.  $\text{Prob}(i^{\text{th}} \text{ and } j^{\text{th}} \text{ nodes are participating}) = p_{aa},$   
for all  $i \neq j \in [n].$
3.  $p_{aa} \leq p_a^2,$

and these events from different communication rounds are independent.

We are not fighting for the full generality and believe that more complex sampling strategies can be considered in the analysis. For simplicity, we settle upon Assumption 8. Standard partial participation strategies, including  $s$ -nice sampling, where the server chooses uniformly  $s$  nodes without replacement ( $p_a = s/n$  and  $p_{aa} = s(s-1)/n(n-1)$ ), and independent participation, where each node independently participates with probability  $p_a$  (due to independence, we have  $p_{aa} = p_a^2$ ), satisfy Assumption 8. In the literature,  $s$ -nice sampling is one of the most popular strategies (Zhao et al., 2021a; Richtárik et al., 2021; Reddi et al., 2020; Konečný et al., 2016).

## 3 Motivation and Related Work

The main goal of our paper is to develop a method for the nonconvex distributed optimization that will include three key features: variance reduction of stochastic gradients, compressed communication, and partial participation. We now provide an overview of the literature (see also Table 1 and Table 2).

### 1. Variance reduction of stochastic gradients

It is important to consider finite-sum (2) and stochastic (3) settings because, in machine learning tasks, either the number of local functions  $m$  is huge or the functions  $f_i$  is an expectation of a stochastic function due to the batch normalization (Ioffe and Szegedy, 2015) or random augmentation (Goodfellow et al., 2016), and it is infeasible to calculate the full gradients analytically. Let us recall the results from the nondistributed optimization. In the gradient setting, the optimal oracle complexity

is  $\mathcal{O}(1/\varepsilon)$ , achieved by the vanilla gradient descent (GD) (Carmon et al., 2020; Nesterov, 2018). In the finite-sum setting and stochastic settings, the optimal oracle complexities are  $\mathcal{O}\left(m + \frac{\sqrt{m}}{\varepsilon}\right)$  and  $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon} + \frac{\sigma}{\varepsilon^{3/2}}\right)$  (Fang et al., 2018; Li et al., 2021a; Arjevani et al., 2019), accordingly, achieved by methods SPIDER, SARAH, PAGE, and STORM from (Fang et al., 2018; Nguyen et al., 2017; Li et al., 2021a; Cutkosky and Orabona, 2019).

## 2. Compressed communication

In distributed optimization (Ramesh et al., 2021; Xu et al., 2021), lossy communication compression can be a powerful tool to increase the communication speed between the nodes and the server. Different types of compressors are considered in the literature, including unbiased compressors (Alistarh et al., 2017; Beznosikov et al., 2020; Szlendak et al., 2021), contractive (biased) compressors (Richtárik et al., 2021), 3PC compressors (Richtárik et al., 2022). We will focus on unbiased compressors because methods DASHA and MARINA (Tyurin and Richtárik, 2023; Szlendak et al., 2021; Gorbunov et al., 2021) that employ unbiased compressors provide the current theoretical state-of-the-art (SOTA) communication complexities.

Many methods analyzed optimization methods with the unbiased compressors (Alistarh et al., 2017; Mishchenko et al., 2019; Horváth et al., 2019b; Gorbunov et al., 2021; Tyurin and Richtárik, 2023). In the gradient setting, the methods MARINA and DASHA by Gorbunov et al. (2021) and Tyurin and Richtárik (2023) establish the current SOTA communication complexity, each method needs  $\frac{1+\omega/\sqrt{n}}{\varepsilon}$  communication rounds to get an  $\varepsilon$ -solution. In the finite-sum and stochastic settings, the current SOTA communication complexity is attained by the DASHA method, while maintaining the optimal oracle complexities  $\mathcal{O}\left(m + \frac{\sqrt{m}}{\varepsilon\sqrt{n}}\right)$  and  $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon n} + \frac{\sigma}{\varepsilon^{3/2}n}\right)$  per node.

## 3. Partial participation

From the beginning of federated learning era, the partial participation has been considered to be the essential feature of distributed optimization methods (McMahan et al., 2017; Konečný et al., 2016; Kairouz et al., 2021). However, previously proposed methods have limitations: i) methods MARINA and FedPAGE from (Gorbunov et al., 2021; Zhao et al., 2021b) still require synchronization of all nodes with a small probability. ii) in the stochastic settings, methods FedAvg, SCAFFOLD, and FRECON with the partial participation mechanism (McMahan et al., 2017; Karimireddy et al., 2020b; Zhao et al., 2021a) provide results without variance reduction techniques from (Fang et al., 2018; Li et al., 2021a; Cutkosky and Orabona, 2019) and, therefore, get suboptimal oracle complexities. Note that FRECON and DASHA reduce the variance *only from compressors* (in the partial participation and stochastic setting). iii) in the finite-sum setting, the ZeroSARAH method by Li et al. (2021b) focuses on the homogeneous regime only (the functions  $f_i$  are equal). iv) The MIME method by Karimireddy et al. (2020a) and the CE-LSGD method (for Partial Participation) by the concurrent paper (Patel et al., 2022) consider the online version of the problem (1). Therefore, MIME and CE-LSGD (for Partial Participation) require stricter assumptions, including the bounded inter-client gradient variance assumption. In the finite-sum setting (2), MIME and CE-LSGD obtain a suboptimal oracle complexity  $\mathcal{O}(1/\varepsilon^{3/2})$  while, in the full participation setting, it is possible to get the complexity  $\mathcal{O}(1/\varepsilon)$ .

## 4 Contributions

We propose a new method DASHA-PP for the nonconvex distributed optimization.

- As far as we know, this is the first method that includes three key ingredients of federated learning methods: *variance reduction of stochastic gradients, compressed communication, and partial participation*.
- Moreover, this is the first method that combines *variance reduction of stochastic gradients and partial participation* flawlessly: i) it gets the optimal oracle complexity ii) does not require the participation of all nodes iii) does not require the bounded gradients assumption of the functions  $f_i$ .
- We prove convergence rates and show that this method has *the optimal oracle complexity and the state-of-the-art communication complexity in the partial participation setting*. Moreover, in our work, we observe a nontrivial side-effect from mixing the variance reduction of stochastic gradients and partial participation. It is a general problem not related to our methods or analysis that we discuss in Section C.

---

**Algorithm 1 DASHA-PP**

---

- 1: **Input:** starting point  $x^0 \in \mathbb{R}^d$ , stepsize  $\gamma > 0$ , momentum  $a \in (0, 1]$ , momentum  $b \in (0, 1]$ , probability  $p_{\text{page}} \in (0, 1]$  (only in **DASHA-PP-PAGE**), batch size  $B$  (only in **DASHA-PP-PAGE**, **DASHA-PP-FINITE-MVR** and **DASHA-PP-MVR**), probability  $p_a \in (0, 1]$  that a node is *participating*<sup>(a)</sup>, number of iterations  $T \geq 1$
  - 2: Initialize  $g_i^0 \in \mathbb{R}^d$ ,  $h_i^0 \in \mathbb{R}^d$  on the nodes and  $g^0 = \frac{1}{n} \sum_{i=1}^n g_i^0$  on the server
  - 3: Initialize  $h_{ij}^0 \in \mathbb{R}^d$  on the nodes and take  $h_i^0 = \frac{1}{m} \sum_{j=1}^m h_{ij}^0$  (only in **DASHA-PP-FINITE-MVR**)
  - 4: **for**  $t = 0, 1, \dots, T - 1$  **do**
  - 5:    $x^{t+1} = x^t - \gamma g^t$
  - 6:   Broadcast  $x^{t+1}, x^t$  to all *participating*<sup>(a)</sup> nodes
  - 7:   **for**  $i = 1, \dots, n$  in parallel **do**
  - 8:     **if**  $i^{\text{th}}$  node is *participating*<sup>(a)</sup> **then**
  - 9:       Calculate  $k_i^{t+1}$  using Algorithm 2, 3, 4 or 5
  - 10:        $h_i^{t+1} = h_i^t + \frac{1}{p_a} k_i^{t+1}$
  - 11:        $m_i^{t+1} = C_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right)$
  - 12:        $g_i^{t+1} = g_i^t + m_i^{t+1}$
  - 13:       Send  $m_i^{t+1}$  to the server
  - 14:     **else**
  - 15:        $h_{ij}^{t+1} = h_{ij}^t$  (only in **DASHA-PP-FINITE-MVR**)
  - 16:        $h_i^{t+1} = h_i^t, \quad g_i^{t+1} = g_i^t, \quad m_i^{t+1} = 0$
  - 17:     **end if**
  - 18:   **end for**
  - 19:    $g^{t+1} = g^t + \frac{1}{n} \sum_{i=1}^n m_i^{t+1}$
  - 20: **end for**
  - 21: **Output:**  $\hat{x}^T$  chosen uniformly at random from  $\{x^t\}_{k=0}^{T-1}$
- (a): For the formal description see Section 2.2.
- 

---

**Algorithm 2** Calculate  $k_i^{t+1}$  for **DASHA-PP** in the gradient setting. See line 9 in Alg. 1

---

- 1:  $k_i^{t+1} = \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))$
- 

---

**Algorithm 3** Calculate  $k_i^{t+1}$  for **DASHA-PP-PAGE** in the finite-sum setting. See line 9 in Alg. 1

---

- 1: Generate a random set  $I_i^t$  of size  $B$  from  $[m]$  *with replacement*
  - 2:  $k_i^{t+1} = \begin{cases} \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)), \\ \text{with probability } p_{\text{page}} \text{ on all } \textit{participating} \text{ nodes,} \\ \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)), \\ \text{with probability } 1 - p_{\text{page}} \text{ on all } \textit{participating} \text{ nodes} \end{cases}$
- 

---

**Algorithm 4** Calc.  $k_i^{t+1}$  for **DASHA-PP-FINITE-MVR** in the finite-sum setting. See line 9 in Alg. 1

---

- 1: Generate a random set  $I_i^t$  of size  $B$  from  $[m]$  *without replacement*
  - 2:  $k_{ij}^{t+1} = \begin{cases} \frac{m}{B} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))), & j \in I_i^t, \\ 0, & j \notin I_i^t \end{cases}$
  - 3:  $h_{ij}^{t+1} = h_{ij}^t + \frac{1}{p_a} k_{ij}^{t+1}$
  - 4:  $k_i^{t+1} = \frac{1}{m} \sum_{j=1}^m k_{ij}^{t+1}$
- 

---

**Algorithm 5** Calculate  $k_i^{t+1}$  for **DASHA-PP-MVR** in the stochastic setting. See line 9 in Alg. 1

---

- 1: Generate i.i.d. samples  $\{\xi_{ij}^{t+1}\}_{j=1}^B$  of size  $B$  from  $\mathcal{D}_i$ .
  - 2:  $k_i^{t+1} = \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1}) - b \left( h_i^t - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1}) \right)$
-



## 5 Algorithm Description and Main Challenges Towards Partial Participation

We now present **DASHA-PP** (see Algorithm 1), a family of methods to solve the optimization problem (1). When we started investigating the problem, we took **DASHA** as a baseline method for two reasons: the family of algorithms **DASHA** provides the current state-of-the-art communication complexities in the *non-partial participation* setting, and, unlike **MARINA**, it does not send non-compressed gradients and does not synchronize all nodes. Let us briefly discuss the main idea of **DASHA**, its problem in the *partial participation* setting, and why the refinement of **DASHA** is not an exercise.

In fact, **DASHA** supports the partial participation of nodes *in the gradient setting*. Since the nodes only do the following steps (see full algorithm in Algorithm 6):

$$g_i^{t+1} = g_i^t + C_i (\nabla f_i(x^{t+1}) - (1-a)\nabla f_i(x^t) - ag_i^t).$$

The partial participation mechanism (independent participation from Section 2.2) can be easily implemented here if we redefine the compressor and use another one instead:

$$C_i^p := \begin{cases} \frac{1}{p}C_i, & \text{with pr. } p_a, \\ 0, & \text{with pr. } 1 - p_a. \end{cases} \Rightarrow g_i^{t+1} = \begin{cases} g_i^t + \frac{1}{p_a}C_i (\nabla f_i(x^{t+1}) - (1-a)\nabla f_i(x^t) - ag_i^t), & p_a \\ g_i^t, & 1 - p_a. \end{cases}$$

With probability  $1 - p$ , a node does not update  $g_i$  and does not send anything to the server. The main observation is that we can do this trick since  $g_i^{t+1}$  depends only on the vectors  $x^{t+1}$ ,  $x^t$ , and  $g_i^t$ .

However, we focus our attention on partial participation *in the finite-sum and stochastic settings*. Consider the nodes' steps in **DASHA-MVR** (see Algorithm 7) that is designed for the stochastic setting:

$$h_i^{t+1} = \nabla f_i(x^{t+1}; \xi_i^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})), \quad (6)$$

$$g_i^{t+1} = g_i^t + C_i (h_i^{t+1} - h_i^t - a(g_i^t - h_i^t)). \quad (7)$$

Even if we use the same trick for (7), we still have to update (6) in every iteration of the algorithm since  $g_i^{t+1}$  additionally depends on  $h_i^{t+1}$  and  $h_i^t$ . In other words, if a node does not update  $g_i$  and does not send anything to the server, it still has to update  $h_i$ , what is impossible without the points  $x^{t+1}$  and  $x^t$ . One of the main challenges was to “guess” how to generalize (6) and (7) to the partial participation setting. We now provide a solution:

$$h_i^{t+1} = h_i^t + \frac{1}{p_a}k_i^{t+1}, \quad k_i^{t+1} = \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})), \quad (8)$$

$$g_i^{t+1} = g_i^t + C_i \left( \frac{1}{p_a}k_i^{t+1} - \frac{a}{p_a}(g_i^t - h_i^t) \right) \text{ with pr. } p_a, \text{ and } h_i^{t+1} = h_i^t, \quad g_i^{t+1} = g_i^t \text{ with pr. } 1 - p_a.$$

Now both control variables  $g_i^t$  and  $h_i^t$  do not change with the probability  $p_a$ . However, this required changing the update rule of  $g_i^{t+1}$  and  $h_i^{t+1}$  to make the proof work.

The theoretical analysis of the new algorithm became more complicated: while in **DASHA**, the randomness from compressors is independent of the randomness from stochastic gradients (see (6) and (7)). In (8) (see also main Algorithm 1), these two randomnesses are coupled by the randomness from the partial participation. Going deeper into details, one can compare Lemma I.2 from (Tyurin and Richtárik, 2023) and Lemma 5. While the former lemma does not use the knowledge about the update rule of  $h_i^{t+1}$ , uses only (4), (16), and (17), the latter lemma additionally explicitly uses that  $h_i^{t+1} = h_i^t + \frac{1}{p_a}k_i^{t+1}$ , surgically copes with the expectations  $E_C[\cdot]$  and  $E_{p_a}[\cdot]$  (for instance, it is not trivial in each order one should apply the expectations), and uses the sampling lemma (Lemma 1). The same reasoning applies to other part of the analysis.

At the first reading of the proofs, we suggest the reader follow the proof of Theorem 2 in the gradient setting, which takes a small part of the paper. Although the appendix seems to be dense and large, the size is justified by the fact that we consider four different settings and PL-condition ( $4 \times 2$  tracks of the proofs. The theory is designed so that the proofs do not repeat steps of each other and use one framework).

## 6 Theorems

We now present the convergence rates theorems of **DASHA-PP** in different settings. We will compare the theorems with the results of the current state-of-the-art methods, **MARINA** and **DASHA**, that work

in the full participation setting. Suppose that **MARINA** or **DASHA** converges to  $\varepsilon$ -solution after  $T$  communication rounds. Then, ideally, we would expect the convergence of the new algorithms to  $\varepsilon$ -solution after up to  $T/p_a$  communication rounds due to the partial participation constraints<sup>3</sup>. The detailed analysis of the algorithms under Polyak-Łojasiewicz condition we provide in Section F. Let us define  $\Delta_0 := f(x^0) - f^*$ .

## 6.1 Gradient Setting

**Theorem 2.** Suppose that Assumptions 1, 2, 3, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_a}{2-p_a}$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{np_a^2} + \frac{16}{np_a^2} \left( 1 - \frac{p_{aa}}{p_a} \right) \right]^{1/2} \hat{L} \right)^{-1},$$

and  $g_i^0 = h_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (**DASHA-PP**), then  $\mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Delta_0}{\gamma T}$ .

Let us recall the convergence rate of **MARINA** or **DASHA**, the number of communication rounds to get  $\varepsilon$ -solution equals  $\mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{\sqrt{n}} \hat{L} \right] \right)$ , while the rate of **DASHA-PP** equals  $\mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega+1}{p_a \sqrt{n}} \hat{L} \right] \right)$ . Up to Lipschitz constants factors, we get the degeneration up to  $1/p_a$  factor due to the partial participation.

## 6.2 Finite-Sum Setting

**Theorem 3.** Suppose that Assumptions 1, 2, 3, 4, 7, and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_{page} p_a}{2-p_a}$ , probability  $p_{page} \in (0, 1]$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1-p_{page})L_{\max}^2}{B} \right) + \frac{16}{np_a^2 p_{page}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{page})L_{\max}^2}{B} \right) \right]^{1/2} \right)^{-1}$$

and  $g_i^0 = h_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (**DASHA-PP-PAGE**) then  $\mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Delta_0}{\gamma T}$ .

We now choose  $p_{page}$  to balance heavy full gradient and light mini-batch calculations. Let us define  $\mathbb{1}_{p_a} := \sqrt{1 - \frac{p_{aa}}{p_a}} \in [0, 1]$ . Note that if  $p_a = 1$  then  $p_{aa} = 1$  and  $\mathbb{1}_{p_a} = 0$ .

**Corollary 1.** Let the assumptions from Theorem 3 hold and  $p_{page} = B/(m+B)$ . Then **DASHA-PP-PAGE** needs

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{1}{p_a} \sqrt{\frac{m}{n}} \left( \frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_{\max}}{B} \right) \right] \right) \quad (9)$$

communication rounds to get an  $\varepsilon$ -solution and the expected number of gradient calculations per node equals  $\mathcal{O}(m + BT)$ .

The convergence rate of the current state-of-the-art method **DASHA-PAGE** without partial participation equals  $\mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{\sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \sqrt{\frac{m}{n}} \frac{L_{\max}}{B} \right] \right)$ . Let us closer compare it with (9). As expected, we see that the second term w.r.t.  $\omega$  degenerates up to  $1/p_a$ . Surprisingly, the third term w.r.t.  $\sqrt{m/n}$  can degenerate up to  $\sqrt{B}/p_a$  when  $\hat{L} \approx L_{\max}$ . Hence, in order to keep degeneration up to  $1/p_a$ , one should take the batch size  $B = \mathcal{O}(L_{\max}^2/\hat{L}^2)$ . This interesting effect we analyze separately in Section C. The fact that the degeneration is up to  $1/p_a$  we check numerically in Section A.

In the following corollary, we consider **RandK** compressors (see Definition 5) and show that with the particular choice of parameters, up to the Lipschitz constants factors, **DASHA-PP-PAGE** gets the optimal oracle complexity and SOTA communication complexity. The choice of the compressor is driven by simplicity, and the following analysis can be used for other unbiased compressors.

<sup>3</sup>We check this numerically in Section A.



223 **Corollary 2.** Suppose that assumptions of Corollary 1 hold,  $B \leq \min \left\{ \frac{1}{p_a} \sqrt{\frac{m}{n}}, \frac{L_{\max}^2}{\frac{1}{p_a} \hat{L}^2} \right\}^4$ , and we  
 224 use the unbiased compressor RandK with  $K = \Theta(Bd/\sqrt{m})$ . Then the communication complexity of  
 225 Algorithm 1 is

$$\mathcal{O} \left( d + \frac{L_{\max} \Delta_0 d}{p_a \varepsilon \sqrt{n}} \right), \quad (10)$$

226 and the expected number of gradient calculations per node equals

$$\mathcal{O} \left( m + \frac{L_{\max} \Delta_0 \sqrt{m}}{p_a \varepsilon \sqrt{n}} \right). \quad (11)$$

227 The convergence rate of DASHA-PP-FINITE-MVR is provided in Section E.5. The conclusions are the  
 228 same for the method.

### 229 6.3 Stochastic Setting

230 We define  $h^t := \frac{1}{n} \sum_{i=1}^n h_i^t$ .

231 **Theorem 4.** Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 232  $b \in \left( 0, \frac{p_a}{2-p_a} \right]$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B} \right) + \frac{12}{np_a b} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B} \right) \right]^{1/2} \right)^{-1},$$

233 and  $g_i^0 = h_i^0$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-MVR). Then

$$\begin{aligned} \mathbb{E} \left[ \left\| \nabla f(\bar{x}^T) \right\|^2 \right] &\leq \frac{1}{T} \left[ \frac{2\Delta_0}{\gamma} + \frac{2}{b} \left\| h^0 - \nabla f(x^0) \right\|^2 + \left( \frac{32b\omega(2\omega+1)}{np_a^2} + \frac{4 \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a} \right) \left( \frac{1}{n} \sum_{i=1}^n \left\| h_i^0 - \nabla f_i(x^0) \right\|^2 \right) \right] \\ &+ \left( \frac{48b^2\omega(2\omega+1)}{p_a^2} + \frac{12b}{p_a} \right) \frac{\sigma^2}{nB}. \end{aligned}$$

234 In the next corollary, we choose momentum  $b$  and initialize vectors  $h_i^0$  to get  $\varepsilon$ -solution. Let us define

$$235 \mathbb{1}_{p_a} := \sqrt{1 - \frac{p_{aa}}{p_a}} \in [0, 1].$$

236 **Corollary 3.** Suppose that assumptions from Theorem 4 hold, momentum  $b =$   
 237  $\Theta \left( \min \left\{ \frac{p_a}{\omega} \sqrt{\frac{n\varepsilon B}{\sigma^2}}, \frac{p_a n \varepsilon B}{\sigma^2} \right\} \right)$ ,  $\frac{\sigma^2}{n\varepsilon B} \geq 1$ , and  $h_i^0 = \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0)$  for all  $i \in [n]$ ,  
 238 and batch size  $B_{\text{init}} = \Theta \left( \frac{\sqrt{p_a B}}{b} \right)$ , then Algorithm 1 (DASHA-PP-MVR) needs

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon n}} \left( \frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right] + \frac{\sigma^2}{\sqrt{p_a n \varepsilon B}} \right)$$

239 communication rounds to get an  $\varepsilon$ -solution and the number of stochastic gradient calculations per  
 240 node equals  $\mathcal{O}(B_{\text{init}} + BT)$ .

241 The convergence rate of the DASHA-SYNC-MVR, the state-of-the-art method without partial participa-  
 242 tion, equals  $\mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{\sqrt{\varepsilon n}} \frac{L_\sigma}{B} \right] + \frac{\sigma^2}{n\varepsilon B} \right)$ . Similar to Section 6.2, we see that in  
 243 the regimes when  $\hat{L} \approx L_\sigma$  the third term w.r.t.  $1/\varepsilon^{3/2}$  can degenerate up to  $\sqrt{B}/p_a$ . However, if we take  
 244  $B = \mathcal{O}(L_\sigma^2/\hat{L}^2)$ , then the degeneration of the third term will be up to  $1/p_a$ . This effect we analyze in  
 245 Section C. The fact that the degeneration is up to  $1/p_a$  we check numerically in Section A.

246 In the following corollary, we consider RandK compressors (see Definition 5) and show that with  
 247 the particular choice of parameters, up to the Lipschitz constants factors, DASHA-PP-MVR gets the  
 248 optimal oracle complexity and SOTA communication complexity of DASHA-SYNC-MVR method.

---

<sup>4</sup>If  $\mathbb{1}_{p_a} = 0$ , then  $\frac{L_\sigma^2}{\frac{1}{p_a} \hat{L}^2} = +\infty$

249 **Corollary 4.** Suppose that assumptions of Corollary 3 hold, batch size  $B \leq \min \left\{ \frac{\sigma}{p_a \sqrt{\varepsilon n}}, \frac{L_\sigma^2}{\mathbb{1}_{p_a}^2 \bar{L}^2} \right\}$ ,  
 250 we take RandK compressors with  $K = \Theta \left( \frac{B d \sqrt{\varepsilon n}}{\sigma} \right)$ . Then the communication complexity equals

$$\mathcal{O} \left( \frac{d\sigma}{\sqrt{p_a} \sqrt{n\varepsilon}} + \frac{L_\sigma \Delta_0 d}{p_a \sqrt{n\varepsilon}} \right), \quad (12)$$

251 and the expected number of stochastic gradient calculations per node equals

$$\mathcal{O} \left( \frac{\sigma^2}{\sqrt{p_a} n\varepsilon} + \frac{L_\sigma \Delta_0 \sigma}{p_a \varepsilon^{3/2} n} \right). \quad (13)$$

252 We are aware that the initial batch size  $B_{\text{init}}$  can be suboptimal w.r.t.  $\omega$  in DASHA-PP-MVR in some  
 253 regimes (see also (Tyurin and Richtárik, 2023)). This is a side effect of mixing the variance reduction  
 254 of stochastic gradients and compression. However, Corollary 4 reveals that we can escape these  
 255 regimes by choosing the parameter  $K$  of RandK compressors in a particular way. To get the complete  
 256 picture, we analyze the same phenomenon under PL condition (see Section F) and provide a new  
 257 method DASHA-PP-SYNC-MVR (see Section G).

## 258 References

- 259 Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). QSGD: Communication-  
 260 efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Process-*  
 261 *ing Systems (NIPS)*, pages 1709–1720.
- 262 Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower  
 263 bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.
- 264 Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020). On biased compression for  
 265 distributed learning. *arXiv preprint arXiv:2002.12410*.
- 266 Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary  
 267 points i. *Mathematical Programming*, 184(1):71–120.
- 268 Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transac-*  
 269 *tions on Intelligent Systems and Technology (TIST)*, 2(3):1–27.
- 270 Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex SGD.  
 271 *arXiv preprint arXiv:1905.10018*.
- 272 Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). SPIDER: Near-optimal non-convex optimization  
 273 via stochastic path integrated differential estimator. In *NeurIPS Information Processing Systems*.
- 274 Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT  
 275 Press.
- 276 Gorbunov, E., Burlachenko, K., Li, Z., and Richtárik, P. (2021). MARINA: Faster non-convex  
 277 distributed learning with compression. In *38th International Conference on Machine Learning*.
- 278 Horváth, S., Ho, C.-Y., Horvath, L., Sahu, A. N., Canini, M., and Richtárik, P. (2019a). Natural  
 279 compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*.
- 280 Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. (2019b). Stochastic distributed  
 281 learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*.
- 282 Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by  
 283 reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–  
 284 456. PMLR.
- 285 Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K.,  
 286 Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated  
 287 learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

288 Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T.  
289 (2020a). Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint*  
290 *arXiv:2008.03606*.

291 Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020b). Scaffold:  
292 Stochastic controlled averaging for federated learning. In *International Conference on Machine*  
293 *Learning*, pages 5132–5143. PMLR.

294 Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated  
295 learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

296 Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated  
297 optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–  
298 450.

299 Li, Z., Bao, H., Zhang, X., and Richtárik, P. (2021a). PAGE: A simple and optimal probabilistic  
300 gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*,  
301 pages 6286–6295. PMLR.

302 Li, Z., Hanzely, S., and Richtárik, P. (2021b). ZeroSARAH: Efficient nonconvex finite-sum optimiza-  
303 tion with zero full gradient computation. *arXiv preprint arXiv:2103.01447*.

304 Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. (2017). Deep gradient compression: Reducing  
305 the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

306 McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-  
307 efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*,  
308 pages 1273–1282. PMLR.

309 Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. (2019). Distributed learning with  
310 compressed gradient differences. *arXiv preprint arXiv:1901.09269*.

311 Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N. R., Ganger, G. R., Gibbons,  
312 P. B., and Zaharia, M. (2019). PipeDream: generalized pipeline parallelism for dnn training. In  
313 *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pages 1–15.

314 Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

315 Nguyen, L., Liu, J., Scheinberg, K., and Takáč, M. (2017). SARAH: A novel method for machine  
316 learning problems using stochastic recursive gradient. In *The 34th International Conference on*  
317 *Machine Learning*.

318 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein,  
319 N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library.  
320 In *Advances in Neural Information Processing Systems (NeurIPS)*.

321 Patel, K. K., Wang, L., Woodworth, B., Bullins, B., and Srebro, N. (2022). Towards optimal commu-  
322 nication complexity in distributed non-convex optimization. In *Advances in Neural Information*  
323 *Processing Systems*.

324 Ramaswamy, S., Mathews, R., Rao, K., and Beaufays, F. (2019). Federated learning for emoji  
325 prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*.

326 Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021).  
327 Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

328 Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan,  
329 H. B. (2020). Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

330 Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021). EF21: A new, simpler, theoretically better, and  
331 practically faster error feedback. In *Neural Information Processing Systems, 2021*.

332 Richtárik, P., Sokolov, I., Fatkhullin, I., Gasanov, E., Li, Z., and Gorbunov, E. (2022). 3PC: Three  
333 point compressors for communication-efficient distributed training and a better theory for lazy  
334 aggregation. *arXiv preprint arXiv:2202.00998*.

335 Sapio, A., Canini, M., Ho, C.-Y., Nelson, J., Kalnis, P., Kim, C., Krishnamurthy, A., Moshref, M.,  
336 Ports, D. R., and Richtárik, P. (2019). Scaling distributed machine learning with in-network  
337 aggregation. *arXiv preprint arXiv:1903.06701*.

338 Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified SGD with memory. *Advances in*  
339 *Neural Information Processing Systems*, 31.

340 Suresh, A. T., Sun, Z., Ro, J. H., and Yu, F. (2022). Correlated quantization for distributed mean  
341 estimation and optimization. *arXiv preprint arXiv:2203.04925*.

342 Szlendak, R., Tyurin, A., and Richtárik, P. (2021). Permutation compressors for provably faster  
343 distributed nonconvex optimization. *arXiv preprint arXiv:2110.03300*.

344 Tyurin, A. and Richtárik, P. (2023). DASHA: Distributed nonconvex optimization with commu-  
345 nication compression and optimal oracle complexity. *International Conference on Learning*  
346 *Representations (ICLR)*.

347 Vogels, T., He, L., Koloskova, A., Karimireddy, S. P., Lin, T., Stich, S. U., and Jaggi, M. (2021).  
348 RelaySum for decentralized deep learning on heterogeneous data. *Advances in Neural Information*  
349 *Processing Systems*, 34.

350 Wangni, J., Wang, J., Liu, J., and Zhang, T. (2018). Gradient sparsification for communication-  
351 efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31.

352 Xu, H., Ho, C.-Y., Abdelmoniem, A. M., Dutta, A., Bergou, E. H., Karatsenidis, K., Canini, M.,  
353 and Kalnis, P. (2021). Grace: A compressed communication framework for distributed machine  
354 learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*,  
355 pages 561–572. IEEE.

356 Zhao, H., Burlachenko, K., Li, Z., and Richtárik, P. (2021a). Faster rates for compressed federated  
357 learning with client-variance reduction. *arXiv preprint arXiv:2112.13097*.

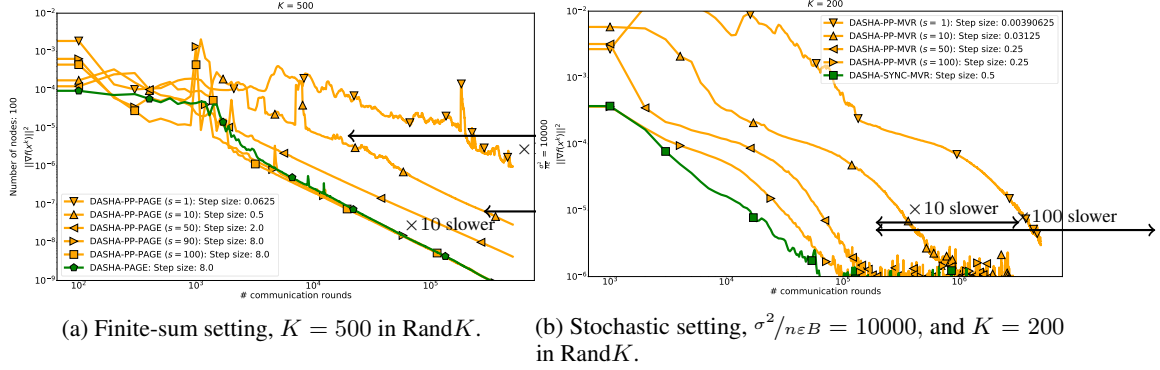
358 Zhao, H., Li, Z., and Richtárik, P. (2021b). FedPAGE: A fast local stochastic gradient method for  
359 communication-efficient federated learning. *arXiv preprint arXiv:2108.04755*.

## 360 Contents

361	<b>1 Introduction</b>	<b>1</b>
362	<b>2 Optimization Problem</b>	<b>1</b>
363	2.1 Unbiased Compressors . . . . .	2
364	2.2 Nodes Partial Participation Assumptions . . . . .	4
365	<b>3 Motivation and Related Work</b>	<b>4</b>
366	<b>4 Contributions</b>	<b>5</b>
367	<b>5 Algorithm Description and Main Challenges Towards Partial Participation</b>	<b>7</b>
368	<b>6 Theorems</b>	<b>7</b>
369	6.1 Gradient Setting . . . . .	8
370	6.2 Finite-Sum Setting . . . . .	8
371	6.3 Stochastic Setting . . . . .	9
372	<b>A Numerical Verification of Theoretical Dependencies</b>	<b>15</b>
373	<b>B Original DASHA and DASHA-MVR Methods</b>	<b>16</b>
374	<b>C Problem of Estimating the Mean in the Partial Participation Setting</b>	<b>17</b>
375	<b>D Auxiliary facts</b>	<b>18</b>
376	D.1 Sampling Lemma . . . . .	18
377	D.2 Compressors Facts . . . . .	20
378	<b>E Proofs of Theorems</b>	<b>20</b>
379	E.1 Standard Lemmas in the Nonconvex Setting . . . . .	21
380	E.2 Generic Lemmas . . . . .	22
381	E.3 Proof for DASHA-PP . . . . .	25
382	E.4 Proof for DASHA-PP-PAGE . . . . .	29
383	E.5 Proof for DASHA-PP-FINITE-MVR . . . . .	40
384	E.6 Proof for DASHA-PP-MVR . . . . .	51
385	<b>F Analysis of DASHA-PP under Polyak-Łojasiewicz Condition</b>	<b>64</b>
386	F.1 Gradient Setting . . . . .	64
387	F.2 Finite-Sum Setting . . . . .	64
388	F.3 Stochastic Setting . . . . .	65
389	F.4 Proofs of Theorems . . . . .	66
390	F.4.1 Standard Lemma under Polyak-Łojasiewicz Condition . . . . .	66
391	F.4.2 Generic Lemma . . . . .	66

392	F.4.3 Proof for DASHA-PP under PL-condition . . . . .	68
393	F.4.4 Proof for DASHA-PP-PAGE under PL-condition . . . . .	70
394	F.4.5 Proof for DASHA-PP-MVR under PL-condition . . . . .	75
395	<b>G Description of DASHA-PP-SYNC-MVR</b>	<b>82</b>
396	G.1 Proof for DASHA-PP-SYNC-MVR . . . . .	84




 Figure 1: Classification task with the *real-sim* dataset.

Our main goal is to verify the dependeces from the theory. We compare **DASHA-PP** with **DASHA**. Clearly, **DASHA-PP** can not generally perform better than **DASHA**. In different settings, we verify that the bigger  $p_a$ , the closer **DASHA-PP** is to **DASHA**, i.e., **DASHA-PP** converges no slower than  $1/p_a$  times.

In all experiments, we take the *real-sim* dataset with dimension  $d = 20,958$  and the number of samples equals 72,309 from LIBSVM datasets (Chang and Lin, 2011) (under the 3-clause BSD license), and randomly split the dataset between  $n = 100$  nodes equally, ignoring residual samples. In the finite-sum setting, we solve a classification problem with functions

$$f_i(x) := \frac{1}{m} \sum_{j=1}^m \left( 1 - \frac{1}{1 + \exp(y_{ij} a_{ij}^\top x)} \right)^2,$$

where  $a_{ij} \in \mathbb{R}^d$  is the feature vector of a sample on the  $i^{\text{th}}$  node,  $y_{ij} \in \{-1, 1\}$  is the corresponding label, and  $m$  is the number of samples on the  $i^{\text{th}}$  node for all  $i \in [n]$ . In the stochastic setting, we consider functions

$$f_i(x_1, x_2) := \mathbb{E}_{j \sim [m]} \left[ -\log \left( \frac{\exp(a_{ij}^\top x_1 y_{ij})}{\sum_{y \in \{1, 2\}} \exp(a_{ij}^\top x_y y)} \right) + \lambda \sum_{y \in \{1, 2\}} \sum_{k=1}^d \frac{\{x_y\}_k^2}{1 + \{x_y\}_k^2} \right],$$

where  $x_1, x_2 \in \mathbb{R}^d$ ,  $\{\cdot\}_k$  is an indexing operation,  $a_{ij} \in \mathbb{R}^d$  is a feature of a sample on the  $i^{\text{th}}$  node,  $y_{ij} \in \{1, 2\}$  is a corresponding label,  $m$  is the number of samples located on the  $i^{\text{th}}$  node, constant  $\lambda = 0.001$  for all  $i \in [n]$ .

The code was written in Python 3.6.8 using PyTorch 1.9 (Paszke et al., 2019). A distributed environment was emulated on a machine with Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz and 64 cores.

We use the standard setting in experiments<sup>5</sup> where all parameters except step sizes are taken as suggested in theory. Step sizes are finetuned from a set  $\{2^i \mid i \in [-10, 10]\}$ . We emulate the partial participation setting using  $s$ -nice sampling with the number of nodes  $n = 100$ . We consider the  $\text{Rand}K$  compressor and take the batch size  $B = 1$ . We plot the relation between communication rounds and values of the norm of gradients at each communication round.

In the finite-sum (Figure 1a) and in the stochastic setting (Figure 1b), we see that the bigger probability  $p_a = s/n$  to 1, the closer **DASHA-PP** to **DASHA**. Moreover, **DASHA-PP** with  $s = 10$  and  $s = 1$  converges approximately  $\times 10$  ( $= 1/p_a$ ) and  $\times 100$  ( $= 1/p_a$ ) times slower, accordingly. Our theory predicts such behavior.

<sup>5</sup>Code: <https://github.com/mysteryresearcher/dasha-partial-participation>

## 420 B Original DASHA and DASHA-MVR Methods

421 To simplify the discussion and explanation from the main part, we present the algorithms from (Tyurin  
422 and Richtárik, 2023)

---

### Algorithm 6 DASHA

---

```

1: Input: starting point  $x^0 \in \mathbb{R}^d$ , stepsize  $\gamma > 0$ , momentum  $a \in (0, 1]$ , number of iterations  $T \geq 1$ 
2: Initialize  $g_i^0 \in \mathbb{R}^d$  on the nodes and  $g^0 = \frac{1}{n} \sum_{i=1}^n g_i^0$  on the server
3: for  $t = 0, 1, \dots, T - 1$  do
4:    $x^{t+1} = x^t - \gamma g^t$ 
5:   Broadcast  $x^{t+1}, x^t$  to all participating(a) nodes
6:   for  $i = 1, \dots, n$  in parallel do
7:      $m_i^{t+1} = \mathcal{C}_i(\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - a(g_i^t - \nabla f_i(x^t)))$ 
8:      $g_i^{t+1} = g_i^t + m_i^{t+1}$ 
9:     Send  $m_i^{t+1}$  to the server
10:  end for
11:   $g^{t+1} = g^t + \frac{1}{n} \sum_{i=1}^n m_i^{t+1}$ 
12: end for
13: Output:  $\hat{x}^T$  chosen uniformly at random from  $\{x^t\}_{k=0}^{T-1}$ 

```

---



---

### Algorithm 7 DASHA-MVR (with batch size $B = 1$ )

---

```

1: Input: starting point  $x^0 \in \mathbb{R}^d$ , stepsize  $\gamma > 0$ , momentums  $a, b \in (0, 1]$ , number of iterations  $T \geq 1$ 
2: Initialize  $g_i^0 \in \mathbb{R}^d$  on the nodes and  $g^0 = \frac{1}{n} \sum_{i=1}^n g_i^0$  on the server
3: for  $t = 0, 1, \dots, T - 1$  do
4:    $x^{t+1} = x^t - \gamma g^t$ 
5:   Broadcast  $x^{t+1}, x^t$  to all participating(a) nodes
6:   for  $i = 1, \dots, n$  in parallel do
7:      $h_i^{t+1} = \nabla f_i(x^{t+1}; \xi_i^{t+1}) + (1 - b)(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})), \quad \xi_i^{t+1} \sim \mathcal{D}_i$ 
8:      $m_i^{t+1} = \mathcal{C}_i(h_i^{t+1} - h_i^t - a(g_i^t - h_i^t))$ 
9:      $g_i^{t+1} = g_i^t + m_i^{t+1}$ 
10:    Send  $m_i^{t+1}$  to the server
11:  end for
12:   $g^{t+1} = g^t + \frac{1}{n} \sum_{i=1}^n m_i^{t+1}$ 
13: end for
14: Output:  $\hat{x}^T$  chosen uniformly at random from  $\{x^t\}_{k=0}^{T-1}$ 

```

---

## C Problem of Estimating the Mean in the Partial Participation Setting

We now provide the example to explain why the only choice of  $B = \mathcal{O}\left(\min\left\{\frac{1}{p_a}\sqrt{\frac{m}{n}}, \frac{L_{\max}^2}{p_a L^2}\right\}\right)$  and

$B = \mathcal{O}\left(\min\left\{\frac{\sigma}{p_a\sqrt{\varepsilon n}}, \frac{L_{\sigma}^2}{p_a^2 \hat{\mathcal{L}}^2}\right\}\right)$  in **DASHA-PP-PAGE** and **DASHA-PP-MVR**, accordingly, guarantees the degeneration up to  $1/p_a$ . This is surprising, because in methods with the variance reduction of stochastic gradients (Li et al., 2021a; Tyurin and Richtárik, 2023) we can take the size of batch size  $B = \mathcal{O}(\sqrt{\frac{m}{n}})$  and  $B = \mathcal{O}\left(\frac{\sigma}{\sqrt{\varepsilon n}}\right)$  and guarantee the optimality. Note that the smaller the batch size  $B$ , the more the server and the nodes have to communicate to get  $\varepsilon$ -solution.

Let us consider the task of estimating the mean of vectors in the distributed setting. Suppose that we have  $n$  nodes, and each of them contains  $m$  vectors  $\{x_{ij}\}_{j=1}^m$ , where  $x_{ij} \in \mathbb{R}^d$  for all  $i \in [n], j \in [m]$ . First, let us consider that each node samples a mini-batch  $I^i$  of size  $B$  with replacement and sends it to the server. Then the server calculates the mean of the mini-batches from nodes. One can easily show that the variance of the estimator is

$$\begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{nB} \sum_{i=1}^n \sum_{j \in I^i} x_{ij} - \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m x_{ij} \right\|^2 \right] \\ = \frac{1}{nB} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \left\| x_{ij} - \frac{1}{m} \sum_{j=1}^m x_{ij} \right\|^2. \end{aligned} \quad (14)$$

Next, we consider the same task in the partial participation setting with  $s$ -nice sampling, i.e., we sample a random set  $S \subset [n]$  of  $s \in [n]$  nodes without replacement and receive the mini-batches only from the sampled nodes. Such sampling of nodes satisfy Assumption 8 with  $p_a = s/n$  and  $p_a = s(s-1)/n(n-1)$ . In this case, the variance of the estimator (See Lemma 1 with  $r_i = 0$  and  $s_i = \sum_{j \in I^i} x_{ij}$ ) is

$$\begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{sB} \sum_{i \in S} \sum_{j \in I^i} x_{ij} - \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m x_{ij} \right\|^2 \right] \\ = \underbrace{\frac{1}{sB} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \left\| x_{ij} - \frac{1}{m} \sum_{j=1}^m x_{ij} \right\|^2}_{\mathcal{L}_{\max}^2} \\ + \underbrace{\frac{n-s}{s(n-1)} \frac{1}{n} \sum_{i=1}^n \left\| \frac{1}{m} \sum_{j=1}^m x_{ij} - \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m x_{ij} \right\|^2}_{\hat{\mathcal{L}}^2}. \end{aligned} \quad (15)$$

Let us assume that  $s \leq n/2$ . Note that (14) scales with any  $B \geq 1$ , while (15) only scales when  $B = \mathcal{O}(\mathcal{L}_{\max}^2/\hat{\mathcal{L}}^2)$ . In other words, for large enough  $B$ , the variance in (15) does not significantly improves with the growth of  $B$  due to the term  $\hat{\mathcal{L}}^2$ . In our proof, due to partial participation, the variance from (15) naturally appears, and we get the same effect. As was mentioned in Sections 6.2 and 6.3, it can be seen in our convergence rate bounds.

## 445 D Auxiliary facts

446 We list auxiliary facts that we use in our proofs:

447 1. For all  $x, y \in \mathbb{R}^d$ , we have

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 \quad (16)$$

448 2. Let us take a *random vector*  $\xi \in \mathbb{R}^d$ , then

$$\mathbb{E} \left[ \|\xi\|^2 \right] = \mathbb{E} \left[ \|\xi - \mathbb{E}[\xi]\|^2 \right] + \|\mathbb{E}[\xi]\|^2. \quad (17)$$

### 449 D.1 Sampling Lemma

450 This section provides a lemma that we regularly use in our proofs, and it is useful for samplings that  
451 satisfy Assumption 8.

452 **Lemma 1.** *Suppose that a set  $S$  is a random subset of a set  $[n]$  such that*

453 1.  $\text{Prob}(i \in S) = p_a, \quad \forall i \in [n],$

454 2.  $\text{Prob}(i \in S, j \in S) = p_{aa}, \quad \forall i \neq j \in [n],$

455 3.  $p_{aa} \leq p_a^2,$

456 where  $p_a \in (0, 1]$  and  $p_{aa} \in [0, 1]$ . Let us take random independent vectors  $s_i \in \mathbb{R}^d$  for all  $i \in [n]$ ,  
457 nonrandom vector  $r_i \in \mathbb{R}^d$  for all  $i \in [n]$ , and random vectors

$$v_i = \begin{cases} r_i + \frac{1}{p_a} s_i, & i \in S, \\ r_i, & i \notin S, \end{cases}$$

458 then

$$\begin{aligned} & \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n v_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n v_i \right] \right\|^2 \right] \\ &= \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E} \left[ \|s_i - \mathbb{E}[s_i]\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\mathbb{E}[s_i]\|^2 + \frac{p_{aa} - p_a^2}{p_a^2} \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[s_i] \right\|^2 \\ &\leq \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E} \left[ \|s_i - \mathbb{E}[s_i]\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\mathbb{E}[s_i]\|^2. \end{aligned}$$

459 *Proof.* Let us define additional constants  $p_{an}$  and  $p_{nn}$ , such that

460 1.  $\text{Prob}(i \in S, j \notin S) = p_{an}, \quad \forall i \neq j \in [n],$

461 2.  $\text{Prob}(i \notin S, j \notin S) = p_{nn}, \quad \forall i \neq j \in [n].$

462 Note, that

$$p_{an} = p_{aa} - p_a \quad (18)$$

463 and

$$p_{nn} = 1 - p_{aa} - 2p_{an}. \quad (19)$$

464 Using the law of total expectation and

$$\mathbb{E}[v_i] = p_a \left( r_i + \mathbb{E} \left[ \frac{1}{p_a} s_i \right] \right) + (1 - p_a) r_i = r_i + \mathbb{E}[s_i],$$

465 we have

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n v_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n v_i \right] \right\|^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \|v_i - (r_i + \mathbb{E}[s_i])\|^2 \right] \\
&\quad + \frac{1}{n^2} \sum_{i \neq j}^n \mathbb{E} [\langle v_i - (r_i + \mathbb{E}[s_i]), v_j - (r_j + \mathbb{E}[s_j]) \rangle] \\
&= \frac{p_a}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \left\| r_i + \frac{1}{p_a} s_i - (r_i + \mathbb{E}[s_i]) \right\|^2 \right] \\
&\quad + \frac{1-p_a}{n^2} \sum_{i=1}^n \|r_i - (r_i + \mathbb{E}[s_i])\|^2 \\
&\quad + \frac{p_{aa}}{n^2} \sum_{i \neq j}^n \mathbb{E} \left[ \left\langle r_i + \frac{1}{p_a} s_i - (r_i + \mathbb{E}[s_i]), r_j + \frac{1}{p_a} s_j - (r_j + \mathbb{E}[s_j]) \right\rangle \right] \\
&\quad + \frac{2p_{an}}{n^2} \sum_{i \neq j}^n \mathbb{E} \left[ \left\langle r_i + \frac{1}{p_a} s_i - (r_i + \mathbb{E}[s_i]), r_j - (r_j + \mathbb{E}[s_j]) \right\rangle \right] \\
&\quad + \frac{p_{nn}}{n^2} \sum_{i \neq j}^n \langle r_i - (r_i + \mathbb{E}[s_i]), r_j - (r_j + \mathbb{E}[s_j]) \rangle.
\end{aligned}$$

466 From the independence of random vectors  $s_i$ , we obtain

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n v_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n v_i \right] \right\|^2 \right] \\
&= \frac{p_a}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \left\| \frac{1}{p_a} s_i - \mathbb{E}[s_i] \right\|^2 \right] \\
&\quad + \frac{1-p_a}{n^2} \sum_{i=1}^n \|\mathbb{E}[s_i]\|^2 \\
&\quad + \frac{p_{aa}(1-p_a)^2}{n^2 p_a^2} \sum_{i \neq j}^n \langle \mathbb{E}[s_i], \mathbb{E}[s_j] \rangle \\
&\quad + \frac{2p_{an}(p_a-1)}{n^2 p_a} \sum_{i \neq j}^n \langle \mathbb{E}[s_i], \mathbb{E}[s_j] \rangle \\
&\quad + \frac{p_{nn}}{n^2} \sum_{i \neq j}^n \langle \mathbb{E}[s_i], \mathbb{E}[s_j] \rangle.
\end{aligned}$$

467 Using (18) and (19), we have

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n v_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n v_i \right] \right\|^2 \right] \\
&= \frac{p_a}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \left\| \frac{1}{p_a} s_i - \mathbb{E}[s_i] \right\|^2 \right] \\
&\quad + \frac{1-p_a}{n^2} \sum_{i=1}^n \|\mathbb{E}[s_i]\|^2
\end{aligned}$$

$$\begin{aligned}
& + \frac{p_{aa} - p_a^2}{n^2 p_a^2} \sum_{i \neq j}^n \langle E[s_i], E[s_j] \rangle \\
& \stackrel{(17)}{=} \frac{1}{n^2 p_a} \sum_{i=1}^n E \left[ \|s_i - E[s_i]\|^2 \right] \\
& + \frac{1 - p_a}{n^2 p_a} \sum_{i=1}^n \|E[s_i]\|^2 \\
& + \frac{p_{aa} - p_a^2}{n^2 p_a^2} \sum_{i \neq j}^n \langle E[s_i], E[s_j] \rangle \\
& = \frac{1}{n^2 p_a} \sum_{i=1}^n E \left[ \|s_i - E[s_i]\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|E[s_i]\|^2 \\
& + \frac{p_{aa} - p_a^2}{p_a^2} \left\| \frac{1}{n} \sum_{i=1}^n E[s_i] \right\|^2.
\end{aligned}$$

468 Finally, using that  $p_{aa} \leq p_a^2$ , we have

$$\begin{aligned}
& E \left[ \left\| \frac{1}{n} \sum_{i=1}^n v_i - E \left[ \frac{1}{n} \sum_{i=1}^n v_i \right] \right\|^2 \right] \\
& \leq \frac{1}{n^2 p_a} \sum_{i=1}^n E \left[ \|s_i - E[s_i]\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|E[s_i]\|^2.
\end{aligned}$$

469

□

## 470 D.2 Compressors Facts

471 We define the *RandK* compressor that chooses without replacement  $K$  coordinates, scales them by a  
472 constant factor to preserve unbiasedness and zero-out other coordinates.

**Definition 5.** Let us take a random subset  $S$  from  $[d]$ ,  $|S| = K$ ,  $K \in [d]$ . We say that a stochastic mapping  $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is *RandK* if

$$\mathcal{C}(x) = \frac{d}{K} \sum_{j \in S} x_j e_j,$$

473 where  $\{e_i\}_{i=1}^d$  is the standard unit basis.

474 **Theorem 6.** If  $\mathcal{C}$  is *RandK*, then  $\mathcal{C} \in \mathbb{U} \left( \frac{d}{K} - 1 \right)$ .

475 See the proof in (Beznosikov et al., 2020).

## 476 E Proofs of Theorems

477 There are three different sources of randomness in Algorithm 1: the first one from vectors  $\{k_i^{t+1}\}_{i=1}^n$ ,  
478 the second one from compressors  $\{\mathcal{C}_i\}_{i=1}^n$ , and the third one from availability of nodes. We define  
479  $E_k[\cdot]$ ,  $E_C[\cdot]$  and  $E_{p_a}[\cdot]$  to be conditional expectations w.r.t.  $\{k_i^{t+1}\}_{i=1}^n$ ,  $\{\mathcal{C}_i\}_{i=1}^n$ , and availability,  
480 accordingly, conditioned on all previous randomness. Moreover, we define  $E_{t+1}[\cdot]$  to be a conditional  
481 expectation w.r.t. all randomness in iteration  $t+1$  conditioned on all previous randomness. Note,  
482 that  $E_{t+1}[\cdot] = E_k[E_C[E_{p_a}[\cdot]]]$ .

483 In the case of **DASHA-PP-PAGE**, there are two different sources of randomness from  $\{k_i^{t+1}\}_{i=1}^n$ .  
484 We define  $E_{p_{\text{page}}}[\cdot]$  and  $E_B[\cdot]$  to be conditional expectations w.r.t. the probabilistic switching and  
485 mini-batch indices  $I_i^t$ , accordingly, conditioned on all previous randomness. Note, that  $E_{t+1}[\cdot] =$   
486  $E_B[E_C[E_{p_a}[E_{p_{\text{page}}}[\cdot]]]]$  and  $E_{t+1}[\cdot] = E_B[E_{p_{\text{page}}}[E_C[E_{p_a}[\cdot]]]]$ .



## 487 E.1 Standard Lemmas in the Nonconvex Setting

488 We start the proof of theorems by providing standard lemmas from the nonconvex optimization.

489 **Lemma 2.** Suppose that Assumption 2 holds and let  $x^{t+1} = x^t - \gamma g^t$ . Then for any  $g^t \in \mathbb{R}^d$  and  
 490  $\gamma > 0$ , we have

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2. \quad (20)$$

491 *Proof.* Using  $L$ -smoothness, we have

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2} \|x^{t+1} - x^t\|^2 \\ &= f(x^t) - \gamma \langle \nabla f(x^t), g^t \rangle + \frac{L}{2} \|x^{t+1} - x^t\|^2. \end{aligned}$$

492 Next, due to  $-\langle x, y \rangle = \frac{1}{2} \|x - y\|^2 - \frac{1}{2} \|x\|^2 - \frac{1}{2} \|y\|^2$ , we obtain

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2.$$

493

□

494 **Lemma 3.** Suppose that Assumption 1 holds and

$$\mathbb{E} [f(x^{t+1})] + \gamma \Psi^{t+1} \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] + \gamma \Psi^t + \gamma C, \quad (21)$$

495 where  $\Psi^t$  is a sequence of numbers,  $\Psi^t \geq 0$  for all  $t \in [T]$ , constant  $C \geq 0$ , and constant  $\gamma > 0$ .  
 496 Then

$$\mathbb{E} [\|\nabla f(\hat{x}^T)\|^2] \leq \frac{2\Delta_0}{\gamma T} + \frac{2\Psi^0}{T} + 2C, \quad (22)$$

497 where a point  $\hat{x}^T$  is chosen uniformly from a set of points  $\{x^t\}_{t=0}^{T-1}$ .

498 *Proof.* By unrolling (21) for  $t$  from 0 to  $T - 1$ , we obtain

$$\frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(x^t)\|^2] + \mathbb{E} [f(x^T)] + \gamma \Psi^T \leq f(x^0) + \gamma \Psi^0 + \gamma TC.$$

499 We subtract  $f^*$ , divide inequality by  $\frac{\gamma T}{2}$ , and take into account that  $f(x) \geq f^*$  for all  $x \in \mathbb{R}$ , and  
 500  $\Psi^t \geq 0$  for all  $t \in [T]$ , to get the following inequality:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(x^t)\|^2] \leq \frac{2\Delta_0}{\gamma T} + \frac{2\Psi^0}{T} + 2C.$$

501 It is left to consider the choice of a point  $\hat{x}^T$  to complete the proof of the lemma. □

**Lemma 4.** If  $0 < \gamma \leq (L + \sqrt{A})^{-1}$ ,  $L > 0$ , and  $A \geq 0$ , then

$$\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma A}{2} \geq 0.$$

502 The lemma can be easily checked with the direct calculation.

## 503 E.2 Generic Lemmas

504 **Lemma 5.** Suppose that Assumptions 7 and 8 hold and let us consider sequences  $g_i^{t+1}$ ,  $h_i^{t+1}$ , and  
 505  $k_i^{t+1}$  from Algorithm 1, then

$$\begin{aligned} & \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \\ & \leq \frac{2\omega}{n^2 p_a} \sum_{i=1}^n \|k_i^{t+1}\|^2 + \frac{a^2((2\omega + 1)p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2, \end{aligned} \quad (23)$$

506 and

$$\begin{aligned} & \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \\ & \leq \frac{2\omega}{p_a} \|k_i^{t+1}\|^2 + \left( \frac{a^2(2\omega + 1 - p_a)}{p_a} + (1-a)^2 \right) \|g_i^t - h_i^t\|^2 \quad \forall i \in [n]. \end{aligned} \quad (24)$$

507 *Proof.* First, we estimate  $\mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right]$ :

$$\begin{aligned} & \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \\ & = \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1} - \mathbb{E}_C [\mathbb{E}_{p_a} [g^{t+1} - h^{t+1}]]\|^2 \right] \right] + \|\mathbb{E}_C [\mathbb{E}_{p_a} [g^{t+1} - h^{t+1}]]\|^2, \end{aligned}$$

508 where we used (17). Due to Assumption 8, we have

$$\begin{aligned} & \mathbb{E}_C [\mathbb{E}_{p_a} [g_i^{t+1}]] \\ & = p_a \mathbb{E}_C \left[ g_i^t + \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right] + (1-p_a) g_i^t \\ & = g_i^t + p_a \mathbb{E}_C \left[ \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right] \\ & = g_i^t + k_i^{t+1} - a (g_i^t - h_i^t), \end{aligned}$$

509 and

$$\mathbb{E}_C [\mathbb{E}_{p_a} [h_i^{t+1}]] = p_a \mathbb{E}_C \left[ h_i^t + \frac{1}{p_a} k_i^{t+1} \right] + (1-p_a) h_i^t = h_i^t + k_i^{t+1}.$$

510 Thus, we can get

$$\begin{aligned} & \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \\ & = \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1} - \mathbb{E}_C [\mathbb{E}_{p_a} [g^{t+1} - h^{t+1}]]\|^2 \right] \right] + (1-a)^2 \|g^t - h^t\|^2. \end{aligned}$$

511 Due to the independence of compressors, we can use Lemma 1 with  $r_i = g_i^t - h_i^t$  and  $s_i =$

512  $p_a \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - k_i^{t+1}$ , and obtain

$$\begin{aligned} & \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \\ & \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_C \left[ \left\| p_a \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - k_i^{t+1} - \mathbb{E}_C \left[ p_a \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - k_i^{t+1} \right] \right\|^2 \right] \\ & \quad + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \mathbb{E}_C \left[ p_a \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - k_i^{t+1} \right] \right\|^2 \end{aligned}$$

$$\begin{aligned}
& + (1-a)^2 \|g^t - h^t\|^2 \\
& = \frac{p_a}{n^2} \sum_{i=1}^n \mathbb{E}_{\mathcal{C}} \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right\|^2 \right] \\
& \quad + \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2.
\end{aligned}$$

513 From Assumption 7, we have

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \\
& \leq \frac{\omega p_a}{n^2} \sum_{i=1}^n \left\| \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right\|^2 + \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \\
& = \frac{\omega}{n^2 p_a} \sum_{i=1}^n \|k_i^{t+1} - a (g_i^t - h_i^t)\|^2 + \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \\
& \stackrel{(16)}{\leq} \frac{2\omega}{n^2 p_a} \sum_{i=1}^n \|k_i^{t+1}\|^2 + \frac{a^2 ((2\omega + 1)p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2.
\end{aligned}$$

514 The second inequality can be proved almost in the same way:

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \\
& = \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g_i^{t+1} - h_i^{t+1} - \mathbb{E}_{\mathcal{C}} [\mathbb{E}_{p_a} [g_i^{t+1} - h_i^{t+1}]]\|^2 \right] \right] + \|\mathbb{E}_{\mathcal{C}} [\mathbb{E}_{p_a} [g_i^{t+1} - h_i^{t+1}]]\|^2 \\
& = \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g_i^{t+1} - h_i^{t+1} - g_i^t + a (g_i^t - h_i^t) + h_i^t\|^2 \right] \right] + (1-a)^2 \|g_i^t - h_i^t\|^2 \\
& = p_a \mathbb{E}_{\mathcal{C}} \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \frac{1}{p_a} k_i^{t+1} + a (g_i^t - h_i^t) \right\|^2 \right] \\
& \quad + a^2 (1-p_a) \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g_i^t - h_i^t\|^2 \\
& \stackrel{(17)}{=} p_a \mathbb{E}_{\mathcal{C}} \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right\|^2 \right] \\
& \quad + a^2 \frac{(1-p_a)^2}{p_a} \|g_i^t - h_i^t\|^2 \\
& \quad + a^2 (1-p_a) \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g_i^t - h_i^t\|^2 \\
& \leq \frac{\omega}{p_a} \|k_i^{t+1} - a (g_i^t - h_i^t)\|^2 \\
& \quad + \frac{a^2 (1-p_a)}{p_a} \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g_i^t - h_i^t\|^2 \\
& \stackrel{(16)}{\leq} \frac{2\omega}{p_a} \|k_i^{t+1}\|^2 + \frac{a^2 (2\omega + 1 - p_a)}{p_a} \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g_i^t - h_i^t\|^2.
\end{aligned}$$

515

□

516 **Lemma 6.** Suppose that Assumptions 2, 7, and 8 hold and let us take  $a = \frac{p_a}{2\omega+1}$ , then

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{n p_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& \quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{n p_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] + \frac{4\gamma\omega(2\omega+1)}{n p_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

517 *Proof.* Due to Lemma 2 and the update step from Line 4 in Algorithm 1, we have

$$\begin{aligned}
& \mathbb{E}_{t+1} [f(x^{t+1})] \\
& \leq \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2 \right] \\
& = \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - h^t + h^t - \nabla f(x^t)\|^2 \right] \\
& \stackrel{(17)}{\leq} \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right].
\end{aligned}$$

518 Let us fix some constants  $\kappa, \eta \in [0, \infty)$  that we will define later. Combining the last inequality,  
519 bounds (23), (24) and using the law of total expectation, we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] \\
& + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& = \mathbb{E} [\mathbb{E}_{t+1} [f(x^{t+1})]] \\
& + \kappa \mathbb{E} [\mathbb{E}_C [\mathbb{E}_{p_a} [\|g^{t+1} - h^{t+1}\|^2]]] + \eta \mathbb{E} \left[ \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right] \\
& + \kappa \mathbb{E} \left[ \frac{2\omega}{n^2 p_a} \sum_{i=1}^n \|k_i^{t+1}\|^2 + \frac{a^2((2\omega+1)p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \right] \\
& + \eta \mathbb{E} \left[ \frac{2\omega}{n p_a} \sum_{i=1}^n \|k_i^{t+1}\|^2 + \left( \frac{a^2(2\omega+1-p_a)}{p_a} + (1-a)^2 \right) \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& = \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + (\gamma + \kappa(1-a)^2) \mathbb{E} [\|g^t - h^t\|^2] \\
& + \left( \frac{\kappa a^2((2\omega+1)p_a - p_{aa})}{n p_a^2} + \eta \left( \frac{a^2(2\omega+1-p_a)}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left( \frac{2\kappa\omega}{n p_a} + \frac{2\eta\omega}{p_a} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

520 Now, by taking  $\kappa = \frac{\gamma}{a}$ , we can see that  $\gamma + \kappa(1-a)^2 \leq \kappa$ , and thus

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] \\
& + \frac{\gamma}{a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma}{a} \mathbb{E} [\|g^t - h^t\|^2] \\
& + \left( \frac{\gamma a((2\omega+1)p_a - p_{aa})}{n p_a^2} + \eta \left( \frac{a^2(2\omega+1-p_a)}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left( \frac{2\gamma\omega}{a n p_a} + \frac{2\eta\omega}{p_a} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

521 Next, by taking  $\eta = \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2}$  and considering the choice of  $a$ , one can show that  
 522  $\left( \frac{\gamma a((2\omega+1)p_a - p_{aa})}{np_a^2} + \eta \left( \frac{a^2(2\omega+1-p_a)}{p_a} + (1-a)^2 \right) \right) \leq \eta$ . Thus

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + \left( \frac{2\gamma(2\omega+1)\omega}{np_a^2} + \frac{2\gamma((2\omega+1)p_a - p_{aa})\omega}{np_a^3} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right]. \end{aligned}$$

523 Considering that  $p_{aa} \geq 0$ , we can simplify the last term and get

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + \frac{4\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right]. \end{aligned}$$

524

□

### 525 E.3 Proof for DASHA-PP

526 **Lemma 7.** Suppose that Assumptions 3 and 8 hold. For  $h_i^{t+1}$  and  $k_i^{t+1}$  from Algorithm 1 (DASHA-PP)  
 527 we have

1.

$$\begin{aligned} & \mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \\ & \leq \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} \|x^{t+1} - x^t\|^2 + \frac{2b^2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

2.

$$\begin{aligned} & \mathbb{E}_{p_a} [\|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2] \\ & \leq \frac{2(1-p_a)}{p_a} L_i^2 \|x^{t+1} - x^t\|^2 + \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n]. \end{aligned}$$

3.

$$\|k_i^{t+1}\|^2 \leq 2L_i^2 \|x^{t+1} - x^t\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n].$$

528 *Proof.* First, let us proof the bound for  $\mathbb{E}_k [\mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2]]$ :

$$\mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2]$$

$$= \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \mathbb{E}_{p_a} [h^{t+1}] \right\|^2 \right] + \left\| \mathbb{E}_{p_a} [h^{t+1}] - \nabla f(x^{t+1}) \right\|^2.$$

529 Using

$$\mathbb{E}_{p_a} [h_i^{t+1}] = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))$$

530 and (17), we have

$$\begin{aligned} & \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \\ &= \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \mathbb{E}_{p_a} [h^{t+1}] \right\|^2 \right] + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2. \end{aligned}$$

531 We can use Lemma 1 with  $r_i = h_i^t$  and  $s_i = k_i^{t+1}$  to obtain

$$\begin{aligned} & \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \\ & \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \left\| k_i^{t+1} - k_i^{t+1} \right\|^2 + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| k_i^{t+1} \right\|^2 + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\ & = \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\ & \stackrel{(16)}{\leq} \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2b^2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\ & \leq \frac{2(p_a - p_{aa}) \hat{L}^2}{n p_a^2} \left\| x^{t+1} - x^t \right\|^2 + \frac{2b^2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2. \end{aligned}$$

532 In the last in inequality, we used Assumption 3. Now, we prove the second inequality:

$$\begin{aligned} & \mathbb{E}_{p_a} \left[ \left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \\ &= \mathbb{E}_{p_a} \left[ \left\| h_i^{t+1} - \mathbb{E}_{p_a} [h_i^{t+1}] \right\|^2 \right] + \left\| \mathbb{E}_{p_a} [h_i^{t+1}] - \nabla f_i(x^{t+1}) \right\|^2 \\ &= \mathbb{E}_{p_a} \left[ \left\| h_i^{t+1} - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))) \right\|^2 \right] + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\ &= \frac{(1-p_a)^2}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\ & \quad + (1-p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\ &= \frac{(1-p_a)}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\ & \leq \frac{2(1-p_a)}{p_a} L_i^2 \left\| x^{t+1} - x^t \right\|^2 + \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \left\| h_i^t - \nabla f_i(x^t) \right\|^2. \end{aligned}$$

533 Finally, the third inequality of the theorem follows from (16) and Assumption 3.  $\square$

534 **Theorem 2.** Suppose that Assumptions 1, 2, 3, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_a}{2-p_a}$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{n p_a^2} + \frac{16}{n p_a^2} \left( 1 - \frac{p_{aa}}{p_a} \right) \right]^{1/2} \hat{L} \right)^{-1},$$

535 and  $g_i^0 = h_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP), then  $\mathbb{E} \left[ \left\| \nabla f(\hat{x}^T) \right\|^2 \right] \leq \frac{2\Delta_0}{\gamma T}$ .

536 *Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 6, Lemma 7,  
537 and the law of total expectation, we obtain

$$\mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \left\| g^{t+1} - h^{t+1} \right\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{n p_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \left\| g_i^{t+1} - h_i^{t+1} \right\|^2 \right]$$



$$\begin{aligned}
& + \nu \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
= & \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} \left[ \mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \right] + \rho \mathbb{E} \left[ \mathbb{E}_{p_a} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
\leq & \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \mathbb{E} \left[ 2\hat{L}^2 \|x^{t+1} - x^t\|^2 + 2b^2 \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \nu \mathbb{E} \left[ \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} \|x^{t+1} - x^t\|^2 + \frac{2b^2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \rho \mathbb{E} \left[ \frac{2(1-p_a)\hat{L}^2}{p_a} \|x^{t+1} - x^t\|^2 + \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

538 After rearranging the terms, we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
\leq & \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\omega(2\omega + 1)\hat{L}^2}{np_a^2} - \nu \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} - \rho \frac{2(1-p_a)\hat{L}^2}{p_a} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + (\gamma + \nu(1-b)^2) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b^2\gamma\omega(2\omega + 1)}{np_a^2} + \nu \frac{2b^2(p_a - p_{aa})}{np_a^2} + \rho \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

539 By taking  $\nu = \frac{\gamma}{b}$ , one can show that  $(\gamma + \nu(1-b)^2) \leq \nu$ , and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
\leq & \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\omega(2\omega+1)\widehat{L}^2}{np_a^2} - \frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} - \rho \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b(p_a - p_{aa})}{np_a^2} + \rho \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

540 Note that  $b = \frac{p_a}{2-p_a}$ , thus

$$\begin{aligned}
& \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b(p_a - p_{aa})}{np_a^2} + \rho \left( \frac{2b^2(1-p_a)}{p_a} + (1-b)^2 \right) \right) \\
& \leq \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b(p_a - p_{aa})}{np_a^2} + \rho(1-b) \right).
\end{aligned}$$

541 And if we take  $\rho = \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}$ , then

$$\left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b(p_a - p_{aa})}{np_a^2} + \rho(1-b) \right) \leq \rho,$$

542 and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\omega(2\omega+1)\widehat{L}^2}{np_a^2} - \frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} \right. \\
& \quad \left. - \frac{16b\gamma\omega(2\omega+1)(1-p_a)\widehat{L}^2}{np_a^3} - \frac{4\gamma(p_a - p_{aa})(1-p_a)\widehat{L}^2}{np_a^3} \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

543 Let us simplify the last inequality. First, note that

$$\frac{16b\gamma\omega(2\omega+1)(1-p_a)\widehat{L}^2}{np_a^3} \leq \frac{16\gamma\omega(2\omega+1)\widehat{L}^2}{np_a^2},$$

544 due to  $b \leq p_a$ . Second,

$$\frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} \leq \frac{4\gamma(p_a - p_{aa})\widehat{L}^2}{np_a^3},$$

545 due to  $b \geq \frac{p_a}{2}$ . All in all, we have

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma\omega(2\omega+1)\widehat{L}^2}{np_a^2} - \frac{8\gamma(p_a - p_{aa})\widehat{L}^2}{np_a^3}\right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
\end{aligned}$$

546 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
\end{aligned}$$

547 It is left to apply Lemma 3 with

$$\begin{aligned}
\Psi^t &= \frac{(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad + \frac{1}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\omega(2\omega+1)}{np_a^2} + \frac{2(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right]
\end{aligned}$$

548 to conclude the proof.  $\square$

#### 549 **E.4 Proof for DASHA-PP-PAGE**

550 Let us denote

$$\begin{aligned}
k_{i,1}^{t+1} &:= \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)), \\
k_{i,2}^{t+1} &:= \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)), \\
h_{i,1}^{t+1} &:= \begin{cases} h_i^t + \frac{1}{p_a} k_{i,1}^{t+1}, & i^{\text{th}} \text{ node is participating,} \\ h_i^t, & \text{otherwise,} \end{cases} \\
h_{i,2}^{t+1} &:= \begin{cases} h_i^t + \frac{1}{p_a} k_{i,2}^{t+1}, & i^{\text{th}} \text{ node is participating,} \\ h_i^t, & \text{otherwise,} \end{cases}
\end{aligned}$$

551  $h_1^{t+1} := \frac{1}{n} \sum_{i=1}^n h_{i,1}^{t+1}$ , and  $h_2^{t+1} := \frac{1}{n} \sum_{i=1}^n h_{i,2}^{t+1}$ . Note, that

$$h^{t+1} = \begin{cases} h_1^{t+1}, & \text{with probability } p_{\text{page}}, \\ h_2^{t+1}, & \text{with probability } 1 - p_{\text{page}}. \end{cases}$$

552 **Lemma 8.** Suppose that Assumptions 3, 4, and 8 hold. For  $h_i^{t+1}$  and  $k_i^{t+1}$  from Algorithm 1  
 553 (DASHA-PP-PAGE) we have

1.

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \\ & \leq \left( \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} + \frac{(1 - p_{\text{page}})L_{\max}^2}{np_a B} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{page}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

2.

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\ & \leq \left( \frac{2(1 - p_a)L_i^2}{p_a} + \frac{(1 - p_{\text{page}})L_{\max}^2}{p_a B} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \left( \frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n]. \end{aligned}$$

3.

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|k_i^{t+1}\|^2 \right] \right] \\ & \leq \left( 2L_i^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n]. \end{aligned}$$

554 *Proof.* First, we prove the first inequality of the theorem:

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \\ & = p_{\text{page}} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_1^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_2^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right]. \end{aligned}$$

555 Using

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] = \\ & = p_a h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) + (1 - p_a) h_i^t \\ & = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)). \end{aligned}$$

556 and

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] = \\ & = p_a h_i^t + \mathbb{E}_B \left[ \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) \right] + (1 - p_a) h_i^t \\ & = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t), \end{aligned}$$

557 we obtain

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \\ & \stackrel{(17)}{=} p_{\text{page}} \mathbb{E}_{p_a} \left[ \|h_1^{t+1} - \mathbb{E}_{p_a} [h_1^{t+1}]\|^2 \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_2^{t+1} - \mathbb{E}_{p_a} [h_2^{t+1}]\|^2 \right] \right] \end{aligned}$$

$$\begin{aligned}
& + p_{\text{page}} \left\| \mathbb{E}_{p_a} [h_1^{t+1}] - \nabla f(x^{t+1}) \right\|^2 + (1 - p_{\text{page}}) \left\| \mathbb{E}_B [\mathbb{E}_{p_a} [h_2^{t+1}]] - \nabla f(x^{t+1}) \right\|^2 \\
& = p_{\text{page}} \mathbb{E}_{p_a} \left[ \left\| h_1^{t+1} - \mathbb{E}_{p_a} [h_1^{t+1}] \right\|^2 \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \left\| h_2^{t+1} - \mathbb{E}_{p_a} [h_2^{t+1}] \right\|^2 \right] \\
& + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \left\| h^t - \nabla f(x^t) \right\|^2. \tag{25}
\end{aligned}$$

558 Next, we consider  $\mathbb{E}_{p_a} \left[ \left\| h_1^{t+1} - \mathbb{E}_{p_a} [h_1^{t+1}] \right\|^2 \right]$ . We can use Lemma 1 with  $r_i = h_i^t$  and  $s_i = k_{i,1}^{t+1}$   
559 to obtain

$$\begin{aligned}
& \mathbb{E}_{p_a} \left[ \left\| h_1^{t+1} - \mathbb{E}_{p_a} [h_1^{t+1}] \right\|^2 \right] \\
& \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \left\| k_{i,1}^{t+1} - k_{i,1}^{t+1} \right\|^2 + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| k_{i,1}^{t+1} \right\|^2 \\
& = \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& \stackrel{(16)}{\leq} \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{page}}^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2.
\end{aligned}$$

560 From Assumption 3, we have

$$\begin{aligned}
& \mathbb{E}_{p_a} \left[ \left\| h_1^{t+1} - \mathbb{E}_{p_a} [h_1^{t+1}] \right\|^2 \right] \\
& \leq \frac{2(p_a - p_{aa})\hat{L}^2}{n p_a^2} \left\| x^{t+1} - x^t \right\|^2 + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{page}}^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2. \tag{26}
\end{aligned}$$

561 Now, we prove the bound for  $\mathbb{E}_B \left[ \left\| h_2^{t+1} - \mathbb{E}_B [\mathbb{E}_{p_a} [h_2^{t+1}]] \right\|^2 \right]$ . Considering that mini-  
562 batches in the algorithm are independent, we can use Lemma 1 with  $r_i = h_i^t$  and  $s_i = k_{i,2}^{t+1}$   
563 to obtain

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_2^{t+1} - \mathbb{E}_{p_a} [h_2^{t+1}] \right\|^2 \right] \right] \\
& \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - \mathbb{E}_B [k_{i,2}^{t+1}] \right\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \mathbb{E}_B [k_{i,2}^{t+1}] \right\|^2 \\
& = \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\
& = \frac{1}{n^2 p_a B^2} \sum_{i=1}^n \mathbb{E}_B \left[ \sum_{j \in I_i^t} \left\| (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\
& = \frac{1}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\
& \leq \frac{1}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) \right\|^2 + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2.
\end{aligned}$$

564 Next, we use Assumptions 3 and 4 to get

$$\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_2^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_2^{t+1} \right] \right] \right\|^2 \right] \right] \leq \left( \frac{L_{\max}^2}{np_a B} + \frac{(p_a - p_{aa}) \hat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2. \quad (27)$$

565 Applying (26) and (27) into (25), we get

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right] \right] \\ & \leq p_{\text{page}} \left( \frac{2(p_a - p_{aa}) \hat{L}^2}{np_a^2} \|x^{t+1} - x^t\|^2 + \frac{2(p_a - p_{aa}) b^2}{n^2 p_a^2 p_{\text{page}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right) + \\ & \quad + (1 - p_{\text{page}}) \left( \frac{L_{\max}^2}{np_a B} + \frac{(p_a - p_{aa}) \hat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \|h^t - \nabla f(x^t)\|^2 \\ & \leq \left( \frac{2(p_a - p_{aa}) \hat{L}^2}{np_a^2} + \frac{(1 - p_{\text{page}}) L_{\max}^2}{np_a B} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \frac{2(p_a - p_{aa}) b^2}{n^2 p_a^2 p_{\text{page}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

566 The proof of the second inequality almost repeats the previous one:

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \right] \\ & = p_{\text{page}} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,2}^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \\ & \stackrel{(17)}{=} p_{\text{page}} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,2}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] \right\|^2 \right] \right] \\ & \quad + p_{\text{page}} \left\| \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] - \nabla f_i(x^{t+1}) \right\|^2 + (1 - p_{\text{page}}) \left\| \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] - \nabla f_i(x^{t+1}) \right\|^2 \\ & = p_{\text{page}} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \right] + (1 - p_{\text{page}}) \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,2}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] \right\|^2 \right] \right] \\ & \quad + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2. \quad (28) \end{aligned}$$

567 Let us consider  $\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \right]$ :

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \right] \\ & = \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \\ & = p_a \left\| h_i^t + \frac{1}{p_a} k_{i,1}^{t+1} - \left( h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right) \right\|^2 \\ & \quad + (1 - p_a) \left\| h_i^t - \left( h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right) \right\|^2 \\ & = \frac{(1 - p_a)^2}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\ & \quad + (1 - p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\ & = \frac{1 - p_a}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2. \end{aligned}$$



568 Considering (16) and Assumption 3, we obtain

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,1}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right] \right\|^2 \right] \right] \\ & \leq \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{2(1-p_a)b^2}{p_a p_{\text{page}}^2} \|h_i^t - \nabla f_i(x^t)\|^2. \end{aligned} \quad (29)$$

569 Next, we obtain the bound for  $\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,2}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] \right\|^2 \right] \right]$ :

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_{i,2}^{t+1} - \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ h_{i,2}^{t+1} \right] \right] \right\|^2 \right] \right] \\ & = p_a \mathbb{E}_B \left[ \left\| h_i^t + \frac{1}{p_a} k_{i,2}^{t+1} - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\ & \quad + (1-p_a) \mathbb{E}_B \left[ \left\| h_i^t - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\ & = p_a \mathbb{E}_B \left[ \left\| \frac{1}{p_a} k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\ & \quad + (1-p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\ & \stackrel{(17)}{=} \frac{1}{p_a} \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] + \frac{(1-p_a)^2}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\ & \quad + (1-p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\ & = \frac{1}{p_a} \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] + \frac{1-p_a}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\ & \leq \frac{1}{p_a} \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] + \frac{(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2, \end{aligned} \quad (30)$$

570 where we used Assumption 3. By plugging (29) and (30) into (28), we get

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \right] \\ & \leq p_{\text{page}} \left( \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{2(1-p_a)b^2}{p_a p_{\text{page}}^2} \|h_i^t - \nabla f_i(x^t)\|^2 \right) \\ & \quad + (1-p_{\text{page}}) \left( \frac{1}{p_a} \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] + \frac{(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 \right) \\ & \quad + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\ & \leq \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{1-p_{\text{page}}}{p_a} \mathbb{E}_B \left[ \left\| k_{i,2}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\ & \quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2. \end{aligned}$$

571 From the independence of elements in the mini-batch, we obtain

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \right] \\ & \leq \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{1-p_{\text{page}}}{p_a} \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\ & \quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{1-p_{\text{page}}}{p_a B^2} \mathbb{E}_B \left[ \sum_{j \in I_i^t} \|(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right] \\
&\quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\
&= \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{1-p_{\text{page}}}{m p_a B} \sum_{j=1}^m \|(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \\
&\quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\leq \frac{2(1-p_a)L_i^2}{p_a} \|x^{t+1} - x^t\|^2 + \frac{1-p_{\text{page}}}{m p_a B} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\|^2 \\
&\quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\leq \left( \frac{2(1-p_a)L_i^2}{p_a} + \frac{(1-p_{\text{page}})L_{\max}^2}{p_a B} \right) \|x^{t+1} - x^t\|^2 \\
&\quad + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2,
\end{aligned}$$

572 where we used Assumption 4. Finally, we prove the last inequality:

$$\begin{aligned}
&\mathbb{E}_B \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|k_i^{t+1}\|^2 \right] \right] \\
&= p_{\text{page}} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\
&\quad + (1-p_{\text{page}}) \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) \right\|^2 \right] \\
&\stackrel{(17)}{=} p_{\text{page}} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\
&\quad + (1-p_{\text{page}}) \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
&\quad + (1-p_{\text{page}}) \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
&\stackrel{(16)}{\leq} 2p_{\text{page}} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\quad + (1-p_{\text{page}}) \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
&\quad + (1-p_{\text{page}}) \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
&\leq 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2
\end{aligned}$$

$$+ (1 - p_{\text{page}}) \mathbb{E}_B \left[ \left\| \frac{1}{B} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right].$$

573 Using the independence of elements in the mini-batch, we have

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|k_i^{t+1}\|^2 \right] \right] \\ & \leq 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2 \\ & \quad + \frac{1 - p_{\text{page}}}{B^2} \mathbb{E}_B \left[ \sum_{j \in I_i^t} \|(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right] \\ & = 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2 \\ & \quad + \frac{1 - p_{\text{page}}}{Bm} \sum_{j=1}^m \|(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \\ & \leq 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2 \\ & \quad + \frac{1 - p_{\text{page}}}{Bm} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\|^2 \end{aligned}$$

574 It is left to consider Assumptions 3 and 4 to get

$$\begin{aligned} & \mathbb{E}_B \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|k_i^{t+1}\|^2 \right] \right] \\ & \leq \left( 2L_i^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{p_{\text{page}}} \|h_i^t - \nabla f_i(x^t)\|^2. \end{aligned}$$

575

□

576 **Theorem 3.** Suppose that Assumptions 1, 2, 3, 4, 7, and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_{\text{page}}p_a}{2-p_a}$ ,  
577 probability  $p_{\text{page}} \in (0, 1]$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) + \frac{16}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right]^{1/2} \right)^{-1}$$

578 and  $g_i^0 = h_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-PAGE) then  $\mathbb{E} [\|\nabla f(\hat{x}^T)\|^2] \leq$

579  $\frac{2\Delta_0}{\gamma T}$ .

580 *Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 6, Lemma 8,  
581 and the law of total expectation, we obtain

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & \quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& = \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_{\text{page}}} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \right] \right] \\
& + \nu \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \right] \\
& + \rho \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{page}}} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \left( 2\hat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{p_{\text{page}}} \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \nu \mathbb{E} \left( \left( \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} + \frac{(1-p_{\text{page}})L_{\max}^2}{np_a B} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{page}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \|h^t - \nabla f(x^t)\|^2 \right) \\
& + \rho \mathbb{E} \left( \left( \frac{2(1-p_a)\hat{L}^2}{p_a} + \frac{(1-p_{\text{page}})L_{\max}^2}{p_a B} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right)
\end{aligned}$$

582 After rearranging the terms, we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( 2\hat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right)
\end{aligned}$$

$$\begin{aligned}
& -\nu \left( \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} + \frac{(1 - p_{\text{page}})L_{\max}^2}{np_a B} \right) - \rho \left( \frac{2(1 - p_a)\widehat{L}^2}{p_a} + \frac{(1 - p_{\text{page}})L_{\max}^2}{p_a B} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( \gamma + \nu \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \right) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b^2\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{2\nu(p_a - p_{aa})b^2}{np_a^2 p_{\text{page}}} \right. \\
& \quad \left. + \rho \left( \frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Due to  $b = \frac{p_{\text{page}} p_a}{2 - p_a} \leq p_{\text{page}}$ , one can show that  $\left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \leq 1 - b$ . Thus, if we take  $\nu = \frac{\gamma}{b}$ , then

$$\left( \gamma + \nu \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \right) \leq \gamma + \nu(1 - b) = \nu,$$

583 therefore

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \left( 2\widehat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \frac{\gamma}{b} \left( \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} + \frac{(1 - p_{\text{page}})L_{\max}^2}{np_a B} \right) - \rho \left( \frac{2(1 - p_a)\widehat{L}^2}{p_a} + \frac{(1 - p_{\text{page}})L_{\max}^2}{p_a B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b^2\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{page}}} \right. \\
& \quad \left. + \rho \left( \frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Next, with the choice of  $b = \frac{p_{\text{page}} p_a}{2 - p_a}$ , we ensure that

$$\left( \frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \leq 1 - b.$$

If we take  $\rho = \frac{8b^2\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}}$ , then

$$\left( \frac{8b^2\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{page}}} + \rho \left( \frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \right) \leq \rho,$$

584 therefore

$$\mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right]$$

$$\begin{aligned}
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( 2\hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \right. \\
& \quad \left. - \frac{\gamma}{bnp_a} \left( 2 \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \right. \\
& \quad \left. - \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^3 p_{\text{page}}} + \frac{2\gamma \left( 1 - \frac{p_{\text{aa}}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2(1-p_a) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \right] \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Let us simplify the inequality. First, due to  $b \geq \frac{p_{\text{page}} p_a}{2}$ , we have

$$\frac{\gamma}{bnp_a} \left( 2 \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \leq \frac{4\gamma}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right).$$

585 Second, due to  $b \leq p_a p_{\text{page}}$  and  $p_{\text{aa}} \leq p_a^2$ , we get

$$\begin{aligned}
& \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^3 p_{\text{page}}} + \frac{2\gamma \left( 1 - \frac{p_{\text{aa}}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2(1-p_a) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \\
& \leq \left( \frac{8\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma \left( 1 - \frac{p_{\text{aa}}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2 \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \\
& \leq \frac{16\gamma\omega(2\omega+1)}{np_a^2} \left( \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \\
& \quad + \frac{4\gamma \left( 1 - \frac{p_{\text{aa}}}{p_a} \right)}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \\
& \leq \frac{16\gamma\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right) \\
& \quad + \frac{4\gamma}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{\text{aa}}}{p_a} \right) \hat{L}^2 + \frac{(1-p_{\text{page}})L_{\text{max}}^2}{B} \right).
\end{aligned}$$

586 Combining all bounds together, we obtain the following simplified inequality:

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma\omega(2\omega+1)}{np_a^2} \left( \widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \frac{8\gamma}{np_a^2 p_{\text{page}}} \left( \left(1 - \frac{p_{\text{aa}}}{p_a}\right) \widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

587 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

588 It is left to apply Lemma 3 with

$$\begin{aligned}
\Psi^t &= \frac{(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{1}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2(p_a - p_{\text{aa}})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]
\end{aligned}$$

589 to conclude the proof.  $\square$

590 **Corollary 1.** Let the assumptions from Theorem 3 hold and  $p_{\text{page}} = B/(m+B)$ . Then DASHA-PP-PAGE  
591 needs

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{p_a \sqrt{n}} \left( \widehat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{1}{p_a} \sqrt{\frac{m}{n}} \left( \frac{1_{p_a} \widehat{L}}{\sqrt{B}} + \frac{L_{\max}}{B} \right) \right] \right) \quad (9)$$

592 communication rounds to get an  $\varepsilon$ -solution and the expected number of gradient calculations per  
593 node equals  $\mathcal{O}(m + BT)$ .

594 *Proof.* In the view of Theorem 3, it is enough to do

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \sqrt{\frac{\omega^2}{np_a^2} \left( \widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right)} + \frac{1}{np_a^2 p_{\text{page}}} \left( \left(1 - \frac{p_{\text{aa}}}{p_a}\right) \widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right] \right)$$

595 steps to get  $\varepsilon$ -solution. Using the choice of  $p_{\text{mega}}$  and the definition of  $\mathbb{1}_{p_a}$ , we can get (9).

596 Note that the expected number of gradients calculations at each communication round equals  $p_{\text{mega}}m +$   
597  $(1 - p_{\text{mega}})B = \frac{2mB}{m+B} \leq 2B$ .  $\square$

598 **Corollary 2.** Suppose that assumptions of Corollary 1 hold,  $B \leq \min \left\{ \frac{1}{p_a} \sqrt{\frac{m}{n}}, \frac{L_{\max}^2}{\mathbb{1}_{p_a}^2 L^2} \right\}$ <sup>6</sup>, and we  
599 use the unbiased compressor RandK with  $K = \Theta(B^d/\sqrt{m})$ . Then the communication complexity of

---

<sup>6</sup>If  $\mathbb{1}_{p_a} = 0$ , then  $\frac{L_{\max}^2}{\mathbb{1}_{p_a}^2 L^2} = +\infty$

600 Algorithm 1 is

$$\mathcal{O}\left(d + \frac{L_{\max}\Delta_0 d}{p_a \varepsilon \sqrt{n}}\right), \quad (10)$$

601 and the expected number of gradient calculations per node equals

$$\mathcal{O}\left(m + \frac{L_{\max}\Delta_0 \sqrt{m}}{p_a \varepsilon \sqrt{n}}\right). \quad (11)$$

602 *Proof.* The communication complexity equals

$$\mathcal{O}(d + KT) = \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ KL + K \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + K \frac{1}{p_a} \sqrt{\frac{m}{n}} \left( \frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_{\max}}{B} \right) \right] \right).$$

603 Since  $B \leq \frac{L_{\max}^2}{\mathbb{1}_{p_a}^2 \hat{L}^2}$ , we have  $\frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_{\max}}{B} \leq \frac{2L_{\max}}{B}$  and

$$\mathcal{O}(d + KT) = \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ KL + K \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + K \frac{1}{p_a} \sqrt{\frac{m}{n}} \frac{L_{\max}}{B} \right] \right).$$

604 Note that  $K = \Theta\left(\frac{Bd}{\sqrt{m}}\right) = \mathcal{O}\left(\frac{d}{p_a \sqrt{n}}\right)$  and  $\omega + 1 = \frac{d}{K}$  due to Theorem 6, thus

$$\begin{aligned} \mathcal{O}(d + KT) &= \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ \frac{d}{p_a \sqrt{n}} L + \frac{d}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{d}{p_a \sqrt{n}} L_{\max} \right] \right) \\ &= \mathcal{O}\left(d + \frac{L_{\max}\Delta_0 d}{p_a \varepsilon \sqrt{n}}\right). \end{aligned}$$

605 Using the same reasoning, the expected number of gradient calculations per node equals

$$\begin{aligned} \mathcal{O}(m + BT) &= \mathcal{O}\left(m + \frac{\Delta_0}{\varepsilon} \left[ BL + B \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + B \frac{1}{p_a} \sqrt{\frac{m}{n}} \left( \frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_{\max}}{B} \right) \right] \right) \\ &= \mathcal{O}\left(m + \frac{\Delta_0}{\varepsilon} \left[ BL + B \frac{d}{K p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + B \frac{1}{p_a} \sqrt{\frac{m}{n}} \frac{L_{\max}}{B} \right] \right) \\ &= \mathcal{O}\left(m + \frac{\Delta_0}{\varepsilon} \left[ \frac{1}{p_a} \sqrt{\frac{m}{n}} L + \frac{\sqrt{m}}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{1}{p_a} \sqrt{\frac{m}{n}} L_{\max} \right] \right) \\ &= \mathcal{O}\left(m + \frac{L_{\max}\Delta_0 \sqrt{m}}{p_a \varepsilon \sqrt{n}}\right). \end{aligned}$$

606 □

## 607 E.5 Proof for DASHA-PP-FINITE-MVR

608 **Lemma 9.** Suppose that Assumptions 3, 4, and 8 hold. For  $h_i^{t+1}$ ,  $h_{ij}^{t+1}$  and  $k_i^{t+1}$  from Algorithm 1  
609 (DASHA-PP-FINITE-MVR) we have

1.

$$\begin{aligned} &\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\ &\leq \left( \frac{2L_{\max}^2}{np_a B} + \frac{2(p_a - p_{aa}) \hat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \\ &\quad + \frac{2(p_a - p_{aa}) b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \frac{2b^2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\ &\quad + (1 - b)^2 \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$



2.

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
& \leq \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\
& \quad + \frac{2b^2}{p_a B m} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n].
\end{aligned}$$

3.

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \right] \\
& \leq \frac{2 \left( 1 - \frac{p_a B}{m} \right) L_{\max}^2}{\frac{p_a B}{m}} \|x^{t+1} - x^t\|^2 \\
& \quad + \left( \frac{2 \left( 1 - \frac{p_a B}{m} \right) b^2}{\frac{p_a B}{m}} + (1-b)^2 \right) \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2, \quad \forall i \in [n], \forall j \in [m].
\end{aligned}$$

4.

$$\begin{aligned}
& \mathbb{E}_B \left[ \|k_i^{t+1}\|^2 \right] \\
& \leq \left( \frac{2L_{\max}^2}{B} + 2L_i^2 \right) \|x^{t+1} - x^t\|^2 \\
& \quad + \frac{2b^2}{Bm} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n].
\end{aligned}$$

610 *Proof.* We start by proving the first inequality. Note that

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} [h_i^{t+1}] \right] \\
& = p_a \left( h_i^t + \frac{1}{p_a} \mathbb{E}_B [k_i^{t+1}] \right) + (1-p_a)h_i^t \\
& = h_i^t + \frac{1}{m} \sum_{j=1}^m \frac{B}{m} \cdot \frac{m}{B} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))) + \left( 1 - \frac{B}{m} \right) \cdot 0 \\
& = \nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t)),
\end{aligned}$$

611 thus

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\
& \stackrel{(17)}{=} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \mathbb{E}_B [\mathbb{E}_{p_a} [h^{t+1}]]\|^2 \right] \right] + (1-b)^2 \|h^t - \nabla f(x^t)\|^2.
\end{aligned}$$

612 We can use Lemma 1 with  $r_i = h_i^t$  and  $s_i = k_i^{t+1}$  to obtain

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\
& \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_B \left[ \|k_i^{t+1} - \mathbb{E}_B [k_i^{t+1}]\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\mathbb{E}_B [k_i^{t+1}]\|^2 \\
& \quad + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \\
& = \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_B \left[ \left\| \frac{1}{m} \sum_{j=1}^m k_{ij}^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))) \right\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2.
\end{aligned}$$

613 Next, we again use Lemma 1 with  $r_i = 0$ ,  $s_i = \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))$ ,

614  $p_a = \frac{B}{m}$ , and  $p_{aa} = \frac{B(B-1)}{m(m-1)}$ :

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right] \\
& \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \left( \frac{m-B}{Bm(m-1)} \sum_{j=1}^m \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t)) \right\|^2 \right) \\
& \quad + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& \quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\
& \leq \frac{1}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t)) \right\|^2 \\
& \quad + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& \quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\
& \stackrel{(16)}{\leq} \frac{2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) \right\|^2 + \frac{2b^2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| h_{ij}^t - \nabla f_{ij}(x^t) \right\|^2 \\
& \quad + \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
& \quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2.
\end{aligned}$$

615 Due to Assumptions 3 and 4, we have

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right] \\
& \leq \left( \frac{2L_{\max}^2}{n p_a B} + \frac{2(p_a - p_{aa})\hat{L}^2}{n p_a^2} \right) \left\| x^{t+1} - x^t \right\|^2 \\
& \quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 + \frac{2b^2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \left\| h_{ij}^t - \nabla f_{ij}(x^t) \right\|^2 \\
& \quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2.
\end{aligned}$$

616 Let us get the bound for the second inequality:

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \\
& \stackrel{(17)}{=} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \left\| h_i^{t+1} - (\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t))) \right\|^2 \right] \right] \\
& \quad + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
& = p_a \mathbb{E}_B \left[ \left\| h_i^t + \frac{1}{p_a} h_i^{t+1} - (\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t))) \right\|^2 \right] \\
& \quad + (1-p_a) \left\| h_i^t - (\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t))) \right\|^2 \\
& \quad + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(17)}{=} \frac{1}{p_a} \mathbb{E}_B \left[ \|k_i^{t+1} - \mathbb{E}_B[k_i^{t+1}]\|^2 \right] \\
&\quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

617 Let us use Lemma 1 with  $r_i = 0$ ,  $s_i = \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))$ ,  $p_a = \frac{B}{m}$ , and  
618  $p_{aa} = \frac{B(B-1)}{m(m-1)}$ :

$$\begin{aligned}
&\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
&\leq \frac{1}{p_a} \left( \frac{m-B}{Bm(m-1)} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \right) \\
&\quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\leq \frac{1}{p_a B m} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \\
&\quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\stackrel{(16)}{\leq} \frac{2}{p_a B m} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\|^2 + \frac{2(1-p_a)}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
&\quad + \frac{2b^2}{p_a B m} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\leq \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\
&\quad + \frac{2b^2}{p_a B m} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2,
\end{aligned}$$

619 where we used Assumptions 3 and 4. We continue the proof by considering

$$\begin{aligned}
&\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \right] : \\
&\mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \right] \\
&\stackrel{(17)}{=} \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_{ij}^{t+1} - (\nabla f_{ij}(x^{t+1}) + (1-b)(h_{ij}^t - \nabla f_{ij}(x^t)))\|^2 \right] \right] \\
&\quad + (1-b)^2 \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
&= \frac{p_a B}{m} \mathbb{E}_B \left[ \left\| h_{ij}^t + \frac{m}{B p_a} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))) - (\nabla f_{ij}(x^{t+1}) + (1-b)(h_{ij}^t - \nabla f_{ij}(x^t))) \right\|^2 \right] \\
&\quad + \left( 1 - \frac{p_a B}{m} \right) \|h_{ij}^t - (\nabla f_{ij}(x^{t+1}) + (1-b)(h_{ij}^t - \nabla f_{ij}(x^t)))\|^2 \\
&\quad + (1-b)^2 \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
&= \frac{\left( 1 - \frac{p_a B}{m} \right)^2}{\frac{p_a B}{m}} \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \\
&\quad + \left( 1 - \frac{p_a B}{m} \right) \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2
\end{aligned}$$

$$\begin{aligned}
& + (1-b)^2 \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
& = \frac{\left(1 - \frac{p_a B}{m}\right)}{\frac{p_a B}{m}} \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \\
& + (1-b)^2 \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
& \stackrel{(16)}{\leq} \frac{2\left(1 - \frac{p_a B}{m}\right)}{\frac{p_a B}{m}} \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\|^2 + \left(\frac{2\left(1 - \frac{p_a B}{m}\right)b^2}{\frac{p_a B}{m}} + (1-b)^2\right) \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2.
\end{aligned}$$

621 It is left to consider Assumption 4:

$$\begin{aligned}
& \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \right] \\
& \leq \frac{2\left(1 - \frac{p_a B}{m}\right) L_{\max}^2}{\frac{p_a B}{m}} \|x^{t+1} - x^t\|^2 + \left(\frac{2\left(1 - \frac{p_a B}{m}\right)b^2}{\frac{p_a B}{m}} + (1-b)^2\right) \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2.
\end{aligned}$$

622 Finally, we obtain the bound for the last inequality of the lemma:

$$\begin{aligned}
& \mathbb{E}_B \left[ \|k_i^{t+1}\|^2 \right] \\
& \stackrel{(17)}{=} \mathbb{E}_B \left[ \|k_i^{t+1} - \mathbb{E}_B[k_i^{t+1}]\|^2 \right] \\
& + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2.
\end{aligned}$$

623 Using Lemma 1, we get

$$\begin{aligned}
& \mathbb{E}_B \left[ \|k_i^{t+1}\|^2 \right] \\
& \leq \frac{m-B}{Bm(m-1)} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \\
& + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& \leq \frac{1}{Bm} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b(h_{ij}^t - \nabla f_{ij}(x^t))\|^2 \\
& + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& \stackrel{(16)}{\leq} \frac{2}{Bm} \sum_{j=1}^m \|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\|^2 + 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
& + \frac{2b^2}{Bm} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
& \leq \left( \frac{2L_{\max}^2}{B} + 2L_i^2 \right) \|x^{t+1} - x^t\|^2 \\
& + \frac{2b^2}{Bm} \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2,
\end{aligned}$$

624 where we used Assumptions 3 and 4. □

**Theorem 7.** Suppose that Assumptions 1, 2, 3, 4, 7, and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{\frac{p_a B}{m}}{2 - \frac{p_a B}{m}}$ ,

$$\gamma \leq \left( L + \sqrt{\frac{148\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{L_{\max}^2}{B} \right) + \frac{72m}{np_a^2 B} \left( \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 + \frac{L_{\max}^2}{B} \right)} \right)^{-1},$$

625  $g_i^0 = h_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  and  $h_{ij}^0 = \nabla f_{ij}(x^0)$  for all  $i \in [n], j \in [m]$  in Algorithm 1  
 626 (DASHA-PP-FINITE-MVR) then  $\mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{2\Delta_0}{\gamma T}$ .

627 *Proof.* Let us fix constants  $\nu, \rho, \delta \in [0, \infty)$  that we will define later. Considering Lemma 6, Lemma 9,  
 628 and the law of total expectation, we obtain

$$\begin{aligned}
 & \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
 & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
 & + \delta \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \\
 & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
 & + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
 & + \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \\
 & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
 & + \delta \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \\
 & = \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
 & + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
 & + \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \mathbb{E} \left[ \mathbb{E}_B \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \right] \\
 & + \nu \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \right] \right] \\
 & + \rho \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\
 & + \delta \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \right] \right] \\
 & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
 & + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
 & + \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \mathbb{E} \left[ \left( \frac{2L_{\max}^2}{B} + 2\hat{L}^2 \right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{Bmn} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \frac{2b^2}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]
 \end{aligned}$$

$$\begin{aligned}
& + \nu \mathbb{E} \left( \left( \frac{2L_{\max}^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \frac{2b^2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
& \quad \left. + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \right) \\
& + \rho \mathbb{E} \left( \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad + \frac{2b^2}{p_a B n m} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \Bigg) \\
& + \delta \mathbb{E} \left( \frac{2 \left(1 - \frac{p_a B}{m}\right) L_{\max}^2}{\frac{p_a B}{m}} \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \left( \frac{2 \left(1 - \frac{p_a B}{m}\right) b^2}{\frac{p_a B}{m}} + (1-b)^2 \right) \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right).
\end{aligned}$$

Due to  $b = \frac{\frac{p_a B}{m}}{2 - \frac{p_a B}{m}} \leq \frac{p_a}{2 - p_a}$ , we have

$$\left( \frac{2 \left(1 - \frac{p_a B}{m}\right) b^2}{\frac{p_a B}{m}} + (1-b)^2 \right) \leq 1 - b$$

and

$$\left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \leq 1 - b.$$

629 Moreover, we consider that  $1 - \frac{p_a B}{m} \leq 1$ , therefore

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& + \delta \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{4\gamma\omega(2\omega + 1)}{np_a^2} \mathbb{E} \left[ \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{Bmn} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + \frac{2b^2}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \nu \mathbb{E} \left( \left( \frac{2L_{\max}^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right.
\end{aligned}$$

$$\begin{aligned}
& + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \frac{2b^2}{n^2 p_a B m} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \\
& + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \Big) \\
& + \rho \mathbb{E} \left( \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \left. + \frac{2b^2}{p_a B n m} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 + (1-b) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right) \\
& + \delta \mathbb{E} \left( \frac{2mL_{\max}^2}{p_a B} \|x^{t+1} - x^t\|^2 + (1-b) \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right).
\end{aligned}$$

630 After rearranging the terms, we get

$$\begin{aligned}
& \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
& + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
& + \delta \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2\right] \\
& \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) \right. \\
& \quad \left. - \nu \left( \frac{2L_{\max}^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) - \delta \frac{2mL_{\max}^2}{p_a B} \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
& + (\gamma + \nu(1-b)^2) \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\nu(p_a - p_{aa})b^2}{np_a^2} + \rho(1-b) \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{2\nu b^2}{np_a B} + \frac{2\rho b^2}{p_a B} + \delta(1-b) \right) \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2\right].
\end{aligned}$$

631 Thus, if we take  $\nu = \frac{\gamma}{b}$ , then  $\gamma + \nu(1-b)^2 \leq \nu$  and

$$\begin{aligned}
& \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
& + \delta \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\hat{L}^2 \right) \right. \\
&\quad \quad \left. - \left( \frac{2\gamma L_{\max}^2}{bn p_a B} + \frac{2\gamma(p_a - p_{aa})\hat{L}^2}{bn p_a^2} \right) - \rho \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\hat{L}^2}{p_a} \right) - \delta \frac{2mL_{\max}^2}{p_a B} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{2\gamma b}{np_a B} + \frac{2\rho b^2}{p_a B} + \delta(1-b) \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right].
\end{aligned}$$

Next, if we take  $\rho = \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}$ , then

$$\left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho(1-b) \right) = \rho,$$

632 therefore

$$\begin{aligned}
&\mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
&\quad + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
&\quad + \delta \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \\
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\hat{L}^2 \right) \right. \\
&\quad \quad - \left( \frac{2\gamma L_{\max}^2}{bn p_a B} + \frac{2\gamma(p_a - p_{aa})\hat{L}^2}{bn p_a^2} \right) - \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\hat{L}^2}{p_a} \right) \\
&\quad \quad \left. - \delta \frac{2mL_{\max}^2}{p_a B} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{2\gamma b}{np_a B} + \frac{16b^3\gamma\omega(2\omega+1)}{np_a^3 B} + \frac{4b^2\gamma(p_a - p_{aa})}{nBp_a^3} + \delta(1-b) \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right].
\end{aligned}$$



633 Due to  $b \leq p_a$  and  $\frac{p_a - p_{aa}}{p_a} \leq 1$ , we have

$$\begin{aligned} & \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2B} + \frac{2\gamma b}{np_aB} + \frac{16b^3\gamma\omega(2\omega+1)}{np_a^3B} + \frac{4b^2\gamma(p_a - p_{aa})}{nBp_a^3} \\ & \leq \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2B} + \frac{2\gamma b}{np_aB} + \frac{16b^2\gamma\omega(2\omega+1)}{np_a^2B} + \frac{4\gamma b}{np_aB} \\ & = \frac{24b^2\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma b}{np_aB}. \end{aligned}$$

634 Let us take  $\delta = \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB}$ . Thus

$$\left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2B} + \frac{2\gamma b}{np_aB} + \frac{16b^3\gamma\omega(2\omega+1)}{np_a^3B} + \frac{4b^2\gamma(p_a - p_{aa})}{nBp_a^3} + \delta(1-b) \right) \leq \delta$$

635 and

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\hat{L}^2 \right) \right. \\ & \quad \left. - \left( \frac{2\gamma L_{\max}^2}{bnp_aB} + \frac{2\gamma(p_a - p_{aa})\hat{L}^2}{bnp_a^2} \right) - \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \left( \frac{2L_{\max}^2}{p_aB} + \frac{2(1-p_a)\hat{L}^2}{p_a} \right) \right. \\ & \quad \left. - \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \frac{2mL_{\max}^2}{p_aB} \right] \mathbb{E}[\|x^{t+1} - x^t\|^2] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] \\ & + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2\right]. \end{aligned}$$

636 Let us simplify the term near  $\mathbb{E}[\|x^{t+1} - x^t\|^2]$ . Due to  $b \leq p_a$ ,  $\frac{p_a - p_{aa}}{p_a} \leq 1$ , and  $1 - p_a \leq 1$ , we

637 have

$$\begin{aligned} & \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\hat{L}^2 \right) \\ & + \left( \frac{2\gamma L_{\max}^2}{bnp_aB} + \frac{2\gamma(p_a - p_{aa})\hat{L}^2}{bnp_a^2} \right) \end{aligned}$$

$$\begin{aligned}
& + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \\
& + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{6\gamma}{np_a B} \right) \frac{2mL_{\max}^2}{p_a B} \\
& \leq \frac{12\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) \\
& + \left( \frac{6\gamma L_{\max}^2}{bnp_a B} + \frac{6\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} \right) \\
& + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{6\gamma}{np_a B} \right) \frac{2mL_{\max}^2}{p_a B}
\end{aligned}$$

638 Considering that  $b \leq \frac{p_a B}{m}$  and  $b \geq \frac{p_a B}{2m}$ , we obtain

$$\begin{aligned}
& \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) \\
& + \left( \frac{2\gamma L_{\max}^2}{bnp_a B} + \frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} \right) \\
& + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \left( \frac{2L_{\max}^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \\
& + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{6\gamma}{np_a B} \right) \frac{2mL_{\max}^2}{p_a B} \\
& \leq \frac{36\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) + \left( \frac{18\gamma L_{\max}^2}{bnp_a B} + \frac{6\gamma(p_a - p_{aa})\widehat{L}^2}{bnp_a^2} \right) \\
& \leq \frac{36\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) + \left( \frac{36m\gamma L_{\max}^2}{np_a^2 B^2} + \frac{12m\gamma(p_a - p_{aa})\widehat{L}^2}{Bnp_a^3} \right).
\end{aligned}$$

639 All in all, we have

$$\begin{aligned}
& \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
& + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2 B} + \frac{6\gamma}{np_a B} \right) \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2\right] \\
& \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{36\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2L_{\max}^2}{B} + 2\widehat{L}^2 \right) - \left( \frac{36m\gamma L_{\max}^2}{np_a^2 B^2} + \frac{12m\gamma(p_a - p_{aa})\widehat{L}^2}{Bnp_a^3} \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right]
\end{aligned}$$

$$+ \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right].$$

640 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\ & + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\ & + \left( \frac{24b\gamma\omega(2\omega+1)}{np_a^2B} + \frac{6\gamma}{np_aB} \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right]. \end{aligned}$$

641 It is left to apply Lemma 3 with

$$\begin{aligned} \Psi^t &= \frac{(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ &+ \frac{1}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\ &+ \left( \frac{8b\omega(2\omega+1)}{np_a^2} + \frac{2(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\ &+ \left( \frac{24b\omega(2\omega+1)}{np_a^2B} + \frac{6}{np_aB} \right) \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|h_{ij}^t - \nabla f_{ij}(x^t)\|^2 \right] \end{aligned}$$

642 to conclude the proof.  $\square$

## 643 E.6 Proof for DASHA-PP-MVR

644 Let us denote  $\nabla f_i(x^{t+1}; \xi_i^{t+1}) := \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1})$ .

645 **Lemma 10.** Suppose that Assumptions 3, 5, 6 and 8 hold. For  $h_i^{t+1}$  and  $k_i^{t+1}$  from Algorithm 1  
646 (DASHA-PP-MVR) we have

1.

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \right] \\ & \leq \frac{2b^2\sigma^2}{np_aB} + \left( \frac{2(1-b)^2L_\sigma^2}{np_aB} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \frac{2(p_a - p_{aa})b^2}{n^2p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

2.

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\ & \leq \frac{2b^2\sigma^2}{p_a B} + \left( \frac{2(1-b)^2 L_\sigma^2}{p_a B} + \frac{2(1-p_a) L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n]. \end{aligned}$$

3.

$$\mathbb{E}_k \left[ \|k_i^{t+1}\|^2 \right] \leq \frac{2b^2\sigma^2}{B} + \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2L_i^2 \right) \|x^{t+1} - x^t\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n].$$

647 *Proof.* First, let us proof the bound for  $\mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right]$ :

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\ & = \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \mathbb{E}_k [\mathbb{E}_{p_a} [h^{t+1}]]\|^2 \right] \right] + \|\mathbb{E}_k [\mathbb{E}_{p_a} [h^{t+1}]] - \nabla f(x^{t+1})\|^2. \end{aligned}$$

648 Using

$$\mathbb{E}_k [\mathbb{E}_{p_a} [h_i^{t+1}]] = h_i^t + \mathbb{E}_k [k_i^{t+1}] = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))$$

649 and (17), we have

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\ & = \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \mathbb{E}_k [\mathbb{E}_{p_a} [h^{t+1}]]\|^2 \right] \right] + (1-b)^2 \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

650 We can use Lemma 1 with  $r_i = h_i^t$  and  $s_i = k_i^{t+1}$  to obtain

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \\ & \leq \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \|k_i^{t+1} - \mathbb{E}_k [k_i^{t+1}]\|^2 \right] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\mathbb{E}_k [k_i^{t+1}]\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \\ & = \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})) \right. \\ & \quad \left. - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \\ & \quad + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\ & \quad + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \\ & \stackrel{(16)}{\leq} \frac{2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \|b(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1}))\|^2 \right] \\ & \quad + \frac{2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \|(1-b)(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)))\|^2 \right] \\ & \quad + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\ & \quad + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \\ & = \frac{2b^2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1})\|^2 \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{2(1-b)^2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2. \\
& = \frac{2b^2}{n^2 p_a B^2} \sum_{i=1}^n \sum_{j=1}^B \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^{t+1}) \right\|^2 \right] \\
& + \frac{2(1-b)^2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2.
\end{aligned}$$

651 In the last equality, we use the independence of elements in the mini-batches. Due to Assumption 5,  
652 we get

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right] \\
& \leq \frac{2b^2 \sigma^2}{np_a B} \\
& + \frac{2(1-b)^2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\
& \stackrel{(16)}{\leq} \frac{2b^2 \sigma^2}{np_a B} \\
& + \frac{2(1-b)^2}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2. \\
& = \frac{2b^2 \sigma^2}{np_a B} \\
& + \frac{2(1-b)^2}{n^2 p_a B^2} \sum_{i=1}^n \sum_{j=1}^B \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^t; \xi_{ij}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& + \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
& + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2,
\end{aligned}$$

653 where we use the independence of elements in the mini-batches. Using Assumptions 3 and 6, we  
654 obtain

$$\mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right]$$

$$\begin{aligned}
&\leq \frac{2b^2\sigma^2}{np_a B} + \left( \frac{2(1-b)^2 L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa}) \widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \\
&\quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2.
\end{aligned}$$

655 Now, we prove the second inequality:

$$\begin{aligned}
&\mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
&= \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \mathbb{E}_k [\mathbb{E}_{p_a} [h_i^{t+1}]]\|^2 \right] \right] \\
&\quad + \|\mathbb{E}_k [\mathbb{E}_{p_a} [h_i^{t+1}]] - \nabla f_i(x^{t+1})\|^2 \\
&= \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \right] \\
&\quad + \|h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) - \nabla f_i(x^{t+1})\|^2 \\
&= \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \right] \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&= p_a \mathbb{E}_k \left[ \left\| h_i^t + \frac{1}{p_a} k_i^{t+1} - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))) \right\|^2 \right] \\
&\quad + (1-p_a) \|h_i^t - (h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&= p_a \mathbb{E}_k \left[ \left\| \frac{1}{p_a} k_i^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))) \right\|^2 \right] \\
&\quad + (1-p_a) \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\stackrel{(17)}{=} \frac{1}{p_a} \mathbb{E}_k \left[ \|k_i^{t+1} - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \\
&\quad + \frac{(1-p_a)^2}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-p_a) \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&= \frac{1}{p_a} \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \\
&\quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&= \frac{1}{p_a} \mathbb{E}_k \left[ \|b(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1})) + (1-b)(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)))\|^2 \right] \\
&\quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
&\quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2 \\
&\stackrel{(16)}{\leq} \frac{2b^2}{p_a} \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1})\|^2 \right] \\
&\quad + \frac{2(1-b)^2}{p_a} \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

656 Considering the independence of elements in the mini-batch, we obtain

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
& = \frac{2b^2}{p_a B^2} \sum_{j=1}^B \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^{t+1})\|^2 \right] \\
& \quad + \frac{2(1-b)^2}{p_a B^2} \sum_{j=1}^B \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^t; \xi_{ij}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right] \\
& \quad + \frac{1-p_a}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& \quad + (1-b)^2 \|h_i^t - \nabla f_i(x^t)\|^2. \\
& \stackrel{(16)}{\leq} \frac{2b^2}{p_a B^2} \sum_{j=1}^B \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^{t+1})\|^2 \right] \\
& \quad + \frac{2(1-b)^2}{p_a B^2} \sum_{j=1}^B \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^t; \xi_{ij}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right] \\
& \quad + \frac{2(1-p_a)}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2
\end{aligned}$$

657 Next, we use Assumptions 3, 6, 5, to get

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\
& \leq \frac{2b^2\sigma^2}{p_a B} + \left( \frac{2(1-b)^2 L_\sigma^2}{p_a B} + \frac{2(1-p_a)L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\
& \quad + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

658 It is left to prove the bound for  $\mathbb{E}_k \left[ \|k_i^{t+1}\|^2 \right]$ :

$$\begin{aligned}
& \mathbb{E}_k \left[ \|k_i^{t+1}\|^2 \right] \\
& = \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t; \xi_i^{t+1}))\|^2 \right] \\
& \stackrel{(17)}{=} \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)))\|^2 \right] \\
& \quad + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& = \mathbb{E}_k \left[ \|b(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1})) + (1-b)(\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)))\|^2 \right] \\
& \quad + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\|^2 \\
& \stackrel{(16)}{\leq} 2b^2 \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^{t+1})\|^2 \right] \\
& \quad + 2(1-b)^2 \mathbb{E}_k \left[ \|\nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))\|^2 \right] \\
& \quad + 2 \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

Using Assumptions 3, 6, 5 and the independence of elements in the mini-batch, we get

$$\begin{aligned} & \mathbb{E}_k \left[ \|k_i^{t+1}\|^2 \right] \\ & \leq \frac{2b^2\sigma^2}{B} + \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2L_i^2 \right) \|x^{t+1} - x^t\|^2 + 2b^2 \|h_i^t - \nabla f_i(x^t)\|^2. \end{aligned}$$

660

□

**Theorem 4.** Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 $b \in \left(0, \frac{p_a}{2-p_a}\right]$ ,

$$\gamma \leq \left( L + \left[ \frac{48\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1-b)^2L_\sigma^2}{B} \right) + \frac{12}{np_ab} \left( \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 + \frac{(1-b)^2L_\sigma^2}{B} \right) \right]^{1/2} \right)^{-1},$$

and  $g_i^0 = h_i^0$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-MVR). Then

$$\begin{aligned} \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] & \leq \frac{1}{T} \left[ \frac{2\Delta_0}{\gamma} + \frac{2}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{32b\omega(2\omega+1)}{np_a^2} + \frac{4\left(1 - \frac{p_{aa}}{p_a}\right)}{np_a} \right) \left( \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \right] \\ & + \left( \frac{48b^2\omega(2\omega+1)}{p_a^2} + \frac{12b}{p_a} \right) \frac{\sigma^2}{nB}. \end{aligned}$$

*Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 6, Lemma 10, and the law of total expectation, we obtain

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & = \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \mathbb{E}_k \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \right] \\ & + \nu \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \right] \right] \\ & + \rho \mathbb{E} \left[ \mathbb{E}_B \left[ \mathbb{E}_{p_a} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \end{aligned}$$



$$\begin{aligned}
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{4\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \frac{2b^2\sigma^2}{B} + \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\hat{L}^2 \right) \|x^{t+1} - x^t\|^2 + 2b^2 \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \nu \mathbb{E} \left( \frac{2b^2\sigma^2}{np_a B} + \left( \frac{2(1-b)^2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \right) \\
& + \rho \mathbb{E} \left( \frac{2b^2\sigma^2}{p_a B} + \left( \frac{2(1-b)^2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\hat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right).
\end{aligned}$$

666 After rearranging the terms, we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\hat{L}^2 \right) \right. \\
& \quad \left. - \nu \left( \frac{2(1-b)^2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} \right) - \rho \left( \frac{2(1-b)^2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\hat{L}^2}{p_a} \right) \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + (\gamma + \nu(1-b)^2) \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\nu(p_a - p_{aa})b^2}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \nu \frac{2b^2}{np_a} + \rho \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

667 By taking  $\nu = \frac{\gamma}{b}$ , one can show that  $(\gamma + \nu(1-b)^2) \leq \nu$ , and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right. \\
& \quad \left. - \frac{\gamma}{b} \left( \frac{2(1-b)^2 L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa}) \widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2(1-b)^2 L_\sigma^2}{p_a B} + \frac{2(1-p_a) \widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b}{np_a} + \rho \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

668 Note that  $b \leq \frac{p_a}{2-p_a}$ , thus

$$\begin{aligned}
& \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \\
& \leq \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho(1-b) \right).
\end{aligned}$$

669 And if we take  $\rho = \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}$ , then

$$\left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})b}{np_a^2} + \rho(1-b) \right) \leq \rho,$$

670 and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right. \\
& \quad \left. - \frac{\gamma}{np_a b} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2 \left( 1 - \frac{p_{aa}}{p_a} \right) \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^3} + \frac{2\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2} \right) \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2(1-p_a) \widehat{L}^2 \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b}{np_a} + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

671 Let us simplify the inequality. First, due to  $b \leq p_a$  and  $(1-p_a) \leq \left( 1 - \frac{p_{aa}}{p_a} \right)$ , we have

$$\left( \frac{8b\gamma\omega(2\omega+1)}{np_a^3} + \frac{2\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2} \right) \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2(1-p_a) \widehat{L}^2 \right)$$

$$\begin{aligned}
&= \frac{8b\gamma\omega(2\omega+1)}{np_a^3} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2(1-p_a)\widehat{L}^2 \right) \\
&\quad + \frac{2\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2(1-p_a)\widehat{L}^2 \right) \\
&\leq \frac{8\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2\widehat{L}^2 \right) \\
&\quad + \frac{2\gamma}{np_a b} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2 \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right),
\end{aligned}$$

672 therefore

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{12\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right. \\
&\quad \quad \left. - \frac{3\gamma}{np_a b} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2 \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
&\quad + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b}{np_a} + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B} \\
&= \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \quad \left. - \frac{6\gamma}{np_a b} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
&\quad + \left( \frac{8b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma b}{np_a} + \left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

673 Also, we can simplify the last term:

$$\begin{aligned}
&\left( \frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \\
&= \frac{16b^3\gamma\omega(2\omega+1)}{np_a^3} + \frac{4b^2\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2}
\end{aligned}$$

$$\leq \frac{16b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{4b\gamma}{np_a},$$

674 thus

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma\omega(2\omega+1)}{np_a^2} \left(\frac{(1-b)^2 L_\sigma^2}{B} + \hat{L}^2\right)\right. \\ & \quad \left. - \frac{6\gamma}{np_a b} \left(\frac{(1-b)^2 L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2\right)\right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \left(\frac{24b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{6\gamma b}{np_a}\right) \frac{\sigma^2}{B}. \end{aligned}$$

675 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \left(\frac{24b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{6\gamma b}{np_a}\right) \frac{\sigma^2}{B}. \end{aligned}$$

676 It is left to apply Lemma 3 with

$$\begin{aligned} \Psi^t &= \frac{(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ &+ \frac{1}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{8b\omega(2\omega+1)}{np_a^2} + \frac{2(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \end{aligned}$$

677 and  $C = \left(\frac{24b^2\omega(2\omega+1)}{p_a^2} + \frac{6b}{p_a}\right) \frac{\sigma^2}{nB}$  to conclude the proof.  $\square$

678 **Corollary 3.** Suppose that assumptions from Theorem 4 hold, momentum  $b =$   
679  $\Theta\left(\min\left\{\frac{p_a}{\omega} \sqrt{\frac{n\varepsilon B}{\sigma^2}}, \frac{p_a n \varepsilon B}{\sigma^2}\right\}\right)$ ,  $\frac{\sigma^2}{n\varepsilon B} \geq 1$ , and  $h_i^0 = \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0)$  for all  $i \in [n]$ ,

680 and batch size  $B_{\text{init}} = \Theta\left(\frac{\sqrt{p_a B}}{b}\right)$ , then Algorithm 1 (DASHA-PP-MVR) needs

$$T := \mathcal{O}\left(\frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon n}} \left( \frac{\mathbb{1}_{p_a} \hat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right] + \frac{\sigma^2}{\sqrt{p_a n \varepsilon B}} \right)$$

681 communication rounds to get an  $\varepsilon$ -solution and the number of stochastic gradient calculations per  
682 node equals  $\mathcal{O}(B_{\text{init}} + BT)$ .

683 *Proof.* Using the result from Theorem 4, we have

$$\begin{aligned} & \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \\ & \leq \frac{1}{T} \left[ 2\Delta_0 \left( L + \sqrt{\frac{48\omega(2\omega+1)}{np_a^2}} \left( \hat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B} \right) + \frac{12}{np_a b} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B} \right) \right) \right. \\ & \quad \left. + \frac{2}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{32b\omega(2\omega+1)}{np_a^2} + \frac{4\left(1 - \frac{p_{aa}}{p_a}\right)}{np_a} \right) \left( \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \right] \\ & \quad + \left( \frac{48b^2\omega(2\omega+1)}{p_a^2} + \frac{12b}{p_a} \right) \frac{\sigma^2}{nB} \end{aligned}$$

684 We choose  $b$  to ensure  $\left( \frac{48b^2\omega(2\omega+1)}{p_a^2} + \frac{12b}{p_a} \right) \frac{\sigma^2}{nB} = \Theta(\varepsilon)$ . Note that  $\frac{1}{b} =$

685  $\Theta\left(\max\left\{\frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{n\varepsilon B}}, \frac{\sigma^2}{p_a n \varepsilon B}\right\}\right) \leq \Theta\left(\max\left\{\frac{\omega^2}{p_a}, \frac{\sigma^2}{p_a n \varepsilon B}\right\}\right)$ , thus

$$\begin{aligned} & \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \\ & = \mathcal{O} \left( \frac{1}{T} \left[ \Delta_0 \left( L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \right. \right. \\ & \quad \left. \left. + \frac{1}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{b\omega^2}{np_a^2} + \frac{1}{np_a} \right) \left( \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \right] + \varepsilon \right), \end{aligned}$$

686 where  $\mathbb{1}_{p_a} = \sqrt{1 - \frac{p_{aa}}{p_a}}$ . It enough to take the following  $T$  to get  $\varepsilon$ -solution.

$$\begin{aligned} T = \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \right. \right. \\ \left. \left. + \frac{1}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{b\omega^2}{np_a^2} + \frac{1}{np_a} \right) \left( \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \right] \right). \end{aligned}$$

687 Let us bound the norms:

$$\begin{aligned} \mathbb{E} \left[ \|h^0 - \nabla f(x^0)\|^2 \right] &= \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0) - \nabla f(x^0) \right\|^2 \right] \\ &= \frac{1}{n^2 B_{\text{init}}^2} \sum_{i=1}^n \sum_{k=1}^{B_{\text{init}}} \mathbb{E} \left[ \|\nabla f_i(x^0; \xi_{ik}^0) - \nabla f_i(x^0)\|^2 \right] \end{aligned}$$

$$\leq \frac{\sigma^2}{nB_{\text{init}}}.$$

688 Using the same reasoning, one can get  $\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \|h_i^0 - \nabla f_i(x^0)\|^2 \right] \leq \frac{\sigma^2}{B_{\text{init}}}$ . Combining all inequalities, we have

$$T = \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \right. \right. \\ \left. \left. + \frac{\sigma^2}{bnB_{\text{init}}} + \frac{b\omega^2\sigma^2}{np_a^2 B_{\text{init}}} + \frac{\sigma^2}{np_a B_{\text{init}}} \right] \right).$$

690 Using the choice of  $B_{\text{init}}$  and  $b$ , we obtain

$$T = \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \right. \right. \\ \left. \left. + \frac{\sigma^2}{\sqrt{p_a} n B} + \frac{b^2 \omega^2 \sigma^2}{np_a^{5/2} B} + \frac{b\sigma^2}{p_a^{3/2} n B} \right] \right) \\ = \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \right. \right. \\ \left. \left. + \frac{\sigma^2}{\sqrt{p_a} n B} + \frac{\varepsilon}{\sqrt{p_a}} \right] \right) \\ = \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right] + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} + \frac{1}{\sqrt{p_a}} \right).$$

691 Using  $\frac{\sigma^2}{n\varepsilon B} \geq 1$ , we can conclude the proof of the inequality. The number of stochastic gradients that  
692 each node calculates equals  $B_{\text{init}} + 2BT = \mathcal{O}(B_{\text{init}} + BT)$ .  $\square$

693 **Corollary 4.** Suppose that assumptions of Corollary 3 hold, batch size  $B \leq \min \left\{ \frac{\sigma}{p_a \sqrt{\varepsilon n}}, \frac{L_\sigma^2}{\mathbb{1}_{p_a}^2 \hat{L}^2} \right\}$ ,  
694 we take RandK compressors with  $K = \Theta \left( \frac{Bd\sqrt{\varepsilon n}}{\sigma} \right)$ . Then the communication complexity equals

$$\mathcal{O} \left( \frac{d\sigma}{\sqrt{p_a} \sqrt{n\varepsilon}} + \frac{L_\sigma \Delta_0 d}{p_a \sqrt{n\varepsilon}} \right), \quad (12)$$

695 and the expected number of stochastic gradient calculations per node equals

$$\mathcal{O} \left( \frac{\sigma^2}{\sqrt{p_a} n \varepsilon} + \frac{L_\sigma \Delta_0 \sigma}{p_a \varepsilon^{3/2} n} \right). \quad (13)$$

696 *Proof.* The communication complexity equals

$$\mathcal{O}(d + KT) = \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ KL + K \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + K \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right] + K \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right).$$

697 Due to  $B \leq \frac{L_\sigma^2}{\mathbb{1}_{p_a}^2 \hat{L}^2}$ , we have  $\mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \leq \frac{2L_\sigma}{\sqrt{B}}$  and

$$\mathcal{O}(d + KT) = \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ KL + K \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + K \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \frac{L_\sigma}{\sqrt{B}} \right] + K \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right).$$

698 From Theorem 6, we have  $\omega + 1 = \frac{d}{K}$ . Since  $K = \Theta\left(\frac{Bd\sqrt{\varepsilon n}}{\sigma}\right) = \mathcal{O}\left(\frac{d}{p_a \sqrt{n}}\right)$ , the communication  
699 complexity equals

$$\begin{aligned} \mathcal{O}(d + KT) &= \mathcal{O}\left(d + \frac{\Delta_0}{\varepsilon} \left[ \frac{d}{p_a \sqrt{n}} L + \frac{d}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{d}{p_a \sqrt{n}} L_\sigma \right] + \frac{d\sigma}{\sqrt{p_a} \sqrt{n} \varepsilon} \right) \\ &= \mathcal{O}\left(\frac{d\sigma}{\sqrt{p_a} \sqrt{n} \varepsilon} + \frac{L_\sigma \Delta_0 d}{p_a \sqrt{n} \varepsilon} \right) \end{aligned}$$

700 And the expected number of stochastic gradient calculations per node equals

$$\begin{aligned} &\mathcal{O}(B_{\text{init}} + BT) \\ &= \mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a} n \varepsilon} + \frac{B\omega}{\sqrt{p_a}} \sqrt{\frac{\sigma^2}{n \varepsilon B}} + \frac{\Delta_0}{\varepsilon} \left[ BL + B \frac{\omega}{p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + B \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \left( \mathbb{1}_{p_a} \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right] + B \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right) \\ &= \mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a} n \varepsilon} + \frac{Bd}{K \sqrt{p_a}} \sqrt{\frac{\sigma^2}{n \varepsilon B}} + \frac{\Delta_0}{\varepsilon} \left[ BL + B \frac{d}{K p_a \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + B \sqrt{\frac{\sigma^2}{p_a^2 \varepsilon n^2 B}} \frac{L_\sigma}{\sqrt{B}} \right] + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon} \right) \\ &= \mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a} n \varepsilon} + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon \sqrt{B}} + \frac{\Delta_0}{\varepsilon} \left[ \frac{\sigma}{p_a \sqrt{\varepsilon n}} L + \frac{\sigma}{p_a \sqrt{\varepsilon n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon n}} L_\sigma \right] \right) \\ &= \mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a} n \varepsilon} + \frac{L_\sigma \Delta_0 \sigma}{p_a \varepsilon^{3/2} n} \right). \end{aligned}$$

701

□

## F Analysis of DASHA-PP under Polyak-Łojasiewicz Condition

In this section, we provide the theoretical convergence rates of DASHA-PP under Polyak-Łojasiewicz Condition.

**Assumption 9.** The function  $f$  satisfy (Polyak-Łojasiewicz) PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*), \quad \forall x \in \mathbb{R}, \quad (31)$$

where  $f^* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$ .

Under Polyak-Łojasiewicz condition, a (random) point  $\hat{x}$  is  $\varepsilon$ -solution, if  $\mathbb{E}[f(\hat{x})] - f^* \leq \varepsilon$ .

We now provide the convergence rates of DASHA-PP under PL-condition.

### F.1 Gradient Setting

**Theorem 8.** Suppose that Assumption 1, 2, 3, 7, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_a}{2-p_a}$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega+1)}{np_a^2} + \frac{48}{np_a^2} \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}} \right)^{-1}, \frac{a}{4\mu} \right\},$$

and  $h_i^0 = g_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP), then  $\mathbb{E}[f(x^T)] - f^* \leq (1 - \gamma\mu)^T \Delta_0$ .

Let us provide bounds up to logarithmic factors and use  $\tilde{\mathcal{O}}(\cdot)$  notation. The provided theorem states that to get  $\varepsilon$ -solution DASHA-PP have to run

$$\tilde{\mathcal{O}} \left( \frac{\omega+1}{p_a} + \frac{L}{\mu} + \frac{\omega\hat{L}}{p_a\mu\sqrt{n}} + \frac{\hat{L}}{p_a\mu\sqrt{n}} \right),$$

communication rounds. The method DASHA from (Tyurin and Richtárik, 2023), have to run

$$\tilde{\mathcal{O}} \left( \omega + \frac{L}{\mu} + \frac{\omega\hat{L}}{\mu\sqrt{n}} \right),$$

communication rounds to get  $\varepsilon$ -solution. The difference is the same as in the general nonconvex case (see Section 6.1). Up to Lipschitz constants factors, we get the degeneration up to  $1/p_a$  factor due to the partial participation.

### F.2 Finite-Sum Setting

**Theorem 9.** Suppose that Assumption 1, 2, 3, 7, 4, 8, and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ , probability  $p_{page} = \frac{B}{m+B}$ ,  $b = \frac{p_{page}p_a}{2-p_a}$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1-p_{page})L_{\max}^2}{B} \right) + \frac{48}{np_a^2 p_{page}} \left( \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 + \frac{(1-p_{page})L_{\max}^2}{B} \right)} \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

and  $h_i^0 = g_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-PAGE), then  $\mathbb{E}[f(x^T)] - f^* \leq (1 - \gamma\mu)^T \Delta_0$ .

The provided theorem states that to get  $\varepsilon$ -solution DASHA-PP have to run

$$\tilde{\mathcal{O}} \left( \frac{\omega+1}{p_a} + \frac{m}{p_a B} + \frac{L}{\mu} + \frac{\omega}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{\sqrt{m}}{p_a\mu\sqrt{n}B} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) \right),$$

communication rounds. The method DASHA-PAGE from (Tyurin and Richtárik, 2023), have to run

$$\tilde{\mathcal{O}} \left( \omega + \frac{m}{B} + \frac{L}{\mu} + \frac{\omega}{\mu\sqrt{n}} \left( \hat{L} + \frac{L_{\max}}{\sqrt{B}} \right) + \frac{\sqrt{m}}{\mu\sqrt{n}B} \left( \frac{L_{\max}}{\sqrt{B}} \right) \right),$$

communication rounds to get  $\varepsilon$ -solution. We can guarantee the degeneration up to  $1/p_a$  factor due to the partial participation only if  $B = \mathcal{O}\left(\frac{L_{\max}^2}{L^2}\right)$ . The same conclusion we have in Section 6.2.



### 725 F.3 Stochastic Setting

**Theorem 10.** Suppose that Assumption 1, 2, 3, 7, 5, 6, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 $b \in \left(0, \frac{p_a}{2-p_a}\right]$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega+1)}{np_a^2} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \hat{L}^2 \right)} + \frac{40}{np_a b} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 \right) \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

726 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-MVR), then

$$\begin{aligned} & \mathbb{E} [f(x^T) - f^*] \\ & \leq (1 - \gamma\mu)^T \left( \Delta_0 + \frac{2\gamma}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{40\gamma b\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \\ & \quad + \frac{1}{\mu} \left( \frac{100b^2\omega(2\omega+1)}{p_a^2} + \frac{20b}{p_a} \right) \frac{\sigma^2}{nB}. \end{aligned}$$

727 The provided theorems states that to get  $\varepsilon$ -solution DASHA-PP have to run

$$\tilde{\mathcal{O}} \left( \underbrace{\frac{\omega+1}{p_a} + \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n \varepsilon B}}}_{\mathcal{P}_2} + \frac{\sigma^2}{p_a \mu n \varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_a \mu \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \underbrace{\frac{\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon B}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right)}_{\mathcal{P}_1} \right) \quad (32)$$

728 communication rounds. We take  $b = \Theta \left( \min \left\{ \frac{p_a}{\omega} \sqrt{\frac{\mu n \varepsilon B}{\sigma^2}}, \frac{p_a \mu n \varepsilon B}{\sigma^2} \right\} \right) \geq$   
 729  $\Theta \left( \min \left\{ \frac{p_a}{\omega^2}, \frac{p_a \mu n \varepsilon B}{\sigma^2} \right\} \right).$

730 The method DASHA-SYNC-MVR from (Tyurin and Richtárik, 2023), have to run

$$\tilde{\mathcal{O}} \left( \omega + \frac{\sigma^2}{\mu n \varepsilon B} + \frac{L}{\mu} + \frac{\omega}{\mu \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{n \mu^{3/2} \sqrt{\varepsilon B}} \left( \frac{L_\sigma}{\sqrt{B}} \right) \right) \quad (33)$$

731 communication rounds to get  $\varepsilon$ -solution<sup>7</sup>.

732 In the stochastic setting, the comparison is a little bit more complicated. As in the finite-sum setting,  
 733 we have to take  $B = \mathcal{O} \left( \frac{L_\sigma^2}{\hat{L}^2} \right)$  to guarantee the degeneration up to  $1/p_a$  of the term  $\mathcal{P}_1$  from (32).

734 However, DASHA-PP-MVR has also suboptimal term  $\mathcal{P}_2$ . This suboptimality is tightly connected with  
 735 the suboptimality of  $B_{\text{init}}$  in the general nonconvex case, which we discuss in Section 6.3, and it also  
 736 appears in the analysis of DASHA-MVR (Tyurin and Richtárik, 2023). Let us provide the counterpart  
 737 of Corollary 4. The corollary reveals that we can escape regimes when  $\mathcal{P}_2$  is the bottleneck by  
 738 choosing the parameters of the compressors.

739 **Corollary 5.** Suppose that assumptions of Theorem 10 hold, batch size  $B \leq \min \left\{ \frac{\sigma}{p_a \sqrt{\mu \varepsilon n}}, \frac{L_\sigma^2}{\hat{L}^2} \right\}$ ,  
 740 we take RandK compressors with  $K = \Theta \left( \frac{B d \sqrt{\mu \varepsilon n}}{\sigma} \right)$ . Then the communication complexity equals

$$\tilde{\mathcal{O}} \left( \frac{d\sigma}{p_a \sqrt{\mu \varepsilon n}} + \frac{dL_\sigma}{p_a \mu \sqrt{n}} \right),$$

741 and the expected number of stochastic gradient calculations per node equals

$$\tilde{\mathcal{O}} \left( \frac{\sigma^2}{p_a \mu n \varepsilon} + \frac{\sigma L_\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon}} \right).$$

742 Up to Lipschitz constants, DASHA-PP-MVR has the state-of-the-art oracle complexity under PL-  
 743 condition (see (Li et al., 2021a)). Moreover, DASHA-PP-MVR has the state-of-the-art communication  
 744 complexity of DASHA for a small enough  $\mu$ .

<sup>7</sup>For simplicity, we omitted  $\frac{d}{\zeta_C}$  term from the complexity in the stochastic setting, where  $\zeta_C$  is defined in Definition 12. For instance, for the RandK compressor (see Definition 5 and Theorem 6),  $\zeta_C = K$  and  $\frac{d}{\zeta_C} = \Theta(\omega)$ .

#### 745 **F.4 Proofs of Theorems**

746 The following proofs almost repeat the proofs from Section E. And one of the main changes is that  
747 instead of Lemma 3, we use the following lemma.

##### 748 **F.4.1 Standard Lemma under Polyak-Łojasiewicz Condition**

749 **Lemma 11.** *Suppose that Assumptions 1 and 9 hold and*

$$\mathbb{E} [f(x^{t+1})] + \gamma \Psi^{t+1} \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] + (1 - \gamma\mu)\gamma \Psi^t + \gamma C,$$

750 where  $\Psi^t$  is a sequence of numbers,  $\Psi^t \geq 0$  for all  $t \in [T]$ , constant  $C \geq 0$ , constant  $\mu > 0$ , and  
751 constant  $\gamma \in (0, 1/\mu)$ . Then

$$\mathbb{E} [f(x^T) - f^*] \leq (1 - \gamma\mu)^T ((f(x^0) - f^*) + \gamma \Psi^0) + \frac{C}{\mu}. \quad (34)$$

752 *Proof.* We subtract  $f^*$  and use PL-condition (31) to get

$$\begin{aligned} \mathbb{E} [f(x^{t+1}) - f^*] + \gamma \Psi^{t+1} &\leq \mathbb{E} [f(x^t) - f^*] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] + \gamma \Psi^t + \gamma C \\ &\leq (1 - \gamma\mu) \mathbb{E} [f(x^t) - f^*] + (1 - \gamma\mu)\gamma \Psi^t + \gamma C \\ &= (1 - \gamma\mu) (\mathbb{E} [f(x^t) - f^*] + \gamma \Psi^t) + \gamma C. \end{aligned}$$

753 Unrolling the inequality, we have

$$\begin{aligned} \mathbb{E} [f(x^{t+1}) - f^*] + \gamma \Psi^{t+1} &\leq (1 - \gamma\mu)^{t+1} ((f(x^0) - f^*) + \gamma \Psi^0) + \gamma C \sum_{i=0}^t (1 - \gamma\mu)^i \\ &\leq (1 - \gamma\mu)^{t+1} ((f(x^0) - f^*) + \gamma \Psi^0) + \frac{C}{\mu}. \end{aligned}$$

754 It is left to note that  $\Psi^t \geq 0$  for all  $t \in [T]$ . □

##### 755 **F.4.2 Generic Lemma**

756 We now provide the counterpart of Lemma 6.

757 **Lemma 12.** *Suppose that Assumptions 2, 7, 8 and 9 hold and let us take  $a = \frac{p_a}{2\omega+1}$ , then*

$$\begin{aligned} &\mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ &\leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ &\quad + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ &\quad + \frac{10\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right]. \end{aligned}$$

758 *Proof.* Let us fix some constants  $\kappa, \eta \in [0, \infty)$  that we will define later. Using the same reasoning as  
759 in Lemma 6, we can get

$$\begin{aligned} &\mathbb{E} [f(x^{t+1})] \\ &\quad + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ &\leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \end{aligned}$$

$$\begin{aligned}
& + \left( \gamma + \kappa (1 - a)^2 \right) \mathbb{E} \left[ \|g^t - h^t\|^2 \right] \\
& + \left( \frac{\kappa a^2 ((2\omega + 1) p_a - p_{aa})}{np_a^2} + \eta \left( \frac{a^2 (2\omega + 1 - p_a)}{p_a} + (1 - a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left( \frac{2\kappa\omega}{np_a} + \frac{2\eta\omega}{p_a} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

760 Let us take  $\kappa = \frac{2\gamma}{a}$ . One can show that  $\gamma + \kappa (1 - a)^2 \leq (1 - \frac{a}{2}) \kappa$ , and thus

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] \\
& + \frac{2\gamma}{a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma}{a} \mathbb{E} [\|g^t - h^t\|^2] \\
& + \left( \frac{2\gamma a ((2\omega + 1) p_a - p_{aa})}{np_a^2} + \eta \left( \frac{a^2 (2\omega + 1 - p_a)}{p_a} + (1 - a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left( \frac{4\gamma\omega}{anp_a} + \frac{2\eta\omega}{p_a} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

761 Considering the choice of  $a$ , one can show that  $\left( \frac{a^2 (2\omega + 1 - p_a)}{p_a} + (1 - a)^2 \right) \leq 1 - a$ . If we take  
762  $\eta = \frac{4\gamma((2\omega+1)p_a-p_{aa})}{np_a^2}$ , then  $\left( \frac{2\gamma a ((2\omega + 1) p_a - p_{aa})}{np_a^2} + \eta \left( \frac{a^2 (2\omega + 1 - p_a)}{p_a} + (1 - a)^2 \right) \right) \leq (1 - \frac{a}{2}) \eta$  and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] \\
& + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1) p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left( 1 - \frac{a}{2} \right) \frac{4\gamma((2\omega + 1) p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left( \frac{2\gamma(2\omega + 1)\omega}{np_a^2} + \frac{8\gamma((2\omega + 1) p_a - p_{aa})\omega}{np_a^3} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \\
& \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left( 1 - \frac{a}{2} \right) \frac{4\gamma((2\omega + 1) p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{10\gamma(2\omega + 1)\omega}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right].
\end{aligned}$$

763 It is left to consider that  $\gamma \leq \frac{a}{2\mu}$ , and therefore  $1 - \frac{a}{2} \leq 1 - \gamma\mu$ . □

764 **E.4.3 Proof for DASHA-PP under PL-condition**

**Theorem 8.** Suppose that Assumption 1, 2, 3, 7, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_a}{2-p_a}$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega+1)}{np_a^2} + \frac{48}{np_a^2} \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}} \right)^{-1}, \frac{a}{4\mu} \right\},$$

765 and  $h_i^0 = g_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP), then  $\mathbb{E}[f(x^T)] - f^* \leq$   
 766  $(1 - \gamma\mu)^T \Delta_0$ .

767 *Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 12, Lemma 7,  
 768 and the law of total expectation, we obtain

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}\left[f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2\right] \\ & + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & + \frac{10\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E}\left[2\widehat{L}^2 \|x^{t+1} - x^t\|^2 + 2b^2 \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \nu \mathbb{E}\left[\frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \|x^{t+1} - x^t\|^2 + \frac{2b^2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2\right] \\ & + \rho \mathbb{E}\left[\frac{2(1-p_a)\widehat{L}^2}{p_a} \|x^{t+1} - x^t\|^2 + \left(\frac{2b^2(1-p_a)}{p_a} + (1-b)^2\right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right]. \end{aligned}$$

769 After rearranging the terms, we get

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{20\gamma\omega(2\omega+1)\widehat{L}^2}{np_a^2} - \nu \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} - \rho \frac{2(1-p_a)\widehat{L}^2}{p_a}\right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\ & + (\gamma + \nu(1-b)^2) \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] \\ & + \left(\frac{20b^2\gamma\omega(2\omega+1)}{np_a^2} + \nu \frac{2b^2(p_a - p_{aa})}{np_a^2} + \rho \left(\frac{2b^2(1-p_a)}{p_a} + (1-b)^2\right)\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right]. \end{aligned}$$

770 By taking  $\nu = \frac{2\gamma}{b}$ , one can show that  $(\gamma + \nu(1-b)^2) \leq (1 - \frac{b}{2})\nu$ , and

$$\mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right]$$

$$\begin{aligned}
& + \frac{2\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
\leq & \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{20\gamma\omega(2\omega + 1)\hat{L}^2}{np_a^2} - \frac{4\gamma(p_a - p_{aa})\hat{L}^2}{bnp_a^2} - \rho \frac{2(1 - p_a)\hat{L}^2}{p_a} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( 1 - \frac{b}{2} \right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{20b^2\gamma\omega(2\omega + 1)}{np_a^2} + \frac{4\gamma b(p_a - p_{aa})}{np_a^2} + \rho \left( \frac{2b^2(1 - p_a)}{p_a} + (1 - b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

771 Note that  $b = \frac{p_a}{2 - p_a}$ , thus

$$\begin{aligned}
& \left( \frac{20b^2\gamma\omega(2\omega + 1)}{np_a^2} + \frac{4\gamma b(p_a - p_{aa})}{np_a^2} + \rho \left( \frac{2b^2(1 - p_a)}{p_a} + (1 - b)^2 \right) \right) \\
& \leq \left( \frac{20b^2\gamma\omega(2\omega + 1)}{np_a^2} + \frac{4\gamma b(p_a - p_{aa})}{np_a^2} + \rho(1 - b) \right).
\end{aligned}$$

772 And if we take  $\rho = \frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}$ , then

$$\left( \frac{20b^2\gamma\omega(2\omega + 1)}{np_a^2} + \frac{4\gamma b(p_a - p_{aa})}{np_a^2} + \rho(1 - b) \right) \leq \left( 1 - \frac{b}{2} \right) \rho,$$

773 and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
\leq & \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{20\gamma\omega(2\omega + 1)\hat{L}^2}{np_a^2} - \frac{4\gamma(p_a - p_{aa})\hat{L}^2}{bnp_a^2} \right. \\
& \quad \left. - \frac{80b\gamma\omega(2\omega + 1)(1 - p_a)\hat{L}^2}{np_a^3} - \frac{16\gamma(p_a - p_{aa})(1 - p_a)\hat{L}^2}{np_a^3} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( 1 - \frac{b}{2} \right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( 1 - \frac{b}{2} \right) \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

774 Due to  $\frac{p_a}{2} \leq b \leq p_a$ , we have

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{100\gamma\omega(2\omega + 1)\hat{L}^2}{np_a^2} - \frac{24\gamma(p_a - p_{aa})\hat{L}^2}{np_a^3}\right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left(\frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
\end{aligned}$$

775 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
&\quad + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left(\frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
\end{aligned}$$

776 Note that  $\gamma \leq \frac{a}{4\mu} \leq \frac{p_a}{4\mu} \leq \frac{b}{2\mu}$ , thus  $1 - \frac{b}{2} \leq 1 - \gamma\mu$  and

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
&\quad + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad + (1 - \gamma\mu) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + (1 - \gamma\mu) \left(\frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
\end{aligned}$$

777 In the view of Lemma 11 with

$$\begin{aligned}
\Psi^t &= \frac{2(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{4((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad + \frac{2}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(\frac{40b\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right],
\end{aligned}$$

778 we can conclude the proof of the theorem.  $\square$

#### 779 F.4.4 Proof for DASHA-PP-PAGE under PL-condition

**Theorem 9.** Suppose that Assumption 1, 2, 3, 7, 4, 8, and 9 hold. Let us take  $a = \frac{p_a}{2\omega + 1}$ , probability

$$p_{\text{page}} = \frac{B}{m+B}, b = \frac{p_{\text{page}} p_a}{2 - p_a},$$

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega + 1)}{np_a^2} \left( \hat{L}^2 + \frac{(1 - p_{\text{page}}) L_{\max}^2}{B} \right)} + \frac{48}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}}) L_{\max}^2}{B} \right) \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

780 and  $h_i^0 = g_i^0 = \nabla f_i(x^0)$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-PAGE), then  $\mathbb{E}[f(x^T)] - f^* \leq$   
 781  $(1 - \gamma\mu)^T \Delta_0$ .

782 *Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 12, Lemma 8,  
 783 and the law of total expectation, we obtain

$$\begin{aligned}
 & \mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
 & + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
 & \leq \mathbb{E}\left[f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2\right] \\
 & + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
 & + \frac{10\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2\right] \\
 & + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
 & \leq \mathbb{E}\left[f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2\right] \\
 & + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
 & + \frac{10\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E}\left[\left(2\hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B}\right) \|x^{t+1} - x^t\|^2 + \frac{2b^2}{p_{\text{page}}} \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
 & + \nu \mathbb{E}\left[\left(\frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} + \frac{(1 - p_{\text{page}})L_{\max}^2}{np_a B}\right) \|x^{t+1} - x^t\|^2\right. \\
 & \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{page}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left(p_{\text{page}} \left(1 - \frac{b}{p_{\text{page}}}\right)^2 + (1 - p_{\text{page}})\right) \|h^t - \nabla f(x^t)\|^2\right] \\
 & + \rho \mathbb{E}\left[\left(\frac{2(1 - p_a)\hat{L}^2}{p_a} + \frac{(1 - p_{\text{page}})L_{\max}^2}{p_a B}\right) \|x^{t+1} - x^t\|^2\right. \\
 & \quad \left. + \left(\frac{2(1 - p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left(1 - \frac{b}{p_{\text{page}}}\right)^2 + (1 - p_{\text{page}})\right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right].
 \end{aligned}$$

784 After rearranging the terms, we get

$$\begin{aligned}
 & \mathbb{E}[f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
 & + \nu \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
 & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
 & + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right]
 \end{aligned}$$

$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( 2\widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \nu \left( \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} + \frac{(1-p_{\text{page}})L_{\max}^2}{np_a B} \right) - \rho \left( \frac{2(1-p_a)\widehat{L}^2}{p_a} + \frac{(1-p_{\text{page}})L_{\max}^2}{p_a B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( \gamma + \nu \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \right) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{2\nu(p_a - p_{aa})b^2}{np_a^2 p_{\text{page}}} \right. \\
& \quad \left. + \rho \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Due to  $b = \frac{p_{\text{page}} p_a}{2-p_a} \leq p_{\text{page}}$ , one can show that  $\left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \leq 1-b$ . Thus, if we take  $\nu = \frac{2\gamma}{b}$ , then

$$\left( \gamma + \nu \left( p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \right) \leq \gamma + \nu(1-b) = \left( 1 - \frac{b}{2} \right) \nu,$$

785 therefore

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( 2\widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \frac{2\gamma}{bn p_a} \left( 2 \left( 1 - \frac{p_{aa}}{p_a} \right) \widehat{L}^2 + \frac{(1-p_{\text{page}})L_{\max}^2}{B} \right) - \rho \left( \frac{2(1-p_a)\widehat{L}^2}{p_a} + \frac{(1-p_{\text{page}})L_{\max}^2}{p_a B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( 1 - \frac{b}{2} \right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{4\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{page}}} \right. \\
& \quad \left. + \rho \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Next, with the choice of  $b = \frac{p_{\text{page}} p_a}{2-p_a}$ , we ensure that

$$\left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \leq 1-b.$$

If we take  $\rho = \frac{40b^2\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}}$ , then

$$\left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{4\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{page}}} + \rho \left( \frac{2(1-p_a)b^2}{p_a p_{\text{page}}} + p_{\text{page}} \left( 1 - \frac{b}{p_{\text{page}}} \right)^2 + (1-p_{\text{page}}) \right) \right) \leq \left( 1 - \frac{b}{2} \right) \rho,$$



786 therefore

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( 2\hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \frac{2\gamma}{bnp_a} \left( 2 \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^3 p_{\text{page}}} + \frac{8\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2(1 - p_a) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left( 1 - \frac{b}{2} \right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( 1 - \frac{b}{2} \right) \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

Let us simplify the inequality. First, due to  $b \geq \frac{p_{\text{page}} p_a}{2}$ , we have

$$\frac{2\gamma}{bnp_a} \left( 2 \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \leq \frac{8\gamma}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right).$$

787 Second, due to  $b \leq p_a p_{\text{page}}$  and  $p_{aa} \leq p_a^2$ , we get

$$\begin{aligned}
& \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^3 p_{\text{page}}} + \frac{8\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2(1 - p_a) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \\
& \leq \left( \frac{40\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2 p_{\text{page}}} \right) \left( 2 \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \\
& \leq \frac{80\gamma\omega(2\omega+1)}{np_a^2} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \\
& \quad + \frac{16\gamma \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \\
& \leq \frac{80\gamma\omega(2\omega+1)}{np_a^2} \left( \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \\
& \quad + \frac{16\gamma}{np_a^2 p_{\text{page}}} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right).
\end{aligned}$$

788 Combining all bounds together, we obtain the following inequality:

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2]
\end{aligned}$$

$$\begin{aligned}
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{100\gamma\omega(2\omega + 1)}{np_a^2} \left( \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right. \\
& \quad \left. - \frac{24\gamma}{np_a^2 p_{\text{page}}} \left( \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 + \frac{(1 - p_{\text{page}})L_{\max}^2}{B} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

789 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

790 Note that  $\gamma \leq \frac{b}{2\mu}$ , thus  $1 - \frac{b}{2} \leq 1 - \gamma\mu$  and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{4\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + (1 - \gamma\mu) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + (1 - \gamma\mu) \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right].
\end{aligned}$$

791 It is left to apply Lemma 11 with

$$\begin{aligned}
\Psi^t &= \frac{2(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{4((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{2}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left( \frac{40b\omega(2\omega + 1)}{np_a^2 p_{\text{page}}} + \frac{8(p_a - p_{aa})}{np_a^2 p_{\text{page}}} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]
\end{aligned}$$

792 to conclude the proof.  $\square$

793 **F.4.5 Proof for DASHA-PP-MVR under PL-condition**

**Theorem 10.** Suppose that Assumption 1, 2, 3, 7, 5, 6, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 $b \in \left(0, \frac{p_a}{2-p_a}\right]$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{200\omega(2\omega+1)}{np_a^2} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \widehat{L}^2 \right) + \frac{40}{np_a b} \left( \frac{(1-b)^2 L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right)} \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

794 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 1 (DASHA-PP-MVR), then

$$\begin{aligned} & \mathbb{E} [f(x^T) - f^*] \\ & \leq (1-\gamma\mu)^T \left( \Delta_0 + \frac{2\gamma}{b} \|h^0 - \nabla f(x^0)\|^2 + \left( \frac{40\gamma b\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) \\ & \quad + \frac{1}{\mu} \left( \frac{100b^2\omega(2\omega+1)}{p_a^2} + \frac{20b}{p_a} \right) \frac{\sigma^2}{nB}. \end{aligned}$$

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & \quad + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & \quad + \frac{10\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right]. \end{aligned}$$

795 *Proof.* Let us fix constants  $\nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 12, Lemma 10,  
 796 and the law of total expectation, we obtain

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & \quad + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & \quad + \frac{10\gamma(2\omega+1)\omega}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|k_i^{t+1}\|^2 \right] \\ & \quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma \|h^t - \nabla f(x^t)\|^2 \right] \\ & \quad + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{10\gamma\omega(2\omega+1)}{np_a^2} \mathbb{E} \left[ \frac{2b^2\sigma^2}{B} + \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 + 2b^2 \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \nu \mathbb{E} \left( \frac{2b^2\sigma^2}{np_a B} + \left( \frac{2(1-b)^2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b)^2 \|h^t - \nabla f(x^t)\|^2 \right) \\
& + \rho \mathbb{E} \left( \frac{2b^2\sigma^2}{p_a B} + \left( \frac{2(1-b)^2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right. \\
& \quad \left. + \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right).
\end{aligned}$$

797 After rearranging the terms, we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right. \\
& \quad \left. - \nu \left( \frac{2(1-b)^2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2(1-b)^2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + (\gamma + \nu(1-b)^2) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{2\nu(p_a - p_{aa})b^2}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2} + \nu \frac{2b^2}{np_a} + \rho \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

798 By taking  $\nu = \frac{2\gamma}{b}$ , one can show that  $(\gamma + \nu(1-b)^2) \leq (1 - \frac{b}{2})\nu$ , and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right.
\end{aligned}$$

$$\begin{aligned}
& -\frac{2\gamma}{b} \left( \frac{2(1-b)^2 L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa}) \hat{L}^2}{np_a^2} \right) - \rho \left( \frac{2(1-b)^2 L_\sigma^2}{p_a B} + \frac{2(1-p_a) \hat{L}^2}{p_a} \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma (p_a - p_{aa}) b}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma b}{np_a} + \rho \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

799 Note that  $b \leq \frac{p_a}{2-p_a}$ , thus

$$\begin{aligned}
& \left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma (p_a - p_{aa}) b}{np_a^2} + \rho \left( \frac{2(1-p_a)b^2}{p_a} + (1-b)^2 \right) \right) \\
& \leq \left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma (p_a - p_{aa}) b}{np_a^2} + \rho (1-b) \right).
\end{aligned}$$

800 And if we take  $\rho = \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a-p_{aa})}{np_a^2}$ , then

$$\left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma (p_a - p_{aa}) b}{np_a^2} + \rho (1-b) \right) \leq \rho,$$

801 and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a-p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2\hat{L}^2 \right) \right. \\
& \quad \left. - \frac{2\gamma}{np_a b} \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2 \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^3} + \frac{8\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2} \right) \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 2(1-p_a) \hat{L}^2 \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a-p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{20b^2 \gamma \omega (2\omega + 1)}{np_a^2} + \frac{4\gamma b}{np_a} + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a-p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

802 Let us simplify the inequality. First, due to  $b \leq p_a$  and  $(1-p_a) \leq \left(1 - \frac{p_{aa}}{p_a}\right)$ , we have

$$\left( \frac{40b\gamma\omega(2\omega+1)}{np_a^3} + \frac{2\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2} \right) \left( \frac{2(1-b)^2 L_\sigma^2}{B} + 8(1-p_a) \hat{L}^2 \right)$$

$$\begin{aligned}
&= \frac{40b\gamma\omega(2\omega+1)}{np_a^3} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2(1-p_a)\widehat{L}^2 \right) \\
&\quad + \frac{8\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2(1-p_a)\widehat{L}^2 \right) \\
&\leq \frac{40\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\widehat{L}^2 \right) \\
&\quad + \frac{8\gamma}{np_ab} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2 \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right),
\end{aligned}$$

803 therefore

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\
&\quad + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{50\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2\widehat{L}^2 \right) \right. \\
&\quad \left. - \frac{10\gamma}{np_ab} \left( \frac{2(1-b)^2L_\sigma^2}{B} + 2 \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
&\quad + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{4\gamma b}{np_a} + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B} \\
&\leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
&\quad + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{100\gamma\omega(2\omega+1)}{np_a^2} \left( \frac{(1-b)^2L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \frac{20\gamma}{np_ab} \left( \frac{(1-b)^2L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2 \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
&\quad + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\
&\quad + \left( \frac{20b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{4\gamma b}{np_a} + \left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

804 Also, we can simplify the last term:

$$\begin{aligned}
&\left( \frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \frac{2b^2}{p_a} \\
&= \frac{80b^3\gamma\omega(2\omega+1)}{np_a^3} + \frac{16b^2\gamma \left(1 - \frac{p_{aa}}{p_a}\right)}{np_a^2}
\end{aligned}$$

$$\leq \frac{80b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{16b\gamma}{np_a},$$

805 thus

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{100\gamma\omega(2\omega+1)}{np_a^2} \left(\frac{(1-b)^2 L_\sigma^2}{B} + \widehat{L}^2\right)\right. \\ & \quad \left. - \frac{20\gamma}{np_a b} \left(\frac{(1-b)^2 L_\sigma^2}{B} + \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2\right)\right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\ & + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left(\frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \left(\frac{100b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{20\gamma b}{np_a}\right) \frac{\sigma^2}{B}. \end{aligned}$$

806 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \\ & + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \left(\frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2\right] \\ & + \left(\frac{100b^2\gamma\omega(2\omega+1)}{np_a^2} + \frac{20\gamma b}{np_a}\right) \frac{\sigma^2}{B}. \end{aligned}$$

807 Note that  $\gamma \leq \frac{b}{2\mu}$ , thus  $1 - \frac{b}{2} \leq 1 - \gamma\mu$  and

$$\begin{aligned} & \mathbb{E}[f(x^{t+1})] + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2\right] \\ & + \frac{2\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \left(\frac{40b\gamma\omega(2\omega+1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2}\right) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2\right] \\ & \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\ & + (1-\gamma\mu) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + (1-\gamma\mu) \frac{4\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2\right] \end{aligned}$$

$$\begin{aligned}
& + (1 - \gamma\mu) \frac{2\gamma}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + (1 - \gamma\mu) \left( \frac{40b\gamma\omega(2\omega + 1)}{np_a^2} + \frac{8\gamma(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{100b^2\gamma\omega(2\omega + 1)}{np_a^2} + \frac{20\gamma b}{np_a} \right) \frac{\sigma^2}{B}.
\end{aligned}$$

808 It is left to apply Lemma 11 with

$$\begin{aligned}
\Psi^t &= \frac{2(2\omega + 1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{4((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&+ \frac{2}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \left( \frac{40b\omega(2\omega + 1)}{np_a^2} + \frac{8(p_a - p_{aa})}{np_a^2} \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]
\end{aligned}$$

809 and  $C = \left( \frac{100b^2\omega(2\omega+1)}{p_a^2} + \frac{20b}{p_a} \right) \frac{\sigma^2}{nB}$  to conclude the proof.  $\square$

810 **Corollary 5.** Suppose that assumptions of Theorem 10 hold, batch size  $B \leq \min \left\{ \frac{\sigma}{p_a\sqrt{\mu\varepsilon n}}, \frac{L_\sigma^2}{L^2} \right\}$ ,  
811 we take RandK compressors with  $K = \Theta \left( \frac{Bd\sqrt{\mu\varepsilon n}}{\sigma} \right)$ . Then the communication complexity equals

$$\tilde{\mathcal{O}} \left( \frac{d\sigma}{p_a\sqrt{\mu\varepsilon n}} + \frac{dL_\sigma}{p_a\mu\sqrt{n}} \right),$$

812 and the expected number of stochastic gradient calculations per node equals

$$\tilde{\mathcal{O}} \left( \frac{\sigma^2}{p_a\mu n\varepsilon} + \frac{\sigma L_\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon}} \right).$$

813 *Proof.* In the view of Theorem 10, DASHA-PP have to run

$$\tilde{\mathcal{O}} \left( \frac{\omega + 1}{p_a} + \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + \frac{\sigma^2}{p_a\mu n\varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon B}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right)$$

814 communication rounds in the stochastic settings to get  $\varepsilon$ -solution. Note that  $K = \mathcal{O} \left( \frac{d}{p_a\sqrt{n}} \right)$ .

815 Moreover, we can skip the initialization procedure and initialize  $h_i^0$  and  $g_i^0$ , for instance, with zeros  
816 because the initialization error is under a logarithm. Considering Theorem 6, the communication  
817 complexity equals

$$\begin{aligned}
& \tilde{\mathcal{O}} \left( K \frac{\omega + 1}{p_a} + K \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + K \frac{\sigma^2}{p_a\mu n\varepsilon B} + K \frac{L}{\mu} + K \frac{\omega}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + K \frac{\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon B}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right) \\
&= \tilde{\mathcal{O}} \left( K \frac{\omega + 1}{p_a} + K \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + K \frac{\sigma^2}{p_a\mu n\varepsilon B} + K \frac{L}{\mu} + K \frac{\omega}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + K \frac{\sigma L_\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon B}} \right) \\
&= \tilde{\mathcal{O}} \left( \frac{d}{p_a} + \frac{d}{p_a} \sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + \frac{K\sigma^2}{p_a\mu n\varepsilon B} + \frac{dL}{p_a\mu\sqrt{n}} + \frac{d}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{K\sigma L_\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon B}} \right) \\
&= \tilde{\mathcal{O}} \left( \frac{d}{p_a} + \frac{d\sigma}{p_a\sqrt{\mu n\varepsilon B}} + \frac{d\sigma}{p_a\sqrt{\mu\varepsilon n}} + \frac{dL}{p_a\mu\sqrt{n}} + \frac{d}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{dL_\sigma}{p_a\mu\sqrt{n}} \right) \\
&= \tilde{\mathcal{O}} \left( \frac{d\sigma}{p_a\sqrt{\mu\varepsilon n}} + \frac{dL_\sigma}{p_a\mu\sqrt{n}} \right).
\end{aligned}$$

818 The expected number of stochastic gradient calculations per node equals

$$\tilde{\mathcal{O}} \left( B \frac{\omega + 1}{p_a} + B \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + B \frac{\sigma^2}{p_a\mu n\varepsilon B} + B \frac{L}{\mu} + B \frac{\omega}{p_a\mu\sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + B \frac{\sigma}{p_a n\mu^{3/2}\sqrt{\varepsilon B}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) \right)$$



$$\begin{aligned}
&= \tilde{\mathcal{O}} \left( B \frac{\omega + 1}{p_a} + B \frac{\omega}{p_a} \sqrt{\frac{\sigma^2}{\mu n \varepsilon B}} + B \frac{\sigma^2}{p_a \mu n \varepsilon B} + B \frac{L}{\mu} + B \frac{\omega}{p_a \mu \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + B \frac{\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon B}} \left( \frac{L_\sigma}{\sqrt{B}} \right) \right) \\
&= \tilde{\mathcal{O}} \left( \frac{Bd}{K p_a} + \frac{Bd}{K p_a} \sqrt{\frac{\sigma^2}{\mu n \varepsilon B}} + \frac{\sigma^2}{p_a \mu n \varepsilon} + B \frac{L}{\mu} + \frac{Bd}{K p_a \mu \sqrt{n}} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma L_\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon}} \right) \\
&= \tilde{\mathcal{O}} \left( \frac{\sigma}{p_a \sqrt{\mu \varepsilon n}} + \frac{\sigma^2}{p_a \mu \varepsilon n \sqrt{B}} + \frac{\sigma^2}{p_a \mu n \varepsilon} + \frac{\sigma L}{p_a \mu^{3/2} \sqrt{\varepsilon} n} + \frac{\sigma}{p_a \mu^{3/2} \sqrt{\varepsilon} n} \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma L_\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon}} \right) \\
&= \tilde{\mathcal{O}} \left( \frac{\sigma^2}{p_a \mu n \varepsilon} + \frac{\sigma L_\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon}} \right).
\end{aligned}$$

819

□

## 820 G Description of DASHA-PP-SYNC-MVR

821 By analogy to (Tyurin and Richtárik, 2023), we provide a “synchronized” version of the algorithm.  
 822 With a small probability, participating nodes calculate and send a mega batch without compression.  
 823 This helps us to resolve the suboptimality of DASHA-PP-MVR w.r.t.  $\omega$ . Note that this suboptimality is  
 824 not a problem. We show in Corollary 4 that DASHA-PP-MVR can have the optimal oracle complexity  
 825 and SOTA communication complexity with the particular choices of parameters of the compressors.

---

### Algorithm 8 DASHA-PP-SYNC-MVR

---

```

1: Input: starting point  $x^0 \in \mathbb{R}^d$ , stepsize  $\gamma > 0$ , momentum  $a \in (0, 1]$ , momentum  $b \in$ 
   (0, 1], probability  $p_{\text{mega}} \in (0, 1]$ , batch size  $B'$  and  $B$ , probability  $p_a \in (0, 1]$  that a node is
   participating(a), number of iterations  $T \geq 1$ .
2: Initialize  $g_i^0, h_i^0$  on the nodes and  $g^0 = \frac{1}{n} \sum_{i=1}^n g_i^0$  on the server
3: for  $t = 0, 1, \dots, T - 1$  do
4:    $x^{t+1} = x^t - \gamma g^t$ 
5:    $c^{t+1} = \begin{cases} 1, & \text{with probability } p_{\text{mega}}, \\ 0, & \text{with probability } 1 - p_{\text{mega}} \end{cases}$ 
6:   Broadcast  $x^{t+1}, x^t$  to all participating(a) nodes
7:   for  $i = 1, \dots, n$  in parallel do
8:     if  $i^{\text{th}}$  node is participating(a) then
9:       if  $c^{t+1} = 1$  then
10:        Generate i.i.d. samples  $\{\xi_{ik}^{t+1}\}_{k=1}^{B'}$  of size  $B'$  from  $\mathcal{D}_i$ .
11:         $k_i^{t+1} = \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) - \frac{b}{p_{\text{mega}}} \left( h_i^t - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) \right)$ 
12:         $m_i^{t+1} = \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t)$ 
13:       else
14:        Generate i.i.d. samples  $\{\xi_{ij}^{t+1}\}_{j=1}^B$  of size  $B$  from  $\mathcal{D}_i$ .
15:         $k_i^{t+1} = \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1})$ 
16:         $m_i^{t+1} = C_i \left( \frac{1}{p_a} k_i^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right)$ 
17:       end if
18:        $h_i^{t+1} = h_i^t + \frac{1}{p_a} k_i^{t+1}$ 
19:        $g_i^{t+1} = g_i^t + m_i^{t+1}$ 
20:       Send  $m_i^{t+1}$  to the server
21:     else
22:        $h_i^{t+1} = h_i^t$ 
23:        $m_i^{t+1} = 0$ 
24:        $g_i^{t+1} = g_i^t$ 
25:     end if
26:   end for
27:    $g^{t+1} = g^t + \frac{1}{n} \sum_{i=1}^n m_i^{t+1}$ 
28: end for
29: Output:  $\hat{x}^T$  chosen uniformly at random from  $\{x^t\}_{k=0}^{T-1}$ 
   (a): For the formal description see Section 2.2.

```

---

826 In the following theorem, we provide the convergence rate of DASHA-PP-SYNC-MVR.

**Theorem 11.** Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 $b = \frac{p_{\text{mega}} p_a}{2-p_a}$ , probability  $p_{\text{mega}} \in (0, 1]$ , batch size  $B' \geq B \geq 1$

$$\gamma \leq \left( L + \sqrt{\frac{8(2\omega+1)\omega}{np_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right) + \frac{16}{np_{\text{mega}} p_a^2} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{L_\sigma^2}{B} \right)} \right)^{-1},$$

827 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 8. Then

$$\begin{aligned} \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] &\leq \frac{1}{T} \left[ \frac{2\Delta_0}{\gamma} + \frac{4}{p_{\text{mega}} p_a} \|h^0 - \nabla f(x^0)\|^2 + \frac{4 \left(1 - \frac{p_{aa}}{p_a}\right)}{n p_{\text{mega}} p_a} \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right] \\ &\quad + \frac{12\sigma^2}{nB'}. \end{aligned}$$

828 First, we introduce the expected density of compressors (Gorbunov et al., 2021; Tyurin and Richtárik, 2023).

830 **Definition 12.** The expected density of the compressor  $\mathcal{C}_i$  is  $\zeta_{\mathcal{C}_i} := \sup_{x \in \mathbb{R}^d} \mathbb{E} [\|\mathcal{C}_i(x)\|_0]$ , where  
831  $\|x\|_0$  is the number of nonzero components of  $x \in \mathbb{R}^d$ . Let  $\zeta_{\mathcal{C}} = \max_{i \in [n]} \zeta_{\mathcal{C}_i}$ .

832 Note that  $\zeta_{\mathcal{C}}$  is finite and  $\zeta_{\mathcal{C}} \leq d$ .

833 In the next corollary, we choose particular algorithm parameters to reveal the communication and  
834 oracle complexity.

**Corollary 6.** Suppose that assumptions from Theorem 11 hold, probability  $p_{\text{mega}} = \min \left\{ \frac{\zeta_{\mathcal{C}}}{d}, \frac{n\varepsilon B}{\sigma^2} \right\}$ ,  
batch size  $B' = \Theta \left( \frac{\sigma^2}{n\varepsilon} \right)$ , and  $h_i^0 = g_i^0 = \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0)$  for all  $i \in [n]$ , initial batch size  
 $B_{\text{init}} = \Theta \left( \frac{B}{p_{\text{mega}} \sqrt{p_a}} \right) = \Theta \left( \max \left\{ \frac{Bd}{\sqrt{p_a} \zeta_{\mathcal{C}}}, \frac{\sigma^2}{\sqrt{p_a} n\varepsilon} \right\} \right)$ , then DASHA-PP-SYNC-MVR needs

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \left( \frac{\omega}{p_a \sqrt{n}} + \sqrt{\frac{d}{p_a^2 \zeta_{\mathcal{C}} n}} \right) \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon} n} \left( \frac{\hat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right] + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right).$$

835 communication rounds to get an  $\varepsilon$ -solution, the expected communication complexity is equal to  
836  $\mathcal{O}(d + \zeta_{\mathcal{C}} T)$ , and the expected number of stochastic gradient calculations per node equals  $\mathcal{O}(B_{\text{init}} +$   
837  $BT)$ , where  $\zeta_{\mathcal{C}}$  is the expected density from Definition 12.

838 The main improvement of Corollary 6 over Corollary 3 is the size of the initial batch size  $B_{\text{init}}$ .  
839 However, Corollary 4 reveals that we can avoid regimes when DASHA-PP-MVR is suboptimal.

840 We also provide a theorem under PL-condition (see Assumption 9).

**Theorem 13.** Suppose that Assumptions 1, 2, 3, 5, 6, 7, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  
 $b = \frac{p_{\text{mega}} p_a}{2-p_a}$ , probability  $p_{\text{mega}} \in (0, 1]$ , batch size  $B' \geq B \geq 1$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{16(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \hat{L}^2 \right) + \left( \frac{48L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{24 \left(1 - \frac{p_{aa}}{p_a}\right) \hat{L}^2}{np_{\text{mega}} p_a^2} \right)} \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

841 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 8. Then

$$\begin{aligned} &\mathbb{E} [f(x^T) - f^*] \\ &\leq (1 - \gamma\mu)^T \left( \Delta_0 + \frac{2\gamma}{b} \|h^0 - \nabla f(x^0)\|^2 + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) + \frac{20\sigma^2}{\mu n B'}. \end{aligned}$$

842 Let us provide bounds up to logarithmic factors and use  $\tilde{\mathcal{O}}(\cdot)$  notation.

**Corollary 7.** Suppose that assumptions from Theorem 13 hold, probability  $p_{\text{mega}} =$   
 $\min \left\{ \frac{\zeta_{\mathcal{C}}}{d}, \frac{\mu n \varepsilon B}{\sigma^2} \right\}$ , batch size  $B' = \Theta \left( \frac{\sigma^2}{\mu n \varepsilon} \right)$  then DASHA-PP-SYNC-MVR needs

$$T := \tilde{\mathcal{O}} \left( \frac{\omega+1}{p_a} + \frac{d}{p_a \zeta_{\mathcal{C}}} + \frac{\sigma^2}{p_a \mu n \varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_a \mu \sqrt{n}} \left( \frac{L_\sigma}{\sqrt{B}} + \hat{L} \right) + \left( \frac{\sqrt{d}}{p_a \mu \sqrt{\zeta_{\mathcal{C}} n}} + \frac{\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon B}} \right) \left( \frac{L_\sigma}{\sqrt{B}} + \hat{L} \right) \right).$$

communication rounds to get an  $\varepsilon$ -solution, the expected communication complexity is equal to  $\tilde{\mathcal{O}}(\zeta_c T)$ , and the expected number of stochastic gradient calculations per node equals  $\tilde{\mathcal{O}}(BT)$ , where  $\zeta_c$  is the expected density from Definition 12.

The proof of this corollary almost repeats the proof of Corollary 6. Note that we can skip the initialization procedure and initialize  $h_i^0$  and  $g_i^0$ , for instance, with zeros because the initialization error is under a logarithm.

Let us assume that  $\frac{d}{\zeta_c} = \Theta(\omega)$  (holds for the RandK compressor), then the convergence rate of DASHA-PP-SYNC-MVR is

$$\tilde{\mathcal{O}}\left(\frac{\omega+1}{p_a} + \frac{\sigma^2}{p_a \mu n \varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_a \mu \sqrt{n}} \left(\frac{L_\sigma}{\sqrt{B}} + \hat{L}\right) + \frac{\sigma}{p_a n \mu^{3/2} \sqrt{\varepsilon B}} \left(\frac{L_\sigma}{\sqrt{B}} + \hat{L}\right)\right). \quad (35)$$

Comparing (35) with the rate of DASHA-PP-MVR (32), one can see that DASHA-PP-SYNC-MVR improves the suboptimal term  $\mathcal{P}_2$  from (32). However, Corollary 5 reveals that we can escape these suboptimal regimes by choosing the parameter  $K$  of RandK compressors in a particular way.

### G.1 Proof for DASHA-PP-SYNC-MVR

In this section, we provide the proof of the convergence rate for DASHA-PP-SYNC-MVR. There are four different sources of randomness in Algorithm 8: the first one from random samples  $\xi_i^{t+1}$ , the second one from compressors  $\{\mathcal{C}_i\}_{i=1}^n$ , the third one from availability of nodes, and the fourth one from  $c^{t+1}$ . We define  $\mathbb{E}_k[\cdot]$ ,  $\mathbb{E}_c[\cdot]$ ,  $\mathbb{E}_{p_a}[\cdot]$  and  $\mathbb{E}_{p_{\text{mega}}}[\cdot]$  to be conditional expectations w.r.t.  $\xi_i^{t+1}$ ,  $\{\mathcal{C}_i\}_{i=1}^n$ , availability, and  $c^{t+1}$ , accordingly, conditioned on all previous randomness. Moreover, we define  $\mathbb{E}_{t+1}[\cdot]$  to be a conditional expectation w.r.t. all randomness in iteration  $t+1$  conditioned on all previous randomness.

Let us denote

$$\begin{aligned} k_{i,1}^{t+1} &:= \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) - \frac{b}{p_{\text{mega}}} \left( h_i^t - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) \right), \\ k_{i,2}^{t+1} &:= \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1}), \\ h_{i,1}^{t+1} &:= \begin{cases} h_i^t + \frac{1}{p_a} k_{i,1}^{t+1}, & i^{\text{th}} \text{ node is participating,} \\ h_i^t, & \text{otherwise,} \end{cases} \\ h_{i,2}^{t+1} &:= \begin{cases} h_i^t + \frac{1}{p_a} k_{i,2}^{t+1}, & i^{\text{th}} \text{ node is participating,} \\ h_i^t, & \text{otherwise,} \end{cases} \\ g_{i,1}^{t+1} &:= \begin{cases} g_i^t + \frac{1}{p_a} k_{i,1}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t), & i^{\text{th}} \text{ node is participating,} \\ g_i^t, & \text{otherwise,} \end{cases} \\ g_{i,2}^{t+1} &:= \begin{cases} g_i^t + \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right), & i^{\text{th}} \text{ node is participating,} \\ g_i^t, & \text{otherwise,} \end{cases} \end{aligned}$$

$h_1^{t+1} := \frac{1}{n} \sum_{i=1}^n h_{i,1}^{t+1}$ ,  $h_2^{t+1} := \frac{1}{n} \sum_{i=1}^n h_{i,2}^{t+1}$ ,  $g_1^{t+1} := \frac{1}{n} \sum_{i=1}^n g_{i,1}^{t+1}$ , and  $g_2^{t+1} := \frac{1}{n} \sum_{i=1}^n g_{i,2}^{t+1}$ . Note, that

$$h^{t+1} = \begin{cases} h_1^{t+1}, & c^{t+1} = 1, \\ h_2^{t+1}, & c^{t+1} = 0, \end{cases}$$

and

$$g^{t+1} = \begin{cases} g_1^{t+1}, & c^{t+1} = 1, \\ g_2^{t+1}, & c^{t+1} = 0 \end{cases}$$

First, we will prove two lemmas.

867 **Lemma 13.** Suppose that Assumptions 3, 5, 7 and 8 hold and let us consider sequences  $\{g_i^{t+1}\}_{i=1}^n$   
 868 and  $\{h_i^{t+1}\}_{i=1}^n$  from Algorithm 8, then

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \right] \\ & \leq \frac{2(1-p_{\text{mega}})\omega}{n^2 p_a} \sum_{i=1}^n \|k_{i,2}^{t+1}\|^2 + \left( \frac{(p_a - p_{aa})a^2}{n^2 p_a^2} + \frac{2(1-p_{\text{mega}})a^2\omega}{n^2 p_a} \right) \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \\ & \quad + (1-a)^2 \|g^t - h^t\|^2, \end{aligned}$$

869 and

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \right] \\ & \leq \frac{2(1-p_{\text{mega}})\omega}{p_a} \|k_{i,2}^{t+1}\|^2 + \left( \frac{(1-p_a)a^2}{p_a} + \frac{2(1-p_{\text{mega}})a^2\omega}{p_a} \right) \|g_i^t - h_i^t\|^2 \\ & \quad + (1-a)^2 \|g_i^t - h_i^t\|^2, \quad \forall i \in [n]. \end{aligned}$$

870 *Proof.* First, we get the bound for  $\mathbb{E}_{t+1} [\|g^{t+1} - h^{t+1}\|^2]$ :

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \right] \\ & = p_{\text{mega}} \mathbb{E}_{p_a} \left[ \|g_1^{t+1} - h_1^{t+1}\|^2 \right] + (1-p_{\text{mega}}) \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g_2^{t+1} - h_2^{t+1}\|^2 \right] \right]. \end{aligned}$$

871 Using

$$\mathbb{E}_{p_a} [g_{i,1}^{t+1} - h_{i,1}^{t+1}] = g_i^t + k_{i,1}^{t+1} - a(g_i^t - h_i^t) - h_i^t - k_{i,1}^{t+1} = (1-a)(g_i^t - h_i^t)$$

872 and

$$\mathbb{E}_{\mathcal{C}} [\mathbb{E}_{p_a} [g_{i,2}^{t+1} - h_{i,2}^{t+1}]] = g_i^t + k_{i,2}^{t+1} - a(g_i^t - h_i^t) - h_i^t - k_{i,2}^{t+1} = (1-a)(g_i^t - h_i^t),$$

873 we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \right] \\ & \stackrel{(17)}{=} p_{\text{mega}} \mathbb{E}_{p_a} \left[ \|g_1^{t+1} - h_1^{t+1} - \mathbb{E}_{p_a} [g_1^{t+1} - h_1^{t+1}]\|^2 \right] \\ & \quad + (1-p_{\text{mega}}) \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \|g_2^{t+1} - h_2^{t+1} - \mathbb{E}_{p_a} [g_2^{t+1} - h_2^{t+1}]\|^2 \right] \right] \\ & \quad + (1-a)^2 \|g^t - h^t\|^2. \end{aligned}$$

874 We can use Lemma 1 two times with i)  $r_i = g_i^t - h_i^t$  and  $s_i = -a(g_i^t - h_i^t)$  and ii)  $r_i = g_i^t - h_i^t$  and

875  $s_i = p_a \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - k_{i,2}^{t+1}$ , to obtain

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \right] \\ & \leq \frac{p_{\text{mega}} a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \\ & \quad + (1-p_{\text{mega}}) \left( \frac{1}{n^2 p_a} \sum_{i=1}^n \mathbb{E}_{\mathcal{C}} \left[ \left\| p_a \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - (k_{i,2}^{t+1} - a(g_i^t - h_i^t)) \right\|^2 \right] \right) \\ & \quad + (1-p_{\text{mega}}) \left( \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right) \\ & \quad + (1-a)^2 \|g^t - h^t\|^2 \\ & = \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \end{aligned}$$

$$\begin{aligned}
& + (1 - p_{\text{mega}}) \left( \frac{p_a}{n^2} \sum_{i=1}^n \mathbb{E}_C \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right\|^2 \right] \right) \\
& + (1 - a)^2 \|g^t - h^t\|^2 \\
& \leq \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \\
& + \frac{(1 - p_{\text{mega}}) p_a \omega}{n^2} \sum_{i=1}^n \left\| \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right\|^2 \\
& + (1 - a)^2 \|g^t - h^t\|^2 \\
& = \frac{a^2 (p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \\
& + \frac{(1 - p_{\text{mega}}) \omega}{n^2 p_a} \sum_{i=1}^n \|k_{i,2}^{t+1} - a (g_i^t - h_i^t)\|^2 \\
& + (1 - a)^2 \|g^t - h^t\|^2.
\end{aligned}$$

876 In the last inequality, we use Assumption 7. Next, using (16), we have

$$\begin{aligned}
& \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] \right] \right] \\
& \leq \frac{2(1 - p_{\text{mega}}) \omega}{n^2 p_a} \sum_{i=1}^n \|k_{i,2}^{t+1}\|^2 + \left( \frac{(p_a - p_{aa}) a^2}{n^2 p_a^2} + \frac{2(1 - p_{\text{mega}}) \omega a^2}{n^2 p_a} \right) \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \\
& + (1 - a)^2 \|g^t - h^t\|^2.
\end{aligned}$$

877 The second inequality can be proved almost in the same way:

$$\begin{aligned}
& \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \right] \\
& = p_{\text{mega}} \mathbb{E}_{p_a} \left[ \|g_{i,1}^{t+1} - h_{i,1}^{t+1}\|^2 \right] + (1 - p_{\text{mega}}) \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g_{i,2}^{t+1} - h_{i,2}^{t+1}\|^2 \right] \right] \\
& \stackrel{(17)}{=} p_{\text{mega}} \mathbb{E}_{p_a} \left[ \|g_{i,1}^{t+1} - h_{i,1}^{t+1} - (1 - a) (g_i^t - h_i^t)\|^2 \right] + (1 - p_{\text{mega}}) \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g_{i,2}^{t+1} - h_{i,2}^{t+1}\|^2 \right] \right] \\
& + p_{\text{mega}} (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& = \frac{p_{\text{mega}} (1 - p_a) a^2}{p_a} \|g_i^t - h_i^t\|^2 + (1 - p_{\text{mega}}) \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g_{i,2}^{t+1} - h_{i,2}^{t+1}\|^2 \right] \right] \\
& + p_{\text{mega}} (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& \stackrel{(17)}{=} \frac{p_{\text{mega}} (1 - p_a) a^2}{p_a} \|g_i^t - h_i^t\|^2 + (1 - p_{\text{mega}}) \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \|g_{i,2}^{t+1} - h_{i,2}^{t+1} - (1 - a) (g_i^t - h_i^t)\|^2 \right] \right] \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& = \frac{p_{\text{mega}} (1 - p_a) a^2}{p_a} \|g_i^t - h_i^t\|^2 \\
& + (1 - p_{\text{mega}}) p_a \mathbb{E}_C \left[ \left\| g_i^t + \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( h_i^t + \frac{1}{p_a} k_{i,2}^{t+1} \right) - (1 - a) (g_i^t - h_i^t) \right\|^2 \right] \\
& + (1 - p_{\text{mega}}) (1 - p_a) \|g_i^t - h_i^t - (1 - a) (g_i^t - h_i^t)\|^2 \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& = \frac{p_{\text{mega}} (1 - p_a) a^2}{p_a} \|g_i^t - h_i^t\|^2 \\
& + (1 - p_{\text{mega}}) p_a \mathbb{E}_C \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_{i,2}^{t+1} - a (g_i^t - h_i^t) \right) \right\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& + (1 - p_{\text{mega}}) (1 - p_a) a^2 \|g_i^t - h_i^t\|^2 \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& \stackrel{(17)}{=} \left( \frac{p_{\text{mega}}(1 - p_a)a^2}{p_a} + \frac{(1 - p_{\text{mega}})(1 - p_a)a^2}{p_a} \right) \|g_i^t - h_i^t\|^2 \\
& + (1 - p_{\text{mega}}) p_a \mathbb{E} \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right\|^2 \right] \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& = \frac{(1 - p_a)a^2}{p_a} \|g_i^t - h_i^t\|^2 \\
& + (1 - p_{\text{mega}}) p_a \mathbb{E} \left[ \left\| \mathcal{C}_i \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) - \left( \frac{1}{p_a} k_{i,2}^{t+1} - \frac{a}{p_a} (g_i^t - h_i^t) \right) \right\|^2 \right] \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& \leq \frac{(1 - p_a)a^2}{p_a} \|g_i^t - h_i^t\|^2 \\
& + \frac{(1 - p_{\text{mega}})\omega}{p_a} \|k_{i,2}^{t+1} - a(g_i^t - h_i^t)\|^2 \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2 \\
& \stackrel{(16)}{\leq} \frac{2(1 - p_{\text{mega}})\omega}{p_a} \|k_{i,2}^{t+1}\|^2 + \left( \frac{(1 - p_a)a^2}{p_a} + \frac{2(1 - p_{\text{mega}})a^2\omega}{p_a} \right) \|g_i^t - h_i^t\|^2 \\
& + (1 - a)^2 \|g_i^t - h_i^t\|^2.
\end{aligned}$$

878

□

879 **Lemma 14.** Suppose that Assumptions 3, 5, 6 and 8 hold and let us consider sequence  $\{h_i^{t+1}\}_{i=1}^n$   
880 from Algorithm 8, then

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \\
& \leq \frac{2b^2\sigma^2}{np_{\text{mega}}p_a B'} + \left( \frac{2p_{\text{mega}}L_\sigma^2}{np_a B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1 - p_{\text{mega}})L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\hat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \\
& + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2,
\end{aligned}$$

881

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\
& \leq \frac{2b^2\sigma^2}{p_a p_{\text{mega}} B'} + \left( \frac{2p_{\text{mega}}L_\sigma^2}{p_a B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1 - p_{\text{mega}})L_\sigma^2}{p_a B} + \frac{2(1 - p_a)L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\
& + \frac{2(1 - p_a)b^2}{p_{\text{mega}}p_a} \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2, \quad \forall i \in [n],
\end{aligned}$$

882 and

$$\mathbb{E}_k \left[ \|k_{i,2}^{t+1}\|^2 \right] \leq \left( \frac{L_\sigma^2}{B} + L_i^2 \right) \|x^{t+1} - x^t\|^2, \quad \forall i \in [n],$$

883 *Proof.* First, we prove the bound for  $\mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right]$ . Using

$$\mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ h_{i,1}^{t+1} \right] \right]$$

$$\begin{aligned}
&= h_i^t + E_k \left[ \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) - \frac{b}{p_{\text{mega}}} \left( h_i^t - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) \right) \right] \\
&= h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}} (h_i^t - \nabla f_i(x^t))
\end{aligned}$$

884 and

$$\begin{aligned}
&E_k [E_{p_a} [h_{i,2}^{t+1}]] \\
&= h_i^t + E_k \left[ \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1}) \right] \\
&= h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t),
\end{aligned}$$

885 we have

$$\begin{aligned}
&E_k [E_{p_a} [E_{p_{\text{mega}}} [\|h^{t+1} - \nabla f(x^{t+1})\|^2]]] \\
&= p_{\text{mega}} E_k [E_{p_a} [\|h_1^{t+1} - \nabla f(x^{t+1})\|^2]] + (1 - p_{\text{mega}}) E_k [E_{p_a} [\|h_2^{t+1} - \nabla f(x^{t+1})\|^2]] \\
&\stackrel{(17)}{=} p_{\text{mega}} E_k [E_{p_a} [\|h_1^{t+1} - E_k [E_{p_a} [h_1^{t+1}]]\|^2]] + (1 - p_{\text{mega}}) E_k [E_{p_a} [\|h_2^{t+1} - E_k [E_{p_a} [h_2^{t+1}]]\|^2]] \\
&\quad + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2.
\end{aligned}$$

886 We can use Lemma 1 two times with i)  $r_i = h_i^t$  and  $s_i = k_{i,1}^{t+1}$  and ii)  $r_i = h_i^t$  and  $s_i = k_{i,2}^{t+1}$ , to  
887 obtain

$$\begin{aligned}
&E_k [E_{p_a} [E_{p_{\text{mega}}} [\|h^{t+1} - \nabla f(x^{t+1})\|^2]]] \\
&\leq p_{\text{mega}} \left( \frac{1}{n^2 p_a} \sum_{i=1}^n E_k [\|k_{i,1}^{t+1} - E_k [k_{i,1}^{t+1}]\|^2] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \right) \\
&\quad + (1 - p_{\text{mega}}) \left( \frac{1}{n^2 p_a} \sum_{i=1}^n E_k [\|k_{i,2}^{t+1} - E_k [k_{i,2}^{t+1}]\|^2] + \frac{p_a - p_{aa}}{n^2 p_a^2} \sum_{i=1}^n \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \right) \\
&\quad + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2 \\
&\stackrel{(16)}{\leq} \frac{p_{\text{mega}}}{n^2 p_a} \sum_{i=1}^n E_k [\|k_{i,1}^{t+1} - E_k [k_{i,1}^{t+1}]\|^2] \\
&\quad + \frac{1 - p_{\text{mega}}}{n^2 p_a} \sum_{i=1}^n E_k [\|k_{i,2}^{t+1} - E_k [k_{i,2}^{t+1}]\|^2] \\
&\quad + \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
&\quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2.
\end{aligned} \tag{36}$$

888 Let us consider  $E_k [\|k_{i,1}^{t+1} - E_k [k_{i,1}^{t+1}]\|^2]$ .

$$\begin{aligned}
&E_k [\|k_{i,1}^{t+1} - E_k [k_{i,1}^{t+1}]\|^2] \\
&= E_k \left[ \left\| \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) - \frac{b}{p_{\text{mega}}} \left( h_i^t - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) \right) \right\|^2 \right]
\end{aligned}$$



$$\begin{aligned}
& - \left( \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}} (h_i^t - \nabla f_i(x^t)) \right) \Big\|^2 \Big] \\
& = \mathbb{E}_k \left[ \left\| \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) + \frac{b}{p_{\text{mega}}} \left( \frac{1}{B'} \sum_{k=1}^{B'} \nabla f_i(x^t; \xi_{ik}^{t+1}) \right) \right. \right. \\
& \quad \left. \left. - \left( \nabla f_i(x^{t+1}) - \nabla f_i(x^t) + \frac{b}{p_{\text{mega}}} (\nabla f_i(x^t)) \right) \right\|^2 \right] \\
& = \frac{1}{B'^2} \sum_{k=1}^{B'} \mathbb{E}_k \left[ \left\| \frac{b}{p_{\text{mega}}} (\nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \nabla f_i(x^{t+1})) \right. \right. \\
& \quad \left. \left. + \left( 1 - \frac{b}{p_{\text{mega}}} \right) (\nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \nabla f_i(x^t; \xi_{ik}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t))) \right\|^2 \right],
\end{aligned}$$

889 where we used independence of the mini-batch samples. Using (16), we get

$$\begin{aligned}
& \mathbb{E}_k \left[ \left\| k_{i,1}^{t+1} - \mathbb{E}_k [k_{i,1}^{t+1}] \right\|^2 \right] \\
& \leq \frac{2b^2}{B'^2 p_{\text{mega}}^2} \sum_{k=1}^{B'} \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \nabla f_i(x^{t+1}) \right\|^2 \right] \\
& \quad + \frac{2}{B'^2} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 \sum_{k=1}^{B'} \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \nabla f_i(x^t; \xi_{ik}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right].
\end{aligned}$$

890 Due to Assumptions 5 and 6, we have

$$\mathbb{E}_k \left[ \left\| k_{i,1}^{t+1} - \mathbb{E}_k [k_{i,1}^{t+1}] \right\|^2 \right] \leq \frac{2b^2 \sigma^2}{B' p_{\text{mega}}^2} + \frac{2L_\sigma^2}{B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 \|x^{t+1} - x^t\|^2. \quad (37)$$

891 Next, we estimate the bound for  $\mathbb{E}_k \left[ \left\| k_{i,2}^{t+1} - \mathbb{E}_k [k_{i,2}^{t+1}] \right\|^2 \right]$ .

$$\begin{aligned}
& \mathbb{E}_k \left[ \left\| k_{i,2}^{t+1} - \mathbb{E}_k [k_{i,2}^{t+1}] \right\|^2 \right] \\
& = \mathbb{E}_k \left[ \left\| \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B} \sum_{j=1}^B \nabla f_i(x^t; \xi_{ij}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right] \\
& = \frac{1}{B^2} \sum_{j=1}^B \mathbb{E}_k \left[ \left\| \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \nabla f_i(x^t; \xi_{ij}^{t+1}) - (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) \right\|^2 \right].
\end{aligned}$$

892 Due to Assumptions 6, we have

$$\mathbb{E}_k \left[ \left\| k_{i,2}^{t+1} - \mathbb{E}_k [k_{i,2}^{t+1}] \right\|^2 \right] \leq \frac{L_\sigma^2}{B} \|x^{t+1} - x^t\|^2. \quad (38)$$

893 Plugging (37) and (38) into (36), we obtain

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \left\| h^{t+1} - \nabla f(x^{t+1}) \right\|^2 \right] \right] \right] \\
& \leq \frac{p_{\text{mega}}}{np_a} \left( \frac{2b^2 \sigma^2}{B' p_{\text{mega}}^2} + \frac{2L_\sigma^2}{B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 \|x^{t+1} - x^t\|^2 \right) \\
& \quad + \frac{(1 - p_{\text{mega}}) L_\sigma^2}{np_a B} \|x^{t+1} - x^t\|^2 \\
& \quad + \frac{2(p_a - p_{aa})}{n^2 p_a^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2
\end{aligned}$$

$$+ \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2.$$

894 Using Assumption 3, we get

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] \right] \right] \\ & \leq \frac{2b^2 \sigma^2}{n p_{\text{mega}} p_a B'} + \left( \frac{2p_{\text{mega}} L_\sigma^2}{n p_a B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1 - p_{\text{mega}}) L_\sigma^2}{n p_a B} + \frac{2(p_a - p_{aa}) \widehat{L}^2}{n p_a^2} \right) \|x^{t+1} - x^t\|^2 \\ & \quad + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2. \end{aligned}$$

895 Using almost the same derivations, we can prove the second inequality:

$$\begin{aligned} & \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\ & = p_{\text{mega}} \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_{i,1}^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] + (1 - p_{\text{mega}}) \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_{i,2}^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \\ & \stackrel{(17)}{=} p_{\text{mega}} \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_{i,1}^{t+1} - \mathbb{E}_k [h_{i,1}^{t+1}]\|^2 \right] \right] + (1 - p_{\text{mega}}) \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \|h_{i,2}^{t+1} - \mathbb{E}_k [h_{i,2}^{t+1}]\|^2 \right] \right] \\ & \quad + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\ & = p_{\text{mega}} p_a \mathbb{E}_k \left[ \left\| h_i^t + \frac{1}{p_a} k_{i,1}^{t+1} - (h_i^t + \mathbb{E}_k [k_{i,1}^{t+1}]) \right\|^2 \right] \\ & \quad + p_{\text{mega}} (1 - p_a) \|h_i^t - (h_i^t + \mathbb{E}_k [k_{i,1}^{t+1}])\|^2 \\ & \quad + (1 - p_{\text{mega}}) p_a \mathbb{E}_k \left[ \left\| h_i^t + \frac{1}{p_a} k_{i,2}^{t+1} - (h_i^t + \mathbb{E}_k [k_{i,2}^{t+1}]) \right\|^2 \right] \\ & \quad + (1 - p_{\text{mega}}) (1 - p_a) \|h_i^t - (h_i^t + \mathbb{E}_k [k_{i,2}^{t+1}])\|^2 \\ & \quad + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\ & = p_{\text{mega}} p_a \mathbb{E}_k \left[ \left\| \frac{1}{p_a} k_{i,1}^{t+1} - \mathbb{E}_k [k_{i,1}^{t+1}] \right\|^2 \right] \\ & \quad + p_{\text{mega}} (1 - p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\ & \quad + (1 - p_{\text{mega}}) p_a \mathbb{E}_k \left[ \left\| \frac{1}{p_a} k_{i,2}^{t+1} - \mathbb{E}_k [k_{i,2}^{t+1}] \right\|^2 \right] \\ & \quad + (1 - p_{\text{mega}}) (1 - p_a) \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \\ & \quad + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\ & \stackrel{(17)}{=} \frac{p_{\text{mega}}}{p_a} \mathbb{E}_k \left[ \|k_{i,1}^{t+1} - \mathbb{E}_k [k_{i,1}^{t+1}]\|^2 \right] \\ & \quad + \frac{(1 - p_{\text{mega}})}{p_a} \mathbb{E}_k \left[ \|k_{i,2}^{t+1} - \mathbb{E}_k [k_{i,2}^{t+1}]\|^2 \right] \\ & \quad + \frac{p_{\text{mega}} (1 - p_a)}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}} (h_i^t - \nabla f_i(x^t)) \right\|^2 \\ & \quad + \frac{(1 - p_{\text{mega}}) (1 - p_a)}{p_a} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 \end{aligned}$$

$$\begin{aligned}
& + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2 \\
& \stackrel{(16)}{\leq} \frac{p_{\text{mega}}}{p_a} \mathbb{E}_k \left[ \|k_{i,1}^{t+1} - \mathbb{E}_k[k_{i,1}^{t+1}]\|^2 \right] \\
& + \frac{(1 - p_{\text{mega}})}{p_a} \mathbb{E}_k \left[ \|k_{i,2}^{t+1} - \mathbb{E}_k[k_{i,2}^{t+1}]\|^2 \right] \\
& + \frac{2(1 - p_a)}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
& + \frac{2(1 - p_a)b^2}{p_{\text{mega}}p_a} \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

896 Using (37) and (38), we get

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\
& \leq \frac{2b^2\sigma^2}{p_ap_{\text{mega}}B'} + \frac{2p_{\text{mega}}L_\sigma^2}{p_aB'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 \|x^{t+1} - x^t\|^2 \\
& + \frac{(1 - p_{\text{mega}})L_\sigma^2}{p_aB} \|x^{t+1} - x^t\|^2 \\
& + \frac{2(1 - p_a)}{p_a} \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
& + \frac{2(1 - p_a)b^2}{p_{\text{mega}}p_a} \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

897 Next, due to Assumption 3, we obtain

$$\begin{aligned}
& \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \\
& \leq \frac{2b^2\sigma^2}{p_ap_{\text{mega}}B'} + \left( \frac{2p_{\text{mega}}L_\sigma^2}{p_aB'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1 - p_{\text{mega}})L_\sigma^2}{p_aB} + \frac{2(1 - p_a)L_i^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \\
& + \frac{2(1 - p_a)b^2}{p_{\text{mega}}p_a} \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1 - p_{\text{mega}}) \right) \|h_i^t - \nabla f_i(x^t)\|^2.
\end{aligned}$$

898 The third inequality can be proved with the help of (38) and Assumption 3.

$$\begin{aligned}
& \mathbb{E}_k \left[ \|k_{i,2}^{t+1}\|^2 \right] \\
& \stackrel{(17)}{=} \mathbb{E}_k \left[ \|k_{i,2}^{t+1} - \mathbb{E}_k[k_{i,2}^{t+1}]\|^2 \right] + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
& \leq \frac{L_\sigma^2}{B} \|x^{t+1} - x^t\|^2 + \|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\|^2 \\
& \leq \left( \frac{L_\sigma^2}{B} + L_i^2 \right) \|x^{t+1} - x^t\|^2.
\end{aligned}$$

899

□

**Theorem 11.** Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_{\text{mega}}p_a}{2-p_a}$ , probability  $p_{\text{mega}} \in (0, 1]$ , batch size  $B' \geq B \geq 1$

$$\gamma \leq \left( L + \sqrt{\frac{8(2\omega+1)\omega}{np_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right) + \frac{16}{np_{\text{mega}}p_a^2} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{L_\sigma^2}{B} \right)} \right)^{-1},$$

900 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 8. Then

$$\mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] \leq \frac{1}{T} \left[ \frac{2\Delta_0}{\gamma} + \frac{4}{p_{\text{mega}}p_a} \|h^0 - \nabla f(x^0)\|^2 + \frac{4 \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_{\text{mega}}p_a} \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right]$$

$$+ \frac{12\sigma^2}{nB'}.$$

901 *Proof.* Due to Lemma 2 and the update step from Line 4 in Algorithm 8, we have

$$\begin{aligned} & \mathbb{E}_{t+1} [f(x^{t+1})] \\ & \leq \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - \nabla f(x^t)\|^2 \right] \\ & = \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \frac{\gamma}{2} \|g^t - h^t + h^t - \nabla f(x^t)\|^2 \right] \\ & \stackrel{(17)}{\leq} \mathbb{E}_{t+1} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right]. \end{aligned}$$

902 Let us fix constants  $\kappa, \eta, \nu, \rho \in [0, \infty)$  that we will define later. Considering Lemma 13, Lemma 14,  
903 and the law of total expectation, we obtain

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & \quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right] \\ & \quad + \kappa \mathbb{E} \left[ \mathbb{E}_k \left[ \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} [\|g^{t+1} - h^{t+1}\|^2] \right] \right] \right] \right] \\ & \quad + \eta \mathbb{E} \left[ \mathbb{E}_k \left[ \mathbb{E}_C \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \right] \right] \right] \right] \\ & \quad + \nu \mathbb{E} \left[ \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] \right] \right] \right] \\ & \quad + \rho \mathbb{E} \left[ \mathbb{E}_k \left[ \mathbb{E}_{p_a} \left[ \mathbb{E}_{p_{\text{mega}}} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \right] \right] \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right] \\ & \quad + \kappa \mathbb{E} \left( \frac{2(1-p_{\text{mega}})\omega}{np_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \left( \frac{(p_a - p_{aa})a^2}{n^2 p_a^2} + \frac{2(1-p_{\text{mega}})a^2\omega}{n^2 p_a} \right) \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \right) \\ & \quad + \eta \mathbb{E} \left( \frac{2(1-p_{\text{mega}})\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \left( \frac{(1-p_a)a^2}{p_a} + \frac{2(1-p_{\text{mega}})a^2\omega}{p_a} \right) \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \right) \\ & \quad + \nu \mathbb{E} \left( \frac{2b^2\sigma^2}{np_{\text{mega}}p_a B'} + \left( \frac{2p_{\text{mega}}L_\sigma^2}{np_a B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1-p_{\text{mega}})L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2 p_a^2 p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + (1-p_{\text{mega}}) \right) \|h^t - \nabla f(x^t)\|^2 \right) \\ & \quad + \rho \mathbb{E} \left( \frac{2b^2\sigma^2}{p_a p_{\text{mega}} B'} + \left( \frac{2p_{\text{mega}}L_\sigma^2}{p_a B'} \left( 1 - \frac{b}{p_{\text{mega}}} \right)^2 + \frac{(1-p_{\text{mega}})L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right) \end{aligned}$$

$$+ \frac{2(1-p_a)b^2}{np_{\text{mega}}p_a} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + \left( p_{\text{mega}} \left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}}) \right) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \Bigg).$$

Let us simplify the last inequality. Since  $B' \geq B$  and  $b = \frac{p_{\text{mega}}p_a}{2-p_a} \leq p_{\text{mega}}$ , we have  $1 - p_{\text{mega}} \leq 1$ ,

$$\frac{2p_{\text{mega}}L_\sigma^2}{p_aB'} \left(1 - \frac{b}{p_{\text{mega}}}\right)^2 \leq \frac{2p_{\text{mega}}L_\sigma^2}{p_aB},$$

$$\left( p_{\text{mega}} \left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}}) \right) \leq 1 - b,$$

and

$$\left( \frac{2(1-p_a)b^2}{p_{\text{mega}}p_a} + p_{\text{mega}} \left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}}) \right) \leq 1 - b.$$

904 Thus

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} \left[ f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|x^{t+1} - x^t\|^2 + \gamma (\|g^t - h^t\|^2 + \|h^t - \nabla f(x^t)\|^2) \right] \\ & + \kappa \mathbb{E} \left( \frac{2\omega}{np_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \frac{((2\omega + 1)p_a - p_{aa})a^2}{n^2p_a^2} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \right) \\ & + \eta \mathbb{E} \left( \frac{2\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \frac{(2\omega + 1 - p_a)a^2}{p_a} \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 + (1-a)^2 \|g^t - h^t\|^2 \right) \\ & + \nu \mathbb{E} \left( \frac{2b^2\sigma^2}{np_{\text{mega}}p_aB'} + \left( \frac{2L_\sigma^2}{np_aB} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + \frac{2(p_a - p_{aa})b^2}{n^2p_a^2p_{\text{mega}}} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 + (1-b) \|h^t - \nabla f(x^t)\|^2 \right) \\ & + \rho \mathbb{E} \left( \frac{2b^2\sigma^2}{p_ap_{\text{mega}}B'} + \left( \frac{2L_\sigma^2}{p_aB} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \|x^{t+1} - x^t\|^2 \right. \\ & \quad \left. + (1-b) \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right). \end{aligned}$$

905 After rearranging the terms, we get

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{2\kappa\omega}{np_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \frac{2\eta\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + (\gamma + \kappa(1-a)^2) \mathbb{E} [\|g^t - h^t\|^2] \\
&\quad + \left( \kappa \frac{((2\omega+1)p_a - p_{aa})a^2}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad + (\gamma + \nu(1-b)) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'}.
\end{aligned}$$

906 Let us take  $\kappa = \frac{\gamma}{a}$ , thus  $\gamma + \kappa(1-a)^2 \leq \kappa$  and

$$\begin{aligned}
&\mathbb{E} [f(x^{t+1})] + \frac{\gamma}{a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
&\quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{2\gamma\omega}{anp_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \frac{2\eta\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma}{a} \mathbb{E} [\|g^t - h^t\|^2] \\
&\quad + \left( \frac{\gamma((2\omega+1)p_a - p_{aa})a}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad + (\gamma + \nu(1-b)) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'}.
\end{aligned}$$

907 Next, since  $a = \frac{p_a}{2\omega+1}$ , we have  $\left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \leq 1-a$ . We the choice  $\eta =$

908  $\frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2}$ , we guarantee  $\frac{\gamma((2\omega+1)p_a - p_{aa})a}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \leq \eta$  and

$$\begin{aligned}
&\mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
&\quad + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{2\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \frac{2\gamma((2\omega+1)p_a - p_{aa})\omega}{np_a^3} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad + (\gamma + \nu(1-b)) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'} \\
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad + (\gamma + \nu(1-b)) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
&\quad + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
&\quad + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'},
\end{aligned}$$

909 where simplified the term using  $p_{aa} \geq 0$ . Let us take  $\nu = \frac{\gamma}{b}$  to obtain

$$\begin{aligned}
&\mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
&\quad + \frac{\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
&\leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
&\quad - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
&\quad \left. - \left( \frac{2\gamma L_\sigma^2}{bn p_a B} + \frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bn p_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
&\quad + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \frac{\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&\quad + \frac{\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2]
\end{aligned}$$

$$\begin{aligned}
& + \left( \frac{2\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{2\gamma b}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'}.
\end{aligned}$$

910 Next, we take  $\rho = \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}}$ , thus

$$\begin{aligned}
& \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma(2\omega + 1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{2\gamma L_\sigma^2}{bn p_a B} + \frac{2\gamma(p_a - p_{aa})\widehat{L}^2}{bn p_a^2} \right) - \left( \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \right) \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{2\gamma b}{np_{\text{mega}} p_a} + \frac{4\gamma(p_a - p_{aa})b^2}{np_a^3 p_{\text{mega}}} \right) \frac{\sigma^2}{B'}.
\end{aligned}$$

911 Since  $\frac{p_{\text{mega}} p_a}{2} \leq b \leq p_{\text{mega}} p_a$  and  $1 - p_a \leq 1 - \frac{p_{aa}}{p_a} \leq 1$ , we get

$$\begin{aligned}
& \mathbb{E}[f(x^{t+1})] + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^{t+1} - h^{t+1}\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma(2\omega + 1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{4\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{4\gamma(p_a - p_{aa})\widehat{L}^2}{np_{\text{mega}} p_a^3} \right) - \left( \frac{4\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{4\gamma(1-p_a)\widehat{L}^2}{np_{\text{mega}} p_a^2} \right) \right) \mathbb{E}[\|x^{t+1} - x^t\|^2] \\
& + \frac{\gamma(2\omega + 1)}{p_a} \mathbb{E}[\|g^t - h^t\|^2] + \frac{\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E}[\|h^t - \nabla f(x^t)\|^2] + \frac{2\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{6\gamma\sigma^2}{nB'} \\
& \leq \mathbb{E}[f(x^t)] - \frac{\gamma}{2} \mathbb{E}[\|\nabla f(x^t)\|^2]
\end{aligned}$$



$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \left( \frac{8\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{8\gamma \left(1 - \frac{p_{\text{aa}}}{p_a}\right) \widehat{L}^2}{np_{\text{mega}} p_a^2} \right) \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{6\gamma\sigma^2}{nB'}.
\end{aligned}$$

912 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& + \frac{\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{\gamma((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \frac{\gamma}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \frac{2\gamma(p_a - p_{\text{aa}})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{6\gamma\sigma^2}{nB'}.
\end{aligned}$$

913 It is left to apply Lemma 3 with

$$\begin{aligned}
\Psi^t &= \frac{(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{((2\omega+1)p_a - p_{\text{aa}})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
&+ \frac{1}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \frac{2 \left(1 - \frac{p_{\text{aa}}}{p_a}\right)}{np_a p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]
\end{aligned}$$

914 and  $C = \frac{6\sigma^2}{nB'}$  to conclude the proof.  $\square$

**Corollary 6.** Suppose that assumptions from Theorem 11 hold, probability  $p_{\text{mega}} = \min \left\{ \frac{\zeta_c}{d}, \frac{n\varepsilon B}{\sigma^2} \right\}$ , batch size  $B' = \Theta \left( \frac{\sigma^2}{n\varepsilon} \right)$ , and  $h_i^0 = g_i^0 = \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0)$  for all  $i \in [n]$ , initial batch size  $B_{\text{init}} = \Theta \left( \frac{B}{p_{\text{mega}} \sqrt{p_a}} \right) = \Theta \left( \max \left\{ \frac{Bd}{\sqrt{p_a} \zeta_c}, \frac{\sigma^2}{\sqrt{p_a} n \varepsilon} \right\} \right)$ , then DASHA-PP-SYNC-MVR needs

$$T := \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \left( \frac{\omega}{p_a \sqrt{n}} + \sqrt{\frac{d}{p_a^2 \zeta_c n}} \right) \left( \widehat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon} n} \left( \frac{\widehat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right] + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right).$$

915 communication rounds to get an  $\varepsilon$ -solution, the expected communication complexity is equal to  
916  $\mathcal{O}(d + \zeta_c T)$ , and the expected number of stochastic gradient calculations per node equals  $\mathcal{O}(B_{\text{init}} +$   
917  $BT)$ , where  $\zeta_c$  is the expected density from Definition 12.

918 *Proof.* Due to the choice of  $B'$ , we have

$$\begin{aligned} \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] &\leq \frac{1}{T} \left[ 2\Delta_0 \left( L + \sqrt{\frac{8(2\omega+1)\omega}{np_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right) + \frac{16}{np_{\text{mega}}p_a^2} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{L_\sigma^2}{B} \right)} \right) \right. \\ &\quad \left. + \frac{4}{p_{\text{mega}}p_a} \|h^0 - \nabla f(x^0)\|^2 + \frac{4 \left( 1 - \frac{p_{aa}}{p_a} \right)}{np_{\text{mega}}p_a} \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right] \\ &\quad + \frac{2\varepsilon}{3}. \end{aligned}$$

919 Using

$$\mathbb{E} \left[ \|h^0 - \nabla f(x^0)\|^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0) - \nabla f(x^0) \right\|^2 \right] \leq \frac{\sigma^2}{nB_{\text{init}}}$$

920 and

$$\frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \|h_i^0 - \nabla f_i(x^0)\|^2 \right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[ \left\| \frac{1}{B_{\text{init}}} \sum_{k=1}^{B_{\text{init}}} \nabla f_i(x^0; \xi_{ik}^0) - \nabla f_i(x^0) \right\|^2 \right] \leq \frac{\sigma^2}{nB_{\text{init}}},$$

921 we have

$$\begin{aligned} \mathbb{E} \left[ \|\nabla f(\hat{x}^T)\|^2 \right] &\leq \frac{1}{T} \left[ 2\Delta_0 \left( L + \sqrt{\frac{8(2\omega+1)\omega}{np_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right) + \frac{16}{np_{\text{mega}}p_a^2} \left( \left( 1 - \frac{p_{aa}}{p_a} \right) \hat{L}^2 + \frac{L_\sigma^2}{B} \right)} \right) \right. \\ &\quad \left. + \frac{8\sigma^2}{np_{\text{mega}}p_a B_{\text{init}}} \right] \\ &\quad + \frac{2\varepsilon}{3}. \end{aligned}$$

922 Therefore, we can take the following  $T$  to get  $\varepsilon$ -solution.

$$T = \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \sqrt{\frac{\omega^2}{np_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right) + \frac{1}{np_{\text{mega}}p_a^2} \left( \hat{L}^2 + \frac{L_\sigma^2}{B} \right)} \right) + \frac{\sigma^2}{np_{\text{mega}}p_a B_{\text{init}}} \right] \right)$$

923 Considering the choice of  $p_{\text{mega}}$  and  $B_{\text{init}}$ , we obtain

$$\begin{aligned} T &= \mathcal{O} \left( \frac{1}{\varepsilon} \left[ \Delta_0 \left( L + \left( \frac{\omega}{p_a \sqrt{n}} + \sqrt{\frac{d}{p_a^2 \zeta_C n}} \right) \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon n}} \left( \frac{\hat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right) + \frac{\sigma^2}{np_{\text{mega}}p_a B_{\text{init}}} \right] \right) \\ &= \mathcal{O} \left( \frac{\Delta_0}{\varepsilon} \left[ L + \left( \frac{\omega}{p_a \sqrt{n}} + \sqrt{\frac{d}{p_a^2 \zeta_C n}} \right) \left( \hat{L} + \frac{L_\sigma}{\sqrt{B}} \right) + \frac{\sigma}{p_a \sqrt{\varepsilon n}} \left( \frac{\hat{L}}{\sqrt{B}} + \frac{L_\sigma}{B} \right) \right] + \frac{\sigma^2}{\sqrt{p_a} n \varepsilon B} \right). \end{aligned}$$

924 The expected communication complexity equals  $\mathcal{O}(d + p_{\text{mega}}d + (1 - p_{\text{mega}})\zeta_C) =$   
 925  $\mathcal{O}(d + \zeta_C)$  and the expected number of stochastic gradient calculations per node equals  
 926  $\mathcal{O}(B_{\text{init}} + p_{\text{mega}}B' + (1 - p_{\text{mega}})B) = \mathcal{O}(B_{\text{init}} + B)$ .  $\square$

**Theorem 13.** Suppose that Assumptions 1, 2, 3, 5, 6, 7, 8 and 9 hold. Let us take  $a = \frac{p_a}{2\omega+1}$ ,  $b = \frac{p_{\text{mega}}p_a}{2-p_a}$ , probability  $p_{\text{mega}} \in (0, 1]$ , batch size  $B' \geq B \geq 1$ ,

$$\gamma \leq \min \left\{ \left( L + \sqrt{\frac{16(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) + \left( \frac{48L_\sigma^2}{np_{\text{mega}}p_a^2B} + \frac{24 \left( 1 - \frac{p_{aa}}{p_a} \right) \widehat{L}^2}{np_{\text{mega}}p_a^2} \right)} \right)^{-1}, \frac{a}{2\mu}, \frac{b}{2\mu} \right\},$$

927 and  $h_i^0 = g_i^0$  for all  $i \in [n]$  in Algorithm 8. Then

$$\begin{aligned} & \mathbb{E} [f(x^T) - f^*] \\ & \leq (1 - \gamma\mu)^T \left( \Delta_0 + \frac{2\gamma}{b} \|h^0 - \nabla f(x^0)\|^2 + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \frac{1}{n} \sum_{i=1}^n \|h_i^0 - \nabla f_i(x^0)\|^2 \right) + \frac{20\sigma^2}{\mu n B'}. \end{aligned}$$

928 *Proof.* Let us fix constants  $\kappa, \eta, \nu, \rho \in [0, \infty)$  that we will define later. As in the proof of Theorem 11,  
929 we can get

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \kappa \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\ & - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{2\kappa\omega}{np_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \frac{2\eta\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\ & \quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\ & + (\gamma + \kappa(1-a)^2) \mathbb{E} [\|g^t - h^t\|^2] \\ & + \left( \kappa \frac{((2\omega+1)p_a - p_{aa})a^2}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ & + (\gamma + \nu(1-b)) \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\ & + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\ & + \left( \frac{2\nu b^2}{np_{\text{mega}}p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'}. \end{aligned}$$

930 Let us take  $\kappa = \frac{2\gamma}{a}$ , thus  $\gamma + \kappa(1-a)^2 \leq (1 - \frac{a}{2})\kappa$  and

$$\begin{aligned} & \mathbb{E} [f(x^{t+1})] + \frac{2\gamma}{a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \eta \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\ & + \nu \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\ & \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\ & - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega}{anp_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \frac{2\eta\omega}{p_a} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \end{aligned}$$

$$\begin{aligned}
& -\nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma}{a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] \\
& + \left( \frac{2\gamma((2\omega+1)p_a - p_{aa})a}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + (\gamma + \nu(1-b)) \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'}.
\end{aligned}$$

931 Next, since  $a = \frac{p_a}{2\omega+1}$ , we have  $\left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \leq 1-a$ . We the choice  $\eta =$   
932  $\frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2}$ , we guarantee  $\frac{\gamma((2\omega+1)p_a - p_{aa})a}{np_a^2} + \eta \left( \frac{(2\omega+1-p_a)a^2}{p_a} + (1-a)^2 \right) \leq (1-\frac{a}{2})\eta$  and

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \nu \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \nu \left( \frac{2L_\sigma^2}{np_a B} + \frac{2(p_a - p_{aa})\widehat{L}^2}{np_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} \left[ \|x^{t+1} - x^t\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] \\
& + \left( 1 - \frac{a}{2} \right) \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + (\gamma + \nu(1-b)) \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] \\
& + \left( \nu \frac{2(p_a - p_{aa})b^2}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{2\nu b^2}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'},
\end{aligned}$$

933 where simplified the term using  $p_{aa} \geq 0$ . Let us take  $\nu = \frac{2\gamma}{b}$  to obtain

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} \left[ \|g^{t+1} - h^{t+1}\|^2 \right] + \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} \left[ \|h^{t+1} - \nabla f(x^{t+1})\|^2 \right] + \rho \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla f(x^t)\|^2 \right]
\end{aligned}$$

$$\begin{aligned}
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{4\gamma L_\sigma^2}{bn p_a B} + \frac{4\gamma(p_a - p_{aa})\widehat{L}^2}{bn p_a^2} \right) - \rho \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{a}{2}\right) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left(1 - \frac{a}{2}\right) \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] \\
& + \left( \frac{4\gamma(p_a - p_{aa})b}{np_a^2 p_{\text{mega}}} + \rho(1-b) \right) \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{4\gamma b}{np_{\text{mega}} p_a} + \frac{2\rho b^2}{p_a p_{\text{mega}}} \right) \frac{\sigma^2}{B'},
\end{aligned}$$

934 Next, we take  $\rho = \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}}$ , thus

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{4\gamma L_\sigma^2}{bn p_a B} + \frac{4\gamma(p_a - p_{aa})\widehat{L}^2}{bn p_a^2} \right) - \left( \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \right) \left( \frac{2L_\sigma^2}{p_a B} + \frac{2(1-p_a)\widehat{L}^2}{p_a} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{a}{2}\right) \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left(1 - \frac{a}{2}\right) \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \left( \frac{4\gamma b}{np_{\text{mega}} p_a} + \frac{16\gamma(p_a - p_{aa})b^2}{np_a^3 p_{\text{mega}}^2} \right) \frac{\sigma^2}{B'},
\end{aligned}$$

935 Since  $\frac{p_{\text{mega}} p_a}{2} \leq b \leq p_{\text{mega}} p_a$  and  $1 - p_a \leq 1 - \frac{p_{aa}}{p_a} \leq 1$ , we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega+1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{2\gamma((2\omega+1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma(2\omega+1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) \right. \\
& \quad \left. - \left( \frac{8\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{8\gamma(p_a - p_{aa})\widehat{L}^2}{np_{\text{mega}} p_a^3} \right) - \left( \frac{16\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{16\gamma(1-p_a)\widehat{L}^2}{np_{\text{mega}} p_a^2} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2]
\end{aligned}$$

$$\begin{aligned}
& + \left(1 - \frac{a}{2}\right) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left(1 - \frac{a}{2}\right) \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{20\gamma\sigma^2}{nB'} \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& - \left( \frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma(2\omega + 1)\omega}{np_a^2} \left( \frac{L_\sigma^2}{B} + \widehat{L}^2 \right) - \left( \frac{24\gamma L_\sigma^2}{np_{\text{mega}} p_a^2 B} + \frac{24\gamma \left(1 - \frac{p_{aa}}{p_a}\right) \widehat{L}^2}{np_{\text{mega}} p_a^2} \right) \right) \mathbb{E} [\|x^{t+1} - x^t\|^2] \\
& + \left(1 - \frac{a}{2}\right) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left(1 - \frac{a}{2}\right) \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{20\gamma\sigma^2}{nB'}.
\end{aligned}$$

936 Using Lemma 4 and the assumption about  $\gamma$ , we get

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + \left(1 - \frac{a}{2}\right) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + \left(1 - \frac{a}{2}\right) \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + \left(1 - \frac{b}{2}\right) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + \left(1 - \frac{b}{2}\right) \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{20\gamma\sigma^2}{nB'}.
\end{aligned}$$

937 Due to  $\gamma \leq \frac{a}{2\mu}$  and  $\gamma \leq \frac{b}{2\mu}$ , we have

$$\begin{aligned}
& \mathbb{E} [f(x^{t+1})] + \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^{t+1} - h^{t+1}\|^2] + \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^{t+1} - h_i^{t+1}\|^2 \right] \\
& + \frac{2\gamma}{b} \mathbb{E} [\|h^{t+1} - \nabla f(x^{t+1})\|^2] + \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^{t+1} - \nabla f_i(x^{t+1})\|^2 \right] \\
& \leq \mathbb{E} [f(x^t)] - \frac{\gamma}{2} \mathbb{E} [\|\nabla f(x^t)\|^2] \\
& + (1 - \gamma\mu) \frac{2\gamma(2\omega + 1)}{p_a} \mathbb{E} [\|g^t - h^t\|^2] + (1 - \gamma\mu) \frac{2\gamma((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\
& + (1 - \gamma\mu) \frac{2\gamma}{b} \mathbb{E} [\|h^t - \nabla f(x^t)\|^2] + (1 - \gamma\mu) \frac{8\gamma(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right] \\
& + \frac{20\gamma\sigma^2}{nB'}.
\end{aligned}$$

938 It is left to apply Lemma 11 with

$$\begin{aligned}\Psi^t &= \frac{2(2\omega + 1)}{p_a} \mathbb{E} \left[ \|g^t - h^t\|^2 \right] + \frac{2((2\omega + 1)p_a - p_{aa})}{np_a^2} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|g_i^t - h_i^t\|^2 \right] \\ &+ \frac{2}{b} \mathbb{E} \left[ \|h^t - \nabla f(x^t)\|^2 \right] + \frac{8(p_a - p_{aa})}{np_a^2 p_{\text{mega}}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|h_i^t - \nabla f_i(x^t)\|^2 \right]\end{aligned}$$

939 and  $C = \frac{20\sigma^2}{nB'}$  to conclude the proof. □