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

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

2 Optimization Problem

Let us consider the nonconvex distributed optimization problem

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

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

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

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

We consider three settings:

1. **Gradient Setting.** The i^{th} node has only access to the gradient $\nabla f_i : \mathbb{R}^d \to \mathbb{R}^d$ 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 $\widehat{L}^2 := \frac{1}{n}\sum_{i=1}^n L_i^2$.*[2]

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

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

where $f_{ij} : \mathbb{R}^d \to \mathbb{R}$ 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,

$$f_i(x) = \mathrm{E}_\xi\left[f_i(x;\xi)\right], \qquad \forall i \in [n], \tag{3}$$

where $f_i : \mathbb{R}^d \times \Omega_\xi \to \mathbb{R}$. 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^{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 ξ_{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., $\mathrm{E}_\xi\left[\nabla f_i(x;\xi)\right] = \nabla f_i(x)$, and $\mathrm{E}_\xi\left[\|\nabla f_i(x;\xi) - \nabla f_i(x)\|^2\right] \leq \sigma^2$, 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., $\mathrm{E}_\xi\left[\|\nabla f_i(x;\xi) - \nabla f_i(y;\xi)\|^2\right] \leq L_\sigma^2\,\|x - y\|^2$.*

We compare algorithms using *the oracle complexity*, i.e., the number of (stochastic) gradients that each node has to calculate to get ε-solution, and *the communication complexity*, i.e., the number of bits that each node has to send to the server to get ε-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 $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is an *unbiased compressor* if there exists $\omega \in \mathbb{R}$ such that

$$\mathrm{E}\left[\mathcal{C}(x)\right] = x, \qquad \mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega\,\|x\|^2, \tag{4}$$

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

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 unbiassed

[1] 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.

[2] Note that $L \leq \widehat{L}$, $\widehat{L} \leq L_{\max}$, and $\widehat{L} \leq L_\sigma$.

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 $\mathcal{O}\left(\frac{\omega}{\sqrt{n}\varepsilon}\right)$?

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

[a] 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.
[b] On average, MARINA provides the compressed communication mechanism with complexity $\mathcal{O}\left(\frac{\omega}{\sqrt{n}\varepsilon}\right)$. However, with a small probability, this method sends non-compressed vectors.
[c] 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.
[d] 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.
[e] 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 ε-solution, while DASHA-PP requires $\mathcal{O}\left(\frac{\sqrt{n}}{\varepsilon s}\right)$ communication rounds (see Theorem 4 with $\omega = 0$, $b = \frac{p_{\text{a}}}{2 - p_{\text{a}}}$, and $p_{\text{a}} = \frac{s}{n}$).

compressors, including Rand$K$ (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.*

2.2 Nodes Partial Participation Assumptions

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

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

1. $$\textbf{Prob}\left(i^{\text{th}} \text{ node is } \textit{participating}\right) = p_{\text{a}} \quad \forall i \in [n],$$

2. $$\textbf{Prob}\left(i^{\text{th}} \text{ and } j^{\text{th}} \text{ nodes are } \textit{participating}\right) = p_{\text{aa}} \quad \forall i \neq j \in [n].$$

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

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

3.
$$p_{\mathrm{aa}} \le p_{\mathrm{a}}^2, \tag{5}$$

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 ($p_{\mathrm{a}} = {}^{s}/_{n}$ and $p_{\mathrm{aa}} = {}^{s(s-1)}/_{n(n-1)}$), and independent participation, where each node independently participates with probability p_{a} (due to independence, we have $p_{\mathrm{aa}} = p_{\mathrm{a}}^2$), 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 $\mathcal{O}\left({}^{1}/_{\varepsilon}\right)$, 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 ε–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 $\mathcal{O}\left(1/\varepsilon^{3/2}\right)$ while, in the full participation setting, it is possible to get the complexity $\mathcal{O}\left(1/\varepsilon\right)$.

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.

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

$$g_i^{t+1} = g_i^t + \mathcal{C}_i\left(\nabla f_i(x^{t+1}) - (1-a)\nabla f_i(x^t) - a g_i^t\right).$$

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 $p_{\text{a}} \in (0, 1]$ that a node is *participating*[(a)], 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** $t = 0, 1, \ldots, T-1$ **do**
5:     $x^{t+1} = x^t - \gamma g^t$
6:     Broadcast x^{t+1}, x^t to all *participating*[(a)] nodes
7:     **for** $i = 1, \ldots, n$ in parallel **do**
8:         **if** i^{th} node is *participating*[(a)] **then**
9:             Calculate k_i^{t+1} using Algorithm 2, 3, 4 or 5
10:             $h_i^{t+1} = h_i^t + \frac{1}{p_{\text{a}}} k_i^{t+1}$
11:             $m_i^{t+1} = \mathcal{C}_i\left(\frac{1}{p_{\text{a}}} k_i^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right)$
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: $k_i^{t+1} = \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right)$

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: $$k_i^{t+1} = \begin{cases} \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}}\left(h_i^t - \nabla f_i(x^t)\right), \\ \qquad \text{with probability } p_{\text{page}} \text{ on all } \textit{participating} \text{ nodes}, \\ \frac{1}{B}\sum_{j \in I_i^t}\left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right), \\ \qquad \text{with probability } 1 - p_{\text{page}} \text{ on all } \textit{participating} \text{ nodes} \end{cases}$$

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: $$k_{ij}^{t+1} = \begin{cases} \frac{m}{B}\left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b\left(h_{ij}^t - \nabla f_{ij}(x^t)\right)\right), & j \in I_i^t, \\ 0, & j \notin I_i^t \end{cases}$$
3: $h_{ij}^{t+1} = h_{ij}^t + \frac{1}{p_{\text{a}}} k_{ij}^{t+1}$
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: $k_i^{t+1} = \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^t; \xi_{ij}^{t+1}) - b\left(h_i^t - \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^t; \xi_{ij}^{t+1})\right)$

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

$$\mathcal{C}_i^p := \begin{cases} \frac{1}{p}\mathcal{C}_i, & \text{with pr. } p_{\text{a}}, \\ 0, & \text{with pr. } 1 - p_{\text{a}}. \end{cases} \Rightarrow g_i^{t+1} = \begin{cases} g_i^t + \frac{1}{p_{\text{a}}}\mathcal{C}_i\left(\nabla f_i(x^{t+1}) - (1-a)\nabla f_i(x^t) - a g_i^t\right), p_{\text{a}} \\ g_i^t, & \text{with pr. } 1 - p_{\text{a}}. \end{cases}$$

With probability $1 - p$, a node does not update $g_i^{\cdot}$ 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:

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

$$g_i^{t+1} = g_i^t + \mathcal{C}_i\left(h_i^{t+1} - h_i^t - a\left(g_i^t - h_i^t\right)\right). \tag{7}$$

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 $g_i^{\cdot}$ and does not send anything to the server, it still has to update $h_i^{\cdot}$, 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 (DASHA-PP-MVR with the batch size $B = 1$):

$$\begin{aligned} &h_i^{t+1} = h_i^t + \frac{1}{p_{\text{a}}}k_i^{t+1},\ k_i^{t+1} = \nabla f_i(x^{t+1}; \xi_i^{t+1}) - \nabla f_i(x^t; \xi_i^{t+1}) - b\left(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})\right), \\ &g_i^{t+1} = g_i^t + \mathcal{C}_i\left(\frac{1}{p_{\text{a}}}k_i^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right) \text{with } p_{\text{a}},\ \text{and } h_i^{t+1} = h_i^t,\ g_i^{t+1} = g_i^t \text{ with } 1 - p_{\text{a}}. \end{aligned} \tag{8}$$

Now both control variables g_i^t and h_i^t do not change with the probability $1 - p_{\text{a}}$. However, when the i^{th} node participates, the update rule of g_i^{t+1} and h_i^{t+1} was modified to make the proof work. When $p_{\text{a}} = 1$ (no partial participation), the updates rule from (8) reduce to (6) and (7).

The theoretical analysis of the new algorithm became more complicated: unlike (6) and (7), the control variables h_i^{t+1} and g_i^{t+1} in (8) (see also main Algorithm 1) 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. The former lemma does not use the knowledge about the update rule of h_i^{t+1}, works with one expectation $\mathrm{E}_{\mathcal{C}}\left[\cdot\right]$, uses only (4), (12), and (13). The latter lemma additionally requires and uses the structure of the update rule of h_i^{t+1}, surgically copes with the expectations $\mathrm{E}_{\mathcal{C}}\left[\cdot\right]$ and $\mathrm{E}_{p_{\text{a}}}\left[\cdot\right]$ (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 parts of the analysis and the finite-sum setting: the generalization of the previous algorithm and the additional randomness from the partial participation required us to rethink the previous proofs.

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 PŁ-condition (4×2 tracks of 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 ε-solution after T communication rounds. Then, ideally, we would expect the convergence of the new algorithms to ε-solution after up to T/p_{a} communication rounds due to the partial participation constraints[3]. The detailed analysis of the algorithms under Polyak-Łojasiewicz condition we provide in Section F. Let us define $\Delta_0 := f(x^0) - f^*$.

[3]We check this numerically in Section A.

6.1 Gradient Setting

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

$$\gamma \leq \left(L + \left[\frac{48\omega\left(2\omega+1\right)}{np_{\rm a}^2} + \frac{16}{np_{\rm a}^2}\left(1 - \frac{p_{\rm aa}}{p_{\rm a}}\right)\right]^{1/2} \widehat{L}\right)^{-1},$$

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

Let us recall the convergence rate of MARINA or DASHA, the number of communication rounds to get ε-solution equals $\mathcal{O}\left(\frac{\Delta_0}{\varepsilon}\left[L + \frac{\omega}{\sqrt{n}}\widehat{L}\right]\right)$, while the rate of DASHA-PP equals $\mathcal{O}\left(\frac{\Delta_0}{\varepsilon}\left[L + \frac{\omega+1}{p_{\rm a}\sqrt{n}}\widehat{L}\right]\right)$. Up to Lipschitz constants factors, we get the degeneration up to $1/p_{\rm a}$ factor due to the partial participation. This is the expected result since each worker sends useful information only with the probability $p_{\rm a}$.

6.2 Finite-Sum Setting

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

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

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

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

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

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

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

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

In the following corollary, we consider Rand$K$ compressors[4] (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. Indeed, comparing the following result with (Tyurin and Richtárik, 2023)[Corollary 6.6], one can see that we get the degeneration up to $1/p_{\rm a}$ factor, which is expected in the partial participation setting. Note that the complexities improve with the number of workers n.

Corollary 2. *Suppose that assumptions of Corollary 1 hold,* $B \leq \min\left\{\frac{1}{p_{\rm a}}\sqrt{\frac{m}{n}}, \frac{L_{\max}^2}{\mathbb{1}_{p_{\rm a}}^2\widehat{L}^2}\right\}$[5]*, and we use the unbiased compressor Rand$K$ with* $K = \Theta\left(Bd/\sqrt{m}\right)$. *Then the communication complexity of*

[4]The choice of the compressor is driven by simplicity, and the following analysis can be used for other unbiased compressors.

[5]If $\mathbb{1}_{p_{\rm a}} = 0$, then $\frac{L_\sigma^2}{\mathbb{1}_{p_{\rm a}}^2\widehat{L}^2} = +\infty$

Algorithm 1 is $\mathcal{O}\left(d + \frac{L_{\max}\Delta_0 d}{p_a \varepsilon \sqrt{n}}\right)$, *and the expected number of gradient calculations per node equals* $\mathcal{O}\left(m + \frac{L_{\max}\Delta_0\sqrt{m}}{p_a\varepsilon\sqrt{n}}\right)$.

The convergence rate of DASHA-PP-FINITE-MVR is provided in Section E.5.

6.3 Stochastic Setting

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

Theorem 4. *Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take* $a = \frac{p_a}{2\omega+1}$, $b \in \left(0, \frac{p_a}{2-p_a}\right]$, $\gamma \leq \left(L + \left[\frac{48\omega(2\omega+1)}{np_a^2}\left(\widehat{L}^2 + \frac{(1-b)^2L_\sigma^2}{B}\right) + \frac{12}{np_a b}\left(\left(1 - \frac{p_{aa}}{p_a}\right)\widehat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B}\right)\right]^{1/2}\right)^{-1}$, *and* $g_i^0 = h_i^0$ *for all* $i \in [n]$ *in Algorithm 1 (DASHA-PP-MVR). Then*

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

In the next corollary, we choose momentum b and initialize vectors h_i^0 to get ε-solution.

Corollary 3. *Suppose that assumptions from Theorem 4 hold, momentum* $b = \Theta\left(\min\left\{\frac{p_a}{\omega}\sqrt{\frac{n\varepsilon B}{\sigma^2}}, \frac{p_a n\varepsilon B}{\sigma^2}\right\}\right)$, $\frac{\sigma^2}{n\varepsilon B} \geq 1$, *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* $i \in [n]$, *and batch size* $B_{\text{init}} = \Theta\left(\frac{\sqrt{p_a}B}{b}\right)$, *then Algorithm 1 (DASHA-PP-MVR) needs*

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

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

The convergence rate of the DASHA-SYNC-MVR, the state-of-the-art method without partial participation, equals $\mathcal{O}\left(\frac{\Delta_0}{\varepsilon}\left[L + \frac{\omega}{\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma}{\sqrt{\varepsilon n}}\frac{L_\sigma}{B}\right] + \frac{\sigma^2}{n\varepsilon B}\right)$. Similar to Section 6.2, we see that in the regimes when $\widehat{L} \approx L_\sigma$ the third term w.r.t. $1/\varepsilon^{3/2}$ can degenerate up to $\sqrt{B}/p_a$. However, if we take $B = \mathcal{O}\left(L_\sigma^2/\widehat{L}^2\right)$, then the degeneration of the third term will be up to $1/p_a$. This effect we analyze 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 Rand$K$ compressors (see Definition 5) and show that with the particular choice of parameters, up to the Lipschitz constants factors, DASHA-PP-MVR gets the optimal oracle complexity and SOTA communication complexity of DASHA-SYNC-MVR method. Indeed, comparing the following result with (Tyurin and Richtárik, 2023)[Corollary 6.9], one can see that we get the degeneration up to $1/p_a$ factor, which is expected in the partial participation setting. Note that the complexities improve with the number of workers n.

Corollary 4. *Suppose that assumptions of Corollary 3 hold, batch size* $B \leq \min\left\{\frac{\sigma}{p_a\sqrt{\varepsilon n}}, \frac{L_\sigma^2}{\mathbb{1}_{p_a}^2\widehat{L}^2}\right\}$, *we take Rand*$K$ *compressors with* $K = \Theta\left(\frac{Bd\sqrt{\varepsilon n}}{\sigma}\right)$. *Then the communication complexity equals* $\mathcal{O}\left(\frac{d\sigma}{\sqrt{p_a}\sqrt{n\varepsilon}} + \frac{L_\sigma\Delta_0 d}{p_a\sqrt{n\varepsilon}}\right)$, *and the expected number of stochastic gradient calculations per node equals* $\mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a}n\varepsilon} + \frac{L_\sigma\Delta_0\sigma}{p_a\varepsilon^{3/2}n}\right)$.

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

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020). On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex SGD. *arXiv preprint arXiv:1905.10018*.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *NeurIPS Information Processing Systems*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT Press.

Gorbunov, E., Burlachenko, K., Li, Z., and Richtárik, P. (2021). MARINA: Faster non-convex distributed learning with compression. In *38th International Conference on Machine Learning*.

Horváth, S., Ho, C.-Y., Horvath, L., Sahu, A. N., Canini, M., and Richtárik, P. (2019a). Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*.

Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. (2019b). Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tyurin, A. and Richtárik, P. (2023). DASHA: Distributed nonconvex optimization with communication compression and optimal oracle complexity. *International Conference on Learning Representations (ICLR)*.

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

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

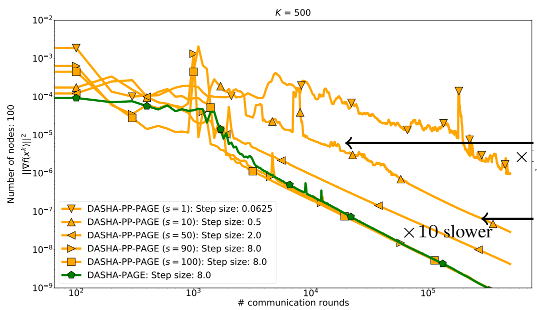
Xu, H., Ho, C.-Y., Abdelmoniem, A. M., Dutta, A., Bergou, E. H., Karatsenidis, K., Canini, M., and Kalnis, P. (2021). Grace: A compressed communication framework for distributed machine learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, pages 561–572. IEEE.

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

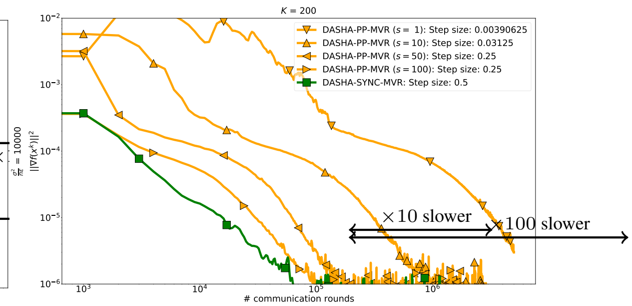
Zhao, H., Li, Z., and Richtárik, P. (2021b). FedPAGE: A fast local stochastic gradient method for communication-efficient federated learning. *arXiv preprint arXiv:2108.04755*.

Contents

A Numerical Verification of Theoretical Dependencies



(a) Finite-sum setting, $K = 500$ in Rand$K$.



(b) Stochastic setting, $\sigma^2/n\varepsilon B = 10000$, and $K = 200$ in Rand$K$.

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 $1/p_{\text{a}}$ 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^{th} node, $y_{ij} \in \{-1, 1\}$ is the corresponding label, and m is the number of samples on the i^{th} node for all $i \in [n]$. In the stochastic setting, we consider functions

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

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^{th} node, $y_{ij} \in \{1, 2\}$ is a corresponding label, m is the number of samples located on the i^{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[6] where all parameters except step sizes are taken as suggested in theory. Step sizes are finetuned from a set $\{2^i \,|\, i \in [-10, 10]\}$. We emulate the partial participation setting using s-nice sampling with the number of nodes $n = 100$. We consider the 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 $p_{\text{a}} = s/n$ to 1, the closer DASHA-PP to DASHA. Moreover, DASHA-PP with $s = 10$ and $s = 1$ converges approximately $\times 10$ $(= 1/p_{\text{a}})$ and $\times 100$ $(= 1/p_{\text{a}})$ times slower, accordingly. Our theory predicts such behavior.

[6]Code: https://github.com/mysteryresearcher/dasha-partial-participation

B Original DASHA and DASHA-MVR Methods

To simplify the discussion and explanation from the main part, we present the algorithms from (Tyurin 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** $t = 0, 1, \ldots, T-1$ **do**
4: $\quad x^{t+1} = x^t - \gamma g^t$
5: $\quad$ Broadcast x^{t+1}, x^t to all *participating*[(a)] nodes
6: $\quad$ **for** $i = 1, \ldots, n$ in parallel **do**
7: $\quad\quad m_i^{t+1} = \mathcal{C}_i\left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - a\left(g_i^t - \nabla f_i(x^t)\right)\right)$
8: $\quad\quad g_i^{t+1} = g_i^t + m_i^{t+1}$
9: $\quad\quad$ Send m_i^{t+1} to the server
10: $\quad$ **end for**
11: $\quad g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n m_i^{t+1}$
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** $t = 0, 1, \ldots, T-1$ **do**
4: $\quad x^{t+1} = x^t - \gamma g^t$
5: $\quad$ Broadcast x^{t+1}, x^t to all *participating*[(a)] nodes
6: $\quad$ **for** $i = 1, \ldots, n$ in parallel **do**
7: $\quad\quad h_i^{t+1} = \nabla f_i(x^{t+1}; \xi_i^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t; \xi_i^{t+1})), \quad \xi_i^{t+1} \sim \mathcal{D}_i$
8: $\quad\quad m_i^{t+1} = \mathcal{C}_i\left(h_i^{t+1} - h_i^t - a\left(g_i^t - h_i^t\right)\right)$
9: $\quad\quad g_i^{t+1} = g_i^t + m_i^{t+1}$
10: $\quad\quad$ Send m_i^{t+1} to the server
11: $\quad$ **end for**
12: $\quad g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n m_i^{t+1}$
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 $B = \mathcal{O}\left(\min\left\{\frac{1}{p_{\mathrm{a}}}\sqrt{\frac{m}{n}}, \frac{L_{\max}^2}{\mathbb{1}_{p_{\mathrm{a}}}^2 \widehat{L}^2}\right\}\right)$ and $B = \mathcal{O}\left(\min\left\{\frac{\sigma}{p_{\mathrm{a}}\sqrt{\varepsilon}n}, \frac{L_{\sigma}^2}{\mathbb{1}_{p_{\mathrm{a}}}^2 \widehat{L}^2}\right\}\right)$ in DASHA-PP-PAGE and DASHA-PP-MVR, accordingly, guarantees the degeneration up to $1/p_{\mathrm{a}}$. 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 $B = \mathcal{O}\left(\sqrt{\frac{m}{n}}\right)$ 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 ε-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

$$\begin{aligned} &\mathrm{E}\left[\left\|\frac{1}{nB}\sum_{i=1}^n\sum_{j\in I^i} x_{ij} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m x_{ij}\right\|^2\right] \qquad (10)\\ &= \frac{1}{nB}\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left\|x_{ij} - \frac{1}{m}\sum_{j=1}^m x_{ij}\right\|^2. \end{aligned}$$

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 $p_{\mathrm{a}} = s/n$ and $p_{\mathrm{a}} = s(s-1)/n(n-1)$. 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

$$\begin{aligned} &\mathrm{E}\left[\left\|\frac{1}{sB}\sum_{i\in S}\sum_{j\in I^i} x_{ij} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m x_{ij}\right\|^2\right] \qquad (11)\\ &= \frac{1}{sB}\underbrace{\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left\|x_{ij} - \frac{1}{m}\sum_{j=1}^m x_{ij}\right\|^2}_{\mathcal{L}_{\max}^2}\\ &\quad + \frac{n-s}{s(n-1)}\underbrace{\frac{1}{n}\sum_{i=1}^n\left\|\frac{1}{m}\sum_{j=1}^m x_{ij} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m x_{ij}\right\|^2}_{\widehat{\mathcal{L}}^2}. \end{aligned}$$

Let us assume that $s \leq n/2$. Note that (10) scales with any $B \geq 1$, while (11) only scales when $B = \mathcal{O}\left(\mathcal{L}_{\max}^2/\widehat{\mathcal{L}}^2\right)$. In other words, for large enough B, the variance in (11) does not significantly improves with the growth of B due to the term $\widehat{\mathcal{L}}^2$. In our proof, due to partial participation, the variance from (11) 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.

D Auxiliary facts

We list auxiliary facts that we use in our proofs:

1. For all $x, y \in \mathbb{R}^d$, we have
$$\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 \tag{12}$$
2. Let us take a *random vector* $\xi \in \mathbb{R}^d$, then
$$\mathrm{E}\left[\|\xi\|^2\right] = \mathrm{E}\left[\|\xi - \mathrm{E}[\xi]\|^2\right] + \|\mathrm{E}[\xi]\|^2. \tag{13}$$

D.1 Sampling Lemma

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

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

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

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

3. $$p_{\text{aa}} \leq p_{\text{a}}^2,$$

where $p_{\text{a}} \in (0,1]$ and $p_{\text{aa}} \in [0,1]$. Let us take random independent vectors $s_i \in \mathbb{R}^d$ for all $i \in [n]$, nonrandom vector $r_i \in \mathbb{R}^d$ for all $i \in [n]$, and random vectors
$$v_i = \begin{cases} r_i + \frac{1}{p_{\text{a}}} s_i, i \in S, \\ r_i, i \notin S, \end{cases}$$
then
$$\begin{aligned}
&\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n v_i - \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n v_i\right]\right\|^2\right] \\
&= \frac{1}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}\left[\|s_i - \mathrm{E}[s_i]\|^2\right] + \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2}\sum_{i=1}^n \|\mathrm{E}[s_i]\|^2 + \frac{p_{\text{aa}} - p_{\text{a}}^2}{p_{\text{a}}^2}\left\|\frac{1}{n}\sum_{i=1}^n \mathrm{E}[s_i]\right\| \\
&\leq \frac{1}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}\left[\|s_i - \mathrm{E}[s_i]\|^2\right] + \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2}\sum_{i=1}^n \|\mathrm{E}[s_i]\|^2.
\end{aligned}$$

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

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

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

Note, that
$$p_{\text{an}} = p_{\text{aa}} - p_{\text{a}} \tag{14}$$
and
$$p_{\text{nn}} = 1 - p_{\text{aa}} - 2p_{\text{an}}. \tag{15}$$
Using the law of total expectation and
$$\mathrm{E}[v_i] = p_{\text{a}}\left(r_i + \mathrm{E}\left[\frac{1}{p_{\text{a}}} s_i\right]\right) + (1 - p_{\text{a}}) r_i = r_i + \mathrm{E}[s_i],$$

468 we have

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

469 From the independence of random vectors s_i, we obtain

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

470 Using (14) and (15), we have

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

$$
\begin{aligned}
&+ \frac{p_{\text{aa}} - p_{\text{a}}^2}{n^2 p_{\text{a}}^2} \sum_{i \neq j}^{n} \left\langle \mathrm{E}\left[s_i\right], \mathrm{E}\left[s_j\right] \right\rangle \\
&\overset{(13)}{=} \frac{1}{n^2 p_{\text{a}}} \sum_{i=1}^{n} \mathrm{E}\left[\left\|s_i - \mathrm{E}\left[s_i\right]\right\|^2\right] \\
&+ \frac{1 - p_{\text{a}}}{n^2 p_{\text{a}}} \sum_{i=1}^{n} \left\|\mathrm{E}\left[s_i\right]\right\|^2 \\
&+ \frac{p_{\text{aa}} - p_{\text{a}}^2}{n^2 p_{\text{a}}^2} \sum_{i \neq j}^{n} \left\langle \mathrm{E}\left[s_i\right], \mathrm{E}\left[s_j\right] \right\rangle \\
&= \frac{1}{n^2 p_{\text{a}}} \sum_{i=1}^{n} \mathrm{E}\left[\left\|s_i - \mathrm{E}\left[s_i\right]\right\|^2\right] \\
&+ \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2} \sum_{i=1}^{n} \left\|\mathrm{E}\left[s_i\right]\right\|^2 \\
&+ \frac{p_{\text{aa}} - p_{\text{a}}^2}{p_{\text{a}}^2} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[s_i\right] \right\|.
\end{aligned}
$$

Finally, using that $p_{\text{aa}} \leq p_{\text{a}}^2$, we have

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

□

D.2 Compressors Facts

We define the Rand$K$ compressor that chooses without replacement K coordinates, scales them by a 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 $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is Rand$K$ if

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

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

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

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

E Proofs of Theorems

There are three different sources of randomness in Algorithm 1: the first one from vectors $\{k_i^{t+1}\}_{i=1}^n$, the second one from compressors $\{\mathcal{C}_i\}_{i=1}^n$, and the third one from availability of nodes. We define $\mathrm{E}_k\left[\cdot\right]$, $\mathrm{E}_{\mathcal{C}}\left[\cdot\right]$ and $\mathrm{E}_{p_{\text{a}}}\left[\cdot\right]$ to be conditional expectations w.r.t. $\{k_i^{t+1}\}_{i=1}^n$, $\{\mathcal{C}_i\}_{i=1}^n$, and availability, accordingly, conditioned on all previous randomness. Moreover, we define $\mathrm{E}_{t+1}\left[\cdot\right]$ to be a conditional expectation w.r.t. all randomness in iteration $t+1$ conditioned on all previous randomness. Note, that $\mathrm{E}_{t+1}\left[\cdot\right] = \mathrm{E}_k\left[\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\cdot\right]\right]\right]$.

In the case of DASHA-PP-PAGE, there are two different sources of randomness from $\{k_i^{t+1}\}_{i=1}^n$. We define $\mathrm{E}_{p_{\text{page}}}\left[\cdot\right]$ and $\mathrm{E}_B\left[\cdot\right]$ to be conditional expectations w.r.t. the probabilistic switching and mini-batch indices I_i^t, accordingly, conditioned on all previous randomness. Note, that $\mathrm{E}_{t+1}\left[\cdot\right] = \mathrm{E}_B\left[\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{page}}}\left[\cdot\right]\right]\right]\right]$ and $\mathrm{E}_{t+1}\left[\cdot\right] = \mathrm{E}_B\left[\mathrm{E}_{p_{\text{page}}}\left[\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\cdot\right]\right]\right]\right]$.

E.1 Standard Lemmas in the Nonconvex Setting

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

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 $\gamma > 0$, we have*

$$f(x^{t+1}) \le f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2. \quad (16)$$

Proof. Using $L-$smoothness, we have

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

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

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

□

Lemma 3. *Suppose that Assumption 1 holds and*

$$\mathrm{E}\left[f(x^{t+1})\right] + \gamma\Psi^{t+1} \le \mathrm{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \gamma\Psi^t + \gamma C, \quad (17)$$

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

$$\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \le \frac{2\Delta_0}{\gamma T} + \frac{2\Psi^0}{T} + 2C, \quad (18)$$

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

Proof. By unrolling (17) for t from 0 to $T-1$, we obtain

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

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

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

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

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

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

The lemma can be easily checked with the direct calculation.

E.2 Generic Lemmas

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

$$\begin{aligned}&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]\right]\\&\leq \frac{2\omega}{n^2 p_{\mathrm{a}}}\sum_{i=1}^n\left\|k_i^{t+1}\right\|^2+\frac{a^2((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{n^2p_{\mathrm{a}}^2}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2+(1-a)^2\left\|g^t-h^t\right\|^2,\end{aligned}\tag{19}$$

and

$$\begin{aligned}&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\right]\\&\leq \frac{2\omega}{p_{\mathrm{a}}}\left\|k_i^{t+1}\right\|^2+\left(\frac{a^2(2\omega+1-p_{\mathrm{a}})}{p_{\mathrm{a}}}+(1-a)^2\right)\left\|g_i^t-h_i^t\right\|^2\quad\forall i\in[n].\end{aligned}\tag{20}$$

Proof. First, we estimate $\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]\right]$:

$$\begin{aligned}&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]\right]\\&=\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g^{t+1}-h^{t+1}-\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[g^{t+1}-h^{t+1}\right]\right]\right\|^2\right]\right]+\left\|\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[g^{t+1}-h^{t+1}\right]\right]\right\|^2,\end{aligned}$$

where we used (13). Due to Assumption 8, we have

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

and

$$\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_i^{t+1}\right]\right]=p_{\mathrm{a}}\mathrm{E}_{\mathcal{C}}\left[h_i^t+\frac{1}{p_{\mathrm{a}}}k_i^{t+1}\right]+(1-p_{\mathrm{a}})h_i^t=h_i^t+k_i^{t+1}.$$

Thus, we can get

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

Due to the independence of compressors, we can use Lemma 1 with $r_i=g_i^t-h_i^t$ and $s_i=p_{\mathrm{a}}\mathcal{C}_i\left(\frac{1}{p_{\mathrm{a}}}k_i^{t+1}-\frac{a}{p_{\mathrm{a}}}\left(g_i^t-h_i^t\right)\right)-k_i^{t+1}$, and obtain

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

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

From Assumption 7, we have

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g^{t+1} - h^{t+1}\right\|^2\right]\right] \\
&\leq \frac{\omega p_{\mathrm{a}}}{n^2} \sum_{i=1}^n \left\|\frac{1}{p_{\mathrm{a}}} k_i^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right)\right\|^2 + \frac{a^2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g^t - h^t\right\|^2 \\
&= \frac{\omega}{n^2 p_{\mathrm{a}}} \sum_{i=1}^n \left\|k_i^{t+1} - a\left(g_i^t - h_i^t\right)\right\|^2 + \frac{a^2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g^t - h^t\right\|^2 \\
&\overset{(12)}{\leq} \frac{2\omega}{n^2 p_{\mathrm{a}}} \sum_{i=1}^n \left\|k_i^{t+1}\right\|^2 + \frac{a^2\left((2\omega+1) p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g^t - h^t\right\|^2 .
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right]\right] \\
&= \mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_i^{t+1} - h_i^{t+1} - \mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[g_i^{t+1} - h_i^{t+1}\right]\right]\right\|^2\right]\right] + \left\|\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[g_i^{t+1} - h_i^{t+1}\right]\right]\right\|^2 \\
&= \mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_i^{t+1} - h_i^{t+1} - g_i^t + a\left(g_i^t - h_i^t\right) + h_i^t\right\|^2\right]\right] + (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
&= p_{\mathrm{a}} \mathrm{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(\frac{1}{p_{\mathrm{a}}} k_i^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right)\right) - \frac{1}{p_{\mathrm{a}}} k_i^{t+1} + a\left(g_i^t - h_i^t\right)\right\|^2\right] \\
&\quad + a^2(1-p_{\mathrm{a}}) \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
&\overset{(13)}{=} p_{\mathrm{a}} \mathrm{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(\frac{1}{p_{\mathrm{a}}} k_i^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right)\right) - \left(\frac{1}{p_{\mathrm{a}}} k_i^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right)\right)\right\|^2\right] \\
&\quad + a^2 \frac{(1-p_{\mathrm{a}})^2}{p_{\mathrm{a}}} \left\|g_i^t - h_i^t\right\|^2 \\
&\quad + a^2(1-p_{\mathrm{a}}) \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
&\leq \frac{\omega}{p_{\mathrm{a}}} \left\|k_i^{t+1} - a\left(g_i^t - h_i^t\right)\right\|^2 \\
&\quad + \frac{a^2(1-p_{\mathrm{a}})}{p_{\mathrm{a}}} \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
&\overset{(12)}{\leq} \frac{2\omega}{p_{\mathrm{a}}} \left\|k_i^{t+1}\right\|^2 + \frac{a^2(2\omega+1-p_{\mathrm{a}})}{p_{\mathrm{a}}} \left\|g_i^t - h_i^t\right\|^2 + (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 .
\end{aligned}
$$

$\square$

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

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

Proof. Due to Lemma 2 and the update step from Line 5 in Algorithm 1, we have

$$
\begin{aligned}
&\mathrm{E}_{t+1}\left[f(x^{t+1})\right] \\
&\leq \mathrm{E}_{t+1}\left[f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2\right] \\
&= \mathrm{E}_{t+1}\left[f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - h^t + h^t - \nabla f(x^t)\right\|^2\right] \\
&\overset{(13)}{\leq} \mathrm{E}_{t+1}\left[f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \gamma\left(\left\|g^t - h^t\right\|^2 + \left\|h^t - \nabla f(x^t)\right\|^2\right]\right).
\end{aligned}
$$

Let us fix some constants $\kappa, \eta \in [0, \infty)$ that we will define later. Combining the last inequality, bounds (19), (20) and using the law of total expectation, we get

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

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

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

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

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

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

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

□

E.3 Proof for DASHA-PP

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

1.
$$
\begin{aligned}
&\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&\leq \frac{2\,(p_{\mathrm{a}} - p_{\mathrm{aa}})\,\widehat{L}^2}{np_{\mathrm{a}}^2}\left\|x^{t+1} - x^t\right\|^2 + \frac{2b^2\,(p_{\mathrm{a}} - p_{\mathrm{aa}})}{n^2p_{\mathrm{a}}^2}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2 + (1-b)^2\left\|h^t - \nabla f(x^t)\right\|^2.
\end{aligned}
$$

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

3.
$$
\left\|k_i^{t+1}\right\|^2 \leq 2L_i^2\left\|x^{t+1} - x^t\right\|^2 + 2b^2\left\|h_i^t - \nabla f_i(x^t)\right\|^2, \quad \forall i \in [n].
$$

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

$$
\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right]
$$

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

Using

$$\mathrm{E}_{p_{\mathrm{a}}}\left[h_i^{t+1}\right] = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))$$

and (13), we have

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

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

$$\begin{aligned}
&\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&\leq \frac{1}{n^2 p_{\mathrm{a}}}\sum_{i=1}^n \left\|k_i^{t+1} - k_i^{t+1}\right\|^2 + \frac{p_{\mathrm{a}} - p_{\mathrm{aa}}}{n^2 p_{\mathrm{a}}^2}\sum_{i=1}^n \left\|k_i^{t+1}\right\|^2 + (1-b)^2\left\|h^t - \nabla f(x^t)\right\|^2 \\
&= \frac{p_{\mathrm{a}} - p_{\mathrm{aa}}}{n^2 p_{\mathrm{a}}^2}\sum_{i=1}^n \left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right)\right\|^2 + (1-b)^2\left\|h^t - \nabla f(x^t)\right\|^2 \\
&\overset{(12)}{\leq} \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2}\sum_{i=1}^n \left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 + \frac{2b^2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2 + (1-b)^2\left\|h^t - \nabla f(x^t)\right\|^2 \\
&\leq \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)\widehat{L}^2}{n p_{\mathrm{a}}^2}\left\|x^{t+1} - x^t\right\|^2 + \frac{2b^2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2 + (1-b)^2\left\|h^t - \nabla f(x^t)\right\|^2.
\end{aligned}$$

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

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

Finally, the third inequality of the theorem follows from (12) and Assumption 3. □

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

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

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

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

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

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

541 After rearranging the terms, we get

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

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

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

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

Note that $b=\frac{p_{\mathrm{a}}}{2-p_{\mathrm{a}}}$, thus

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

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

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

and

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

Let us simplify the last inequality. First, note that

$$
\frac{16b\gamma\omega(2\omega+1)(1-p_{\mathrm{a}})\widehat{L}^2}{np_{\mathrm{a}}^3}\leq\frac{16\gamma\omega(2\omega+1)\widehat{L}^2}{np_{\mathrm{a}}^2},
$$

due to $b\leq p_{\mathrm{a}}$. Second,

$$
\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{bnp_{\mathrm{a}}^2}\leq\frac{4\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^3},
$$

due to $b\geq\frac{p_{\mathrm{a}}}{2}$. All in all, we have

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

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

Using Lemma 4 and the assumption about γ, we get

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

It is left to apply Lemma 3 with

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

to conclude the proof. □

E.4 Proof for DASHA-PP-PAGE

Let us denote

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

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

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

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

1.

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

2.

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

3.

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

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

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

Using

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

and

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

we obtain

$$\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\mathrm{page}}}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\right]\right]\\
&\overset{(13)}{=}p_{\mathrm{page}}\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_1^{t+1}-\mathrm{E}_{p_{\mathrm{a}}}\left[h_1^{t+1}\right]\right\|^2\right]+(1-p_{\mathrm{page}})\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_2^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_2^{t+1}\right]\right]\right\|^2\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&+ p_{\text{page}} \left\| \mathrm{E}_{p_{\text{a}}} \left[h_1^{t+1} \right] - \nabla f(x^{t+1}) \right\|^2 + (1 - p_{\text{page}}) \left\| \mathrm{E}_B \left[\mathrm{E}_{p_{\text{a}}} \left[h_2^{t+1} \right] \right] - \nabla f(x^{t+1}) \right\|^2 \\
&= p_{\text{page}} \mathrm{E}_{p_{\text{a}}} \left[\left\| h_1^{t+1} - \mathrm{E}_{p_{\text{a}}} \left[h_1^{t+1} \right] \right\|^2 \right] + (1 - p_{\text{page}}) \mathrm{E}_B \left[\mathrm{E}_{p_{\text{a}}} \left[\left\| h_2^{t+1} - \mathrm{E}_B \left[\mathrm{E}_{p_{\text{a}}} \left[h_2^{t+1} \right] \right] \right\|^2 \right] \right] \\
&+ \left(p_{\text{page}} \left(1 - \frac{b}{p_{\text{page}}} \right)^2 + (1 - p_{\text{page}}) \right) \left\| h^t - \nabla f(x^t) \right\|^2 .
\end{aligned} \tag{21}
$$

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

$$
\begin{aligned}
&\mathrm{E}_{p_{\text{a}}} \left[\left\| h_1^{t+1} - \mathrm{E}_{p_{\text{a}}} \left[h_1^{t+1} \right] \right\|^2 \right] \\
&\leq \frac{1}{n^2 p_{\text{a}}} \sum_{i=1}^n \left\| k_{i,1}^{t+1} - k_{i,1}^{t+1} \right\|^2 + \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2} \sum_{i=1}^n \left\| k_{i,1}^{t+1} \right\|^2 \\
&= \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{page}}} \left(h_i^t - \nabla f_i(x^t) \right) \right\|^2 \\
&\overset{(12)}{\leq} \frac{2 \left(p_{\text{a}} - p_{\text{aa}} \right)}{n^2 p_{\text{a}}^2} \sum_{i=1}^n \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2 \left(p_{\text{a}} - p_{\text{aa}} \right) b^2}{n^2 p_{\text{a}}^2 p_{\text{page}}^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 .
\end{aligned}
$$

From Assumption 3, we have

$$
\begin{aligned}
&\mathrm{E}_{p_{\text{a}}} \left[\left\| h_1^{t+1} - \mathrm{E}_{p_{\text{a}}} \left[h_1^{t+1} \right] \right\|^2 \right] \\
&\leq \frac{2 \left(p_{\text{a}} - p_{\text{aa}} \right) \widehat{L}^2}{n p_{\text{a}}^2} \left\| x^{t+1} - x^t \right\|^2 + \frac{2 \left(p_{\text{a}} - p_{\text{aa}} \right) b^2}{n^2 p_{\text{a}}^2 p_{\text{page}}^2} \sum_{i=1}^n \left\| h_i^t - \nabla f_i(x^t) \right\|^2 .
\end{aligned} \tag{22}
$$

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

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

Next, we use Assumptions 3 and 4 to get

$$\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_2^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_2^{t+1}\right]\right]\right\|^2\right]\right] \leq \left(\frac{L_{\max}^2}{np_{\mathrm{a}}B}+\frac{(p_{\mathrm{a}}-p_{\mathrm{aa}})\,\widehat{L}^2}{np_{\mathrm{a}}^2}\right)\left\|x^{t+1}-x^t\right\|^2. \tag{23}$$

Applying (22) and (23) into (21), we get

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

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

$$\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\text{page}}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\right]\\
&=p_{\text{page}}\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]+(1-p_{\text{page}})\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\\
&\overset{(13)}{=}p_{\text{page}}\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,1}^{t+1}\right]\right]\right\|^2\right]\right]+(1-p_{\text{page}})\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,2}^{t+1}\right]\right]\right\|^2\right]\right]\\
&\quad+p_{\text{page}}\left\|\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,1}^{t+1}\right]\right]-\nabla f_i(x^{t+1})\right\|^2+(1-p_{\text{page}})\left\|\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,2}^{t+1}\right]\right]-\nabla f_i(x^{t+1})\right\|^2\\
&=p_{\text{page}}\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,1}^{t+1}\right]\right]\right\|^2\right]\right]+(1-p_{\text{page}})\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,2}^{t+1}\right]\right]\right\|^2\right]\right]\\
&\quad+\left(p_{\text{page}}\left(1-\frac{b}{p_{\text{page}}}\right)^2+(1-p_{\text{page}})\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2.
\end{aligned}\tag{24}$$

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

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

Considering (12) and Assumption 3, we obtain

$$
\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,1}^{t+1}\right]\right]\right\|^2\right]\right] \\
&\leq \frac{2\left(1-p_{\mathrm{a}}\right) L_i^2}{p_{\mathrm{a}}}\left\|x^{t+1}-x^t\right\|^2+\frac{2\left(1-p_{\mathrm{a}}\right) b^2}{p_{\mathrm{a}} p_{\text{page}}^2}\left\|h_i^t-\nabla f_i(x^t)\right\|^2 .
\end{aligned} \tag{25}
$$

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

$$
\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,2}^{t+1}\right]\right]\right\|^2\right]\right] \\
&= p_{\mathrm{a}}\mathrm{E}_B\left[\left\|h_i^t+\frac{1}{p_{\mathrm{a}}}k_{i,2}^{t+1}-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad+(1-p_{\mathrm{a}})\mathrm{E}_B\left[\left\|h_i^t-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&= p_{\mathrm{a}}\mathrm{E}_B\left[\left\|\frac{1}{p_{\mathrm{a}}}k_{i,2}^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad+\left(1-p_{\mathrm{a}}\right)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\overset{(13)}{=} \frac{1}{p_{\mathrm{a}}}\mathrm{E}_B\left[\left\|k_{i,2}^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right]+\frac{\left(1-p_{\mathrm{a}}\right)^2}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\quad+\left(1-p_{\mathrm{a}}\right)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&= \frac{1}{p_{\mathrm{a}}}\mathrm{E}_B\left[\left\|k_{i,2}^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right]+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\leq \frac{1}{p_{\mathrm{a}}}\mathrm{E}_B\left[\left\|k_{i,2}^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right]+\frac{\left(1-p_{\mathrm{a}}\right)L_i^2}{p_{\mathrm{a}}}\left\|x^{t+1}-x^t\right\|^2,
\end{aligned} \tag{26}
$$

where we used Assumption 3. By plugging (25) and (26) into (24), we get

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

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

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

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

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

$$
\begin{aligned}
& \mathrm{E}_B\left[\mathrm{E}_{p_{\text{page}}}\left[\left\|k_i^{t+1}\right\|^2\right]\right] \\
&= p_{\text{page}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-\frac{b}{p_{\text{page}}}\left(h_i^t-\nabla f_i(x^t)\right)\right\|^2 \\
&\quad+(1-p_{\text{page}}) \mathrm{E}_B\left[\left\|\frac{1}{B} \sum_{j \in I_i^t}\left(\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right)\right\|^2\right] \\
&\overset{(13)}{=} p_{\text{page}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-\frac{b}{p_{\text{page}}}\left(h_i^t-\nabla f_i(x^t)\right)\right\|^2 \\
&\quad+(1-p_{\text{page}}) \mathrm{E}_B\left[\left\|\frac{1}{B} \sum_{j \in I_i^t}\left(\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right)-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad+\left(1-p_{\text{page}}\right)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\overset{(12)}{\leq} 2 p_{\text{page}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2+\frac{2 b^2}{p_{\text{page}}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\
&\quad+(1-p_{\text{page}}) \mathrm{E}_B\left[\left\|\frac{1}{B} \sum_{j \in I_i^t}\left(\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right)-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad+\left(1-p_{\text{page}}\right)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\leq 2\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2+\frac{2 b^2}{p_{\text{page}}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2
\end{aligned}
$$

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

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

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

.

It it left to consider Assumptions 3 and 4 to get

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

□

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

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

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

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

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

$$
\begin{aligned}
&+\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|k_i^{t+1}\right\|^2\right]\\
&+\nu\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\rho\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
=&\ \mathrm{E}\left[f(x^t)-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\gamma\left\|h^t-\nabla f(x^t)\right\|^2\right]\\
&+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&+\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{page}}}\left[\frac{1}{n}\sum_{i=1}^n\left\|k_i^{t+1}\right\|^2\right]\right]\right]\\
&+\nu\mathrm{E}\left[\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\mathrm{page}}}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\right]\right]\right]\\
&+\rho\mathrm{E}\left[\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\mathrm{page}}}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\right]\right]\\
\leq&\ \mathrm{E}\left[f(x^t)-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\gamma\left\|h^t-\nabla f(x^t)\right\|^2\right]\\
&+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&+\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\left(2\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\left\|x^{t+1}-x^t\right\|^2+\frac{2b^2}{p_{\mathrm{page}}}\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\\
&+\nu\mathrm{E}\Bigg(\left(\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^2}+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{np_{\mathrm{a}}B}\right)\left\|x^{t+1}-x^t\right\|^2\\
&\qquad+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b^2}{n^2p_{\mathrm{a}}^2p_{\mathrm{page}}}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2+\left(p_{\mathrm{page}}\left(1-\frac{b}{p_{\mathrm{page}}}\right)^2+(1-p_{\mathrm{page}})\right)\left\|h^t-\nabla f(x^t)\right\|^2\Bigg)\\
&+\rho\mathrm{E}\Bigg(\left(\frac{2\left(1-p_{\mathrm{a}}\right)\widehat{L}^2}{p_{\mathrm{a}}}+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{p_{\mathrm{a}}B}\right)\left\|x^{t+1}-x^t\right\|^2\\
&\qquad+\left(\frac{2\left(1-p_{\mathrm{a}}\right)b^2}{p_{\mathrm{a}}p_{\mathrm{page}}}+p_{\mathrm{page}}\left(1-\frac{b}{p_{\mathrm{page}}}\right)^2+(1-p_{\mathrm{page}})\right)\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\Bigg)
\end{aligned}
$$

585 After rearranging the terms, we get

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

$$
-\nu\left(\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^{2}}{np_{\mathrm{a}}^{2}}+\frac{(1-p_{\mathrm{page}})L_{\max}^{2}}{np_{\mathrm{a}}B}\right)-\rho\left(\frac{2\left(1-p_{\mathrm{a}}\right)\widehat{L}^{2}}{p_{\mathrm{a}}}+\frac{(1-p_{\mathrm{page}})L_{\max}^{2}}{p_{\mathrm{a}}B}\right)\Bigg)\mathrm{E}\left[\left\|x^{t+1}-x^{t}\right\|^{2}\right]
$$

$$
+\left(\gamma+\nu\left(p_{\mathrm{page}}\left(1-\frac{b}{p_{\mathrm{page}}}\right)^{2}+(1-p_{\mathrm{page}})\right)\right)\mathrm{E}\left[\left\|h^{t}-\nabla f(x^{t})\right\|^{2}\right]
$$

$$
+\Bigg(\frac{8b^{2}\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}+\frac{2\nu\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b^{2}}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}
$$

$$
+\rho\left(\frac{2\left(1-p_{\mathrm{a}}\right)b^{2}}{p_{\mathrm{a}}p_{\mathrm{page}}}+p_{\mathrm{page}}\left(1-\frac{b}{p_{\mathrm{page}}}\right)^{2}+(1-p_{\mathrm{page}})\right)\Bigg)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|h_{i}^{t}-\nabla f_{i}(x^{t})\right\|^{2}\right].
$$

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

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

586 therefore

$$
\mathrm{E}\left[f(x^{t+1})\right]+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^{2}\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^{2}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|g_{i}^{t+1}-h_{i}^{t+1}\right\|^{2}\right]
$$

$$
+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^{2}\right]+\rho\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|h_{i}^{t+1}-\nabla f_{i}(x^{t+1})\right\|^{2}\right]
$$

$$
\leq\mathrm{E}\left[f(x^{t})\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^{t})\right\|^{2}\right]
$$

$$
+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t}-h^{t}\right\|^{2}\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^{2}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|g_{i}^{t}-h_{i}^{t}\right\|^{2}\right]
$$

$$
-\Bigg(\frac{1}{2\gamma}-\frac{L}{2}-\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^{2}}\left(2\widehat{L}^{2}+\frac{(1-p_{\mathrm{page}})L_{\max}^{2}}{B}\right)
$$

$$
-\frac{\gamma}{b}\left(\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^{2}}{np_{\mathrm{a}}^{2}}+\frac{(1-p_{\mathrm{page}})L_{\max}^{2}}{np_{\mathrm{a}}B}\right)-\rho\left(\frac{2\left(1-p_{\mathrm{a}}\right)\widehat{L}^{2}}{p_{\mathrm{a}}}+\frac{(1-p_{\mathrm{page}})L_{\max}^{2}}{p_{\mathrm{a}}B}\right)\Bigg)\mathrm{E}\left[\left\|x^{t+1}-x^{t}\right\|^{2}\right]
$$

$$
+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t}-\nabla f(x^{t})\right\|^{2}\right]
$$

$$
+\Bigg(\frac{8b^{2}\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}
$$

$$
+\rho\left(\frac{2\left(1-p_{\mathrm{a}}\right)b^{2}}{p_{\mathrm{a}}p_{\mathrm{page}}}+p_{\mathrm{page}}\left(1-\frac{b}{p_{\mathrm{page}}}\right)^{2}+(1-p_{\mathrm{page}})\right)\Bigg)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|h_{i}^{t}-\nabla f_{i}(x^{t})\right\|^{2}\right].
$$

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

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

If we take $\rho=\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^{2}p_{\mathrm{page}}}$, then

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

587 therefore

$$
\mathrm{E}\left[f(x^{t+1})\right]+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^{2}\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^{2}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|g_{i}^{t+1}-h_{i}^{t+1}\right\|^{2}\right]
$$

$$
\begin{aligned}
&+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\le \mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(2\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\right.\\
&\qquad-\frac{\gamma}{bnp_{\mathrm{a}}}\left(2\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\qquad\left.-\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^3p_{\mathrm{page}}}+\frac{2\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\right)\left(2\left(1-p_{\mathrm{a}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\right)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right].
\end{aligned}
$$

Let us simplify the inequality. First, due to $b \ge \frac{p_{\mathrm{page}}p_{\mathrm{a}}}{2}$, we have

$$
\frac{\gamma}{bnp_{\mathrm{a}}}\left(2\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\le\frac{4\gamma}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\left(\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right).
$$

588 Second, due to $b \le p_{\mathrm{a}}p_{\mathrm{page}}$ and $p_{\mathrm{aa}} \le p_{\mathrm{a}}^2$, we get

$$
\begin{aligned}
&\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^3p_{\mathrm{page}}}+\frac{2\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\right)\left(2\left(1-p_{\mathrm{a}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\le\left(\frac{8\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\right)\left(2\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\le\frac{16\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\quad+\frac{4\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\left(\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\le\frac{16\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\\
&\quad+\frac{4\gamma}{np_{\mathrm{a}}^2p_{\mathrm{page}}}\left(\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2+\frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right).
\end{aligned}
$$

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

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

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

590 Using Lemma 4 and the assumption about γ, we get

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

591 It is left to apply Lemma 3 with

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

592 to conclude the proof. □

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

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

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

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

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

598 steps to get ε-solution. Using the choice of p_{mega} and the definition of $\mathbb{1}_{p_{\mathrm{a}}}$, we can get (9).
599 Note that the expected number of gradients calculations at each communication round equals $p_{\mathrm{mega}}m+$
600 $(1-p_{\mathrm{mega}})B=\frac{2mB}{m+B}\leq 2B.$ □

601 **Corollary 2.** *Suppose that assumptions of Corollary 1 hold, $B\leq\min\left\{\frac{1}{p_{\mathrm{a}}}\sqrt{\frac{m}{n}},\frac{L_{\max}^2}{\mathbb{1}_{p_{\mathrm{a}}}^2\widehat{L}^2}\right\}$[7], and we*
602 *use the unbiased compressor RandK with $K=\Theta\left({Bd}/{\sqrt{m}}\right)$. Then the communication complexity of*
603 *Algorithm 1 is $\mathcal{O}\left(d+\frac{L_{\max}\Delta_0 d}{p_{\mathrm{a}}\varepsilon\sqrt{n}}\right)$, and the expected number of gradient calculations per node equals*
604 $\mathcal{O}\left(m+\frac{L_{\max}\Delta_0\sqrt{m}}{p_{\mathrm{a}}\varepsilon\sqrt{n}}\right)$.

[7] If $\mathbb{1}_{p_{\mathrm{a}}}=0$, then $\frac{L_\sigma^2}{\mathbb{1}_{p_{\mathrm{a}}}^2\widehat{L}^2}=+\infty$

Proof. The communication complexity equals

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

Since $B \leq \frac{L_{\max}^2}{\mathbb{1}_{p_{\mathrm{a}}}^2\widehat{L}^2}$, we have $\frac{\mathbb{1}_{p_{\mathrm{a}}}\widehat{L}}{\sqrt{B}}+\frac{L_{\max}}{B} \leq \frac{2L_{\max}}{B}$ and

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

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

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

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

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

□

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

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 (DASHA-PP-FINITE-MVR) we have*

1.

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

2.

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

3.

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

4.

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

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

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

thus

$$\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\right] \\
&\overset{(13)}{=}\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1}-\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h^{t+1}\right]\right]\right\|^2\right]\right]+(1-b)^2\left\|h^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

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

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

Next, we again use Lemma 1 with $r_i=0$, $s_i=\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)$, $p_{\mathrm{a}}=\frac{B}{m}$, and $p_{\mathrm{aa}}=\frac{B(B-1)}{m(m-1)}$:

$$\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\right]$$

$$
\begin{aligned}
&\leq \frac{1}{n^2 p_{\mathrm{a}}} \sum_{i=1}^{n} \left(\frac{m-B}{Bm(m-1)} \sum_{j=1}^{m} \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b\left(h_{ij}^t - \nabla f_{ij}(x^t)\right) \right\|^2 \right) \\
&\quad + \frac{p_{\mathrm{a}} - p_{\mathrm{aa}}}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^{n} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right) \right\|^2 \\
&\quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\
&\leq \frac{1}{n^2 p_{\mathrm{a}} Bm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) - b\left(h_{ij}^t - \nabla f_{ij}(x^t)\right) \right\|^2 \\
&\quad + \frac{p_{\mathrm{a}} - p_{\mathrm{aa}}}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^{n} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right) \right\|^2 \\
&\quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 \\
&\overset{(12)}{\leq} \frac{2}{n^2 p_{\mathrm{a}} Bm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t) \right\|^2 + \frac{2b^2}{n^2 p_{\mathrm{a}} Bm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| h_{ij}^t - \nabla f_{ij}(x^t) \right\|^2 \\
&\quad + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^{n} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) \right\|^2 + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right) b^2}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^{n} \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
&\quad + (1-b)^2 \left\| h^t - \nabla f(x^t) \right\|^2 .
\end{aligned}
$$

Due to Assumptions 3 and 4, we have

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

Let us get the bound for the second inequality:

$$
\begin{aligned}
&\mathrm{E}_B \left[\mathrm{E}_{p_{\mathrm{a}}} \left[\left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right] \\
&\overset{(13)}{=} \mathrm{E}_B \left[\mathrm{E}_{p_{\mathrm{a}}} \left[\left\| h_i^{t+1} - \left(\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t)) \right) \right\|^2 \right] \right] \\
&\quad + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
&= p_{\mathrm{a}} \mathrm{E}_B \left[\left\| h_i^t + \frac{1}{p_{\mathrm{a}}} k_i^{t+1} - \left(\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t)) \right) \right\|^2 \right] \\
&\quad + (1-p_{\mathrm{a}}) \left\| h_i^t - \left(\nabla f_i(x^{t+1}) + (1-b)(h_i^t - \nabla f_i(x^t)) \right) \right\|^2 \\
&\quad + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 \\
&\overset{(13)}{=} \frac{1}{p_{\mathrm{a}}} \mathrm{E}_B \left[\left\| k_i^{t+1} - \mathrm{E}_B \left[k_i^{t+1} \right] \right\|^2 \right] \\
&\quad + \frac{1-p_{\mathrm{a}}}{p_{\mathrm{a}}} \left\| \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t)) \right\|^2 \\
&\quad + (1-b)^2 \left\| h_i^t - \nabla f_i(x^t) \right\|^2 .
\end{aligned}
$$

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

$$
\mathrm{E}_B \left[\mathrm{E}_{p_{\mathrm{a}}} \left[\left\| h_i^{t+1} - \nabla f_i(x^{t+1}) \right\|^2 \right] \right]
$$

$$
\begin{aligned}
&\le \frac{1}{p_{\mathrm{a}}}\left(\frac{m-B}{Bm(m-1)}\sum_{j=1}^{m}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)\right\|^2\right)\\
&\quad+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2\\
&\quad+(1-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&\le \frac{1}{p_{\mathrm{a}}Bm}\sum_{j=1}^{m}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)\right\|^2\\
&\quad+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2\\
&\quad+(1-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&\overset{(12)}{\le} \frac{2}{p_{\mathrm{a}}Bm}\sum_{j=1}^{m}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right\|^2+\frac{2(1-p_{\mathrm{a}})}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\\
&\quad+\frac{2b^2}{p_{\mathrm{a}}Bm}\sum_{j=1}^{m}\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2+\left(\frac{2\left(1-p_{\mathrm{a}}\right)b^2}{p_{\mathrm{a}}}+(1-b)^2\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&\le \left(\frac{2L_{\max}^2}{p_{\mathrm{a}}B}+\frac{2(1-p_{\mathrm{a}})L_i^2}{p_{\mathrm{a}}}\right)\left\|x^{t+1}-x^t\right\|^2\\
&\quad+\frac{2b^2}{p_{\mathrm{a}}Bm}\sum_{j=1}^{m}\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2+\left(\frac{2\left(1-p_{\mathrm{a}}\right)b^2}{p_{\mathrm{a}}}+(1-b)^2\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2,
\end{aligned}
$$

622 where we used Assumptions 3 and 4. We continue the proof by considering
623 $\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{ij}^{t+1}-\nabla f_{ij}(x^{t+1})\right\|^2\right]\right]$:

$$
\begin{aligned}
&\mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{ij}^{t+1}-\nabla f_{ij}(x^{t+1})\right\|^2\right]\right]\\
&\overset{(13)}{=} \mathrm{E}_B\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{ij}^{t+1}-\left(\nabla f_{ij}(x^{t+1})+(1-b)(h_{ij}^t-\nabla f_{ij}(x^t))\right)\right\|^2\right]\right]\\
&\quad+(1-b)^2\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2\\
&=\frac{p_{\mathrm{a}}B}{m}\mathrm{E}_B\left[\left\|h_{ij}^t+\frac{m}{Bp_{\mathrm{a}}}\left(\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)\right)-\left(\nabla f_{ij}(x^{t+1})+(1-b)(h_{ij}^t-\nabla f_{ij}(x^t))\right)\right\|^2\right]\\
&\quad+\left(1-\frac{p_{\mathrm{a}}B}{m}\right)\left\|h_{ij}^t-\left(\nabla f_{ij}(x^{t+1})+(1-b)(h_{ij}^t-\nabla f_{ij}(x^t))\right)\right\|^2\\
&\quad+(1-b)^2\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2\\
&=\frac{\left(1-\frac{p_{\mathrm{a}}B}{m}\right)^2}{\frac{p_{\mathrm{a}}B}{m}}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b(h_{ij}^t-\nabla f_{ij}(x^t))\right\|^2\\
&\quad+\left(1-\frac{p_{\mathrm{a}}B}{m}\right)\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b(h_{ij}^t-\nabla f_{ij}(x^t))\right\|^2\\
&\quad+(1-b)^2\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2\\
&=\frac{\left(1-\frac{p_{\mathrm{a}}B}{m}\right)}{\frac{p_{\mathrm{a}}B}{m}}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b(h_{ij}^t-\nabla f_{ij}(x^t))\right\|^2\\
&\quad+(1-b)^2\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2\\
&\overset{(12)}{\le} \frac{2\left(1-\frac{p_{\mathrm{a}}B}{m}\right)}{\frac{p_{\mathrm{a}}B}{m}}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right\|^2+\left(\frac{2\left(1-\frac{p_{\mathrm{a}}B}{m}\right)b^2}{\frac{p_{\mathrm{a}}B}{m}}+(1-b)^2\right)\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2.
\end{aligned}
$$

It is left to consider Assumption 4:

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

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

$$\begin{aligned}
&\mathrm{E}_B\left[\left\|k_i^{t+1}\right\|^2\right] \\
&\overset{(13)}{=} \mathrm{E}_B\left[\left\|k_i^{t+1}-\mathrm{E}_B\left[k_i^{t+1}\right]\right\|^2\right] \\
&\quad+\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2.
\end{aligned}$$

Using Lemma 1, we get

$$\begin{aligned}
&\mathrm{E}_B\left[\left\|k_i^{t+1}\right\|^2\right] \\
&\leq \frac{m-B}{Bm(m-1)}\sum_{j=1}^m\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)\right\|^2 \\
&\quad+\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2 \\
&\leq \frac{1}{Bm}\sum_{j=1}^m\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)-b\left(h_{ij}^t-\nabla f_{ij}(x^t)\right)\right\|^2 \\
&\quad+\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2 \\
&\overset{(12)}{\leq} \frac{2}{Bm}\sum_{j=1}^m\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right\|^2+2\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\
&\quad+\frac{2b^2}{Bm}\sum_{j=1}^m\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2+2b^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\
&\leq \left(\frac{2L_{\max}^2}{B}+2L_i^2\right)\left\|x^{t+1}-x^t\right\|^2 \\
&\quad+\frac{2b^2}{Bm}\sum_{j=1}^m\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2+2b^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2,
\end{aligned}$$

where we used Assumptions 3 and 4. □

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

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

$g_i^0=h_i^0=\nabla f_i(x^0)$ *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* (DASHA-PP-FINITE-MVR) *then* $\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]\leq\frac{2\Delta_0}{\gamma T}$.

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

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

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

$$
\begin{aligned}
&+ \rho \mathrm{E}\Bigg(\left(\frac{2L_{\max}^2}{p_{\mathrm{a}}B} + \frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\left\|x^{t+1}-x^t\right\|^2 \\
&\qquad + \frac{2b^2}{p_{\mathrm{a}}Bnm}\sum_{i=1}^n\sum_{j=1}^m \left\|h_{ij}^t - \nabla f_{ij}(x^t)\right\|^2 + \left(\frac{2\left(1-p_{\mathrm{a}}\right)b^2}{p_{\mathrm{a}}} + (1-b)^2\right)\frac{1}{n}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2 \Bigg) \\
&+ \delta \mathrm{E}\Bigg(\frac{2\left(1-\frac{p_{\mathrm{a}}B}{m}\right)L_{\max}^2}{\frac{p_{\mathrm{a}}B}{m}}\left\|x^{t+1}-x^t\right\|^2 \\
&\qquad + \left(\frac{2\left(1-\frac{p_{\mathrm{a}}B}{m}\right)b^2}{\frac{p_{\mathrm{a}}B}{m}} + (1-b)^2\right)\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left\|h_{ij}^t - \nabla f_{ij}(x^t)\right\|^2 \Bigg).
\end{aligned}
$$

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

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

and

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

632 Moreover, we consider that $1-\frac{p_{\mathrm{a}}B}{m} \leq 1$, therefore

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

$$+\frac{2b^2}{p_{\mathrm{a}}Bnm}\sum_{i=1}^n\sum_{j=1}^m\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2+(1-b)\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\Bigg)$$

$$+\delta\mathrm{E}\left(\frac{2mL_{\max}^2}{p_{\mathrm{a}}B}\left\|x^{t+1}-x^t\right\|^2+(1-b)\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left\|h_{ij}^t-\nabla f_{ij}(x^t)\right\|^2\right).$$

633 After rearranging the terms, we get

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

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

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

$$- \left(\frac{2\gamma L_{\max}^2}{bnp_{\text{a}}B} + \frac{2\gamma \left(p_{\text{a}} - p_{\text{aa}}\right) \widehat{L}^2}{bnp_{\text{a}}^2} \right) - \rho \left(\frac{2L_{\max}^2}{p_{\text{a}}B} + \frac{2(1-p_{\text{a}})\widehat{L}^2}{p_{\text{a}}} \right) - \delta \frac{2mL_{\max}^2}{p_{\text{a}}B} \Bigg) \mathrm{E}\left[\left\|x^{t+1} - x^t\right\|^2\right]$$
$$+ \frac{\gamma}{b}\mathrm{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right]$$
$$+ \left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2} + \frac{2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)b}{np_{\text{a}}^2} + \rho\left(1-b\right) \right) \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2\right]$$
$$+ \left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2 B} + \frac{2\gamma b}{np_{\text{a}}B} + \frac{2\rho b^2}{p_{\text{a}}B} + \delta\left(1-b\right) \right) \mathrm{E}\left[\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|h_{ij}^t - \nabla f_{ij}(x^t)\right\|^2\right].$$

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

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

635 therefore

$$\mathrm{E}\left[f(x^{t+1})\right] + \frac{\gamma(2\omega+1)}{p_{\text{a}}}\mathrm{E}\left[\left\|g^{t+1} - h^{t+1}\right\|^2\right] + \frac{\gamma((2\omega+1)\,p_{\text{a}} - p_{\text{aa}})}{np_{\text{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right]$$
$$+ \frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] + \left(\frac{8b\gamma\omega(2\omega+1)}{np_{\text{a}}^2} + \frac{2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)}{np_{\text{a}}^2} \right) \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|h_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right]$$
$$+ \delta\mathrm{E}\left[\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|h_{ij}^{t+1} - \nabla f_{ij}(x^{t+1})\right\|^2\right]$$
$$\leq \mathrm{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$$
$$+ \frac{\gamma(2\omega+1)}{p_{\text{a}}}\mathrm{E}\left[\left\|g^t - h^t\right\|^2\right] + \frac{\gamma((2\omega+1)\,p_{\text{a}} - p_{\text{aa}})}{np_{\text{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|g_i^t - h_i^t\right\|^2\right]$$
$$- \Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma\omega(2\omega+1)}{np_{\text{a}}^2}\left(\frac{2L_{\max}^2}{B} + 2\widehat{L}^2\right)$$
$$- \left(\frac{2\gamma L_{\max}^2}{bnp_{\text{a}}B} + \frac{2\gamma \left(p_{\text{a}} - p_{\text{aa}}\right) \widehat{L}^2}{bnp_{\text{a}}^2} \right) - \left(\frac{8b\gamma\omega(2\omega+1)}{np_{\text{a}}^2} + \frac{2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)}{np_{\text{a}}^2} \right) \left(\frac{2L_{\max}^2}{p_{\text{a}}B} + \frac{2(1-p_{\text{a}})\widehat{L}^2}{p_{\text{a}}} \right)$$
$$- \delta \frac{2mL_{\max}^2}{p_{\text{a}}B} \Bigg) \mathrm{E}\left[\left\|x^{t+1} - x^t\right\|^2\right]$$
$$+ \frac{\gamma}{b}\mathrm{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right]$$
$$+ \left(\frac{8b\gamma\omega(2\omega+1)}{np_{\text{a}}^2} + \frac{2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)}{np_{\text{a}}^2} \right) \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2\right]$$
$$+ \left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2 B} + \frac{2\gamma b}{np_{\text{a}}B} + \frac{16b^3\gamma\omega(2\omega+1)}{np_{\text{a}}^3 B} + \frac{4b^2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)}{nBp_{\text{a}}^3} + \delta\left(1-b\right) \right) \mathrm{E}\left[\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|h_{ij}^t - \nabla f_{ij}(x^t)\right\|^2\right].$$

636 Due to $b \leq p_{\text{a}}$ and $\frac{p_{\text{a}} - p_{\text{aa}}}{p_{\text{a}}} \leq 1$, we have

$$\frac{8b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2 B} + \frac{2\gamma b}{np_{\text{a}}B} + \frac{16b^3\gamma\omega(2\omega+1)}{np_{\text{a}}^3 B} + \frac{4b^2\gamma\left(p_{\text{a}} - p_{\text{aa}}\right)}{nBp_{\text{a}}^3}$$
$$\leq \frac{8b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2 B} + \frac{2\gamma b}{np_{\text{a}}B} + \frac{16b^2\gamma\omega(2\omega+1)}{np_{\text{a}}^2 B} + \frac{4\gamma b}{np_{\text{a}}B}$$

$$
= \frac{24 b^2 \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^2 B} + \frac{6\gamma b}{n p_{\mathrm{a}} B}.
$$

Let us take $\delta = \frac{24 b \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^2 B} + \frac{6\gamma}{n p_{\mathrm{a}} B}$. Thus

$$
\left(\frac{8 b^2 \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^2 B} + \frac{2\gamma b}{n p_{\mathrm{a}} B} + \frac{16 b^3 \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^3 B} + \frac{4 b^2 \gamma \left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n B p_{\mathrm{a}}^3} + \delta \left(1 - b\right)\right) \leq \delta
$$

and

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

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

$$
\begin{aligned}
&\frac{4\gamma\omega(2\omega+1)}{n p_{\mathrm{a}}^2}\left(\frac{2 L_{\max}^2}{B} + 2\widehat{L}^2\right) \\
&\quad + \left(\frac{2\gamma L_{\max}^2}{b n p_{\mathrm{a}} B} + \frac{2\gamma \left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right) \widehat{L}^2}{b n p_{\mathrm{a}}^2}\right) \\
&\quad + \left(\frac{8 b \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^2} + \frac{2\gamma \left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n p_{\mathrm{a}}^2}\right)\left(\frac{2 L_{\max}^2}{p_{\mathrm{a}} B} + \frac{2(1 - p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right) \\
&\quad + \left(\frac{24 b \gamma \omega (2\omega+1)}{n p_{\mathrm{a}}^2 B} + \frac{6\gamma}{n p_{\mathrm{a}} B}\right) \frac{2 m L_{\max}^2}{p_{\mathrm{a}} B} \\
&\leq \frac{12\gamma\omega(2\omega+1)}{n p_{\mathrm{a}}^2}\left(\frac{2 L_{\max}^2}{B} + 2\widehat{L}^2\right)
\end{aligned}
$$

$$+\left(\frac{6\gamma L_{\max}^2}{bnp_{\mathrm{a}}B}+\frac{6\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{bnp_{\mathrm{a}}^2}\right)$$
$$+\left(\frac{24b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2B}+\frac{6\gamma}{np_{\mathrm{a}}B}\right)\frac{2mL_{\max}^2}{p_{\mathrm{a}}B}$$

Considering that $b\leq\frac{p_{\mathrm{a}}B}{m}$ and $b\geq\frac{p_{\mathrm{a}}B}{2m}$, we obtain

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

All in all, we have

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

Using Lemma 4 and the assumption about γ, we get

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

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

It is left to apply Lemma 3 with

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

to conclude the proof. □

E.6 Proof for DASHA-PP-MVR

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

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

1.

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

2.

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

3.

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

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

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

Using

$$\mathrm{E}_k\left[\mathrm{E}_{p_\mathrm{a}}\left[h_i^{t+1}\right]\right] = h_i^t + \mathrm{E}_k\left[k_i^{t+1}\right] = h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))$$

and (13), we have

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

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

$$\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_\mathrm{a}}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\right] \\
&\le \frac{1}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|k_i^{t+1}-\mathrm{E}_k\left[k_i^{t+1}\right]\right\|^2\right] + \frac{p_\mathrm{a}-p_\mathrm{aa}}{n^2p_\mathrm{a}^2}\sum_{i=1}^n\left\|\mathrm{E}_k\left[k_i^{t+1}\right]\right\|^2 + (1-b)^2\left\|h^t-\nabla f(x^t)\right\|^2 \\
&= \frac{1}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^t;\xi_i^{t+1}) - b\left(h_i^t - \nabla f_i(x^t;\xi_i^{t+1})\right)\right.\right. \\
&\quad \left.\left. - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right)\right)\right\|^2\right] \\
&\quad + \frac{p_\mathrm{a}-p_\mathrm{aa}}{n^2p_\mathrm{a}^2}\sum_{i=1}^n\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b\left(h_i^t - \nabla f_i(x^t)\right)\right\|^2 \\
&\quad + (1-b)^2\left\|h^t-\nabla f(x^t)\right\|^2 \\
&\overset{(12)}{\le} \frac{2}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|b\left(\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^{t+1})\right)\right\|^2\right] \\
&\quad + \frac{2}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|(1-b)\left(\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^t;\xi_i^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right)\right\|^2\right] \\
&\quad + \frac{p_\mathrm{a}-p_\mathrm{aa}}{n^2p_\mathrm{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 \\
&= \frac{2b^2}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right] \\
&\quad + \frac{2(1-b)^2}{n^2p_\mathrm{a}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^t;\xi_i^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{p_\mathrm{a}-p_\mathrm{aa}}{n^2p_\mathrm{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.
\end{aligned}$$

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

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

$$
\begin{aligned}
&\mathrm{E}_k \left[\mathrm{E}_{p_{\mathrm{a}}} \left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right] \\
&\le \frac{2b^2\sigma^2}{n p_{\mathrm{a}} B} \\
&\quad + \frac{2(1-b)^2}{n^2 p_{\mathrm{a}}} \sum_{i=1}^n \mathrm{E}_k \left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^t;\xi_i^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{p_{\mathrm{a}} - p_{\mathrm{aa}}}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - b(h_i^t - \nabla f_i(x^t))\right\|^2 \\
&\quad + (1-b)^2 \left\|h^t - \nabla f(x^t)\right\|^2 \\
&\overset{(12)}{\le} \frac{2b^2\sigma^2}{n p_{\mathrm{a}} B} \\
&\quad + \frac{2(1-b)^2}{n^2 p_{\mathrm{a}}} \sum_{i=1}^n \mathrm{E}_k \left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1}) - \nabla f_i(x^t;\xi_i^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right) b^2}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2 \\
&\quad + (1-b)^2 \left\|h^t - \nabla f(x^t)\right\|^2. \\
&= \frac{2b^2\sigma^2}{n p_{\mathrm{a}} B} \\
&\quad + \frac{2(1-b)^2}{n^2 p_{\mathrm{a}} B^2} \sum_{i=1}^n \sum_{j=1}^B \mathrm{E}_k \left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1}) - \nabla f_i(x^t;\xi_{ij}^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 + \frac{2\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right) b^2}{n^2 p_{\mathrm{a}}^2} \sum_{i=1}^n \left\|h_i^t - \nabla f_i(x^t)\right\|^2 \\
&\quad + (1-b)^2 \left\|h^t - \nabla f(x^t)\right\|^2,
\end{aligned}
$$

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

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

658 Now, we prove the second inequality:

$$
\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\\
&=\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_i^{t+1}-\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_i^{t+1}\right]\right]\right\|^2\right]\right]\\
&\quad+\left\|\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_i^{t+1}\right]\right]-\nabla f_i(x^{t+1})\right\|^2\\
&=\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_i^{t+1}-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\right]\\
&\quad+\left\|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})\right\|^2\\
&=\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_i^{t+1}-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\right]\\
&\quad+(1-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&=p_{\mathrm{a}}\mathrm{E}_k\left[\left\|h_i^t+\frac{1}{p_{\mathrm{a}}}k_i^{t+1}-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\\
&\quad+(1-p_{\mathrm{a}})\left\|h_i^t-\left(h_i^t+\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\\
&\quad+(1-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&=p_{\mathrm{a}}\mathrm{E}_k\left[\left\|\frac{1}{p_{\mathrm{a}}}k_i^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\\
&\quad+(1-p_{\mathrm{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-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&\overset{(13)}{=}\frac{1}{p_{\mathrm{a}}}\mathrm{E}_k\left[\left\|k_i^{t+1}-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\\
&\quad+\frac{(1-p_{\mathrm{a}})^2}{p_{\mathrm{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_{\mathrm{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-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&=\frac{1}{p_{\mathrm{a}}}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-b\left(h_i^t-\nabla f_i(x^t;\xi_i^{t+1})\right)-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right]\\
&\quad+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{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-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&=\frac{1}{p_{\mathrm{a}}}\mathrm{E}_k\left[\left\|b\left(\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^{t+1})\right)+(1-b)\left(\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right)\right\|^2\right]\\
&\quad+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{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-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2\\
&\overset{(12)}{\leq}\frac{2b^2}{p_{\mathrm{a}}}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\quad+\frac{2(1-b)^2}{p_{\mathrm{a}}}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right]\\
&\quad+\frac{1-p_{\mathrm{a}}}{p_{\mathrm{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-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_\mathrm{a}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right] \\
&= \frac{2b^2}{p_\mathrm{a}B^2}\sum_{j=1}^{B}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1})-\nabla f_i(x^{t+1})\right\|^2\right] \\
&\quad + \frac{2(1-b)^2}{p_\mathrm{a}B^2}\sum_{j=1}^{B}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1})-\nabla f_i(x^{t};\xi_{ij}^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{1-p_\mathrm{a}}{p_\mathrm{a}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2 \\
&\quad + (1-b)^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2. \\
&\overset{(12)}{\leq} \frac{2b^2}{p_\mathrm{a}B^2}\sum_{j=1}^{B}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1})-\nabla f_i(x^{t+1})\right\|^2\right] \\
&\quad + \frac{2(1-b)^2}{p_\mathrm{a}B^2}\sum_{j=1}^{B}\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1})-\nabla f_i(x^{t};\xi_{ij}^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + \frac{2(1-p_\mathrm{a})}{p_\mathrm{a}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 + \left(\frac{2(1-p_\mathrm{a})b^2}{p_\mathrm{a}}+(1-b)^2\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\mathrm{E}_k\left[\left\|k_i^{t+1}\right\|^2\right] \\
&= \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-b\left(h_i^t-\nabla f_i(x^t;\xi_i^{t+1})\right)\right\|^2\right] \\
&\overset{(13)}{=} \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-b\left(h_i^t-\nabla f_i(x^t;\xi_i^{t+1})\right)-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right)\right\|^2\right] \\
&\quad + \left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2 \\
&= \mathrm{E}_k\left[\left\|b\left(\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^{t+1})\right)+(1-b)\left(\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right)\right\|^2\right] \\
&\quad + \left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-b(h_i^t-\nabla f_i(x^t))\right\|^2 \\
&\overset{(12)}{\leq} 2b^2\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^{t+1})\right\|^2\right] \\
&\quad + 2(1-b)^2\mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_i^{t+1})-\nabla f_i(x^t;\xi_i^{t+1})-\left(\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right)\right\|^2\right] \\
&\quad + 2\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 + 2b^2\left\|h_i^t-\nabla f_i(x^t)\right\|^2.
\end{aligned}
$$

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

$$
\mathrm{E}_k\left[\left\|k_i^{t+1}\right\|^2\right]
$$

$$\leq \frac{2b^2\sigma^2}{B} + \left(\frac{2(1-b)^2 L_\sigma^2}{B} + 2L_i^2\right)\left\|x^{t+1} - x^t\right\|^2 + 2b^2\left\|h_i^t - \nabla f_i(x^t)\right\|^2.$$

$\square$

Theorem 4. *Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take* $a = \frac{p_a}{2\omega+1}$, $b \in \left(0, \frac{p_a}{2-p_a}\right]$, $\gamma \leq \left(L + \left[\frac{48\omega(2\omega+1)}{np_a^2}\left(\widehat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B}\right) + \frac{12}{np_a b}\left(\left(1 - \frac{p_{aa}}{p_a}\right)\widehat{L}^2 + \frac{(1-b)^2 L_\sigma^2}{B}\right)\right]^{1/2}\right)^{-1}$, *and* $g_i^0 = h_i^0$ *for all* $i \in [n]$ *in Algorithm 1 (DASHA-PP-MVR). Then*

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

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

$$
\begin{aligned}
&+\nu \mathrm{E}\Bigg(\frac{2b^2\sigma^2}{np_{\mathrm{a}}B}+\left(\frac{2(1-b)^2L_\sigma^2}{np_{\mathrm{a}}B}+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^2}\right)\left\|x^{t+1}-x^t\right\|^2\\
&\qquad+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b^2}{n^2p_{\mathrm{a}}^2}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2+(1-b)^2\left\|h^t-\nabla f(x^t)\right\|^2\Bigg)\\
&+\rho \mathrm{E}\Bigg(\frac{2b^2\sigma^2}{p_{\mathrm{a}}B}+\left(\frac{2(1-b)^2L_\sigma^2}{p_{\mathrm{a}}B}+\frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\left\|x^{t+1}-x^t\right\|^2\\
&\qquad+\left(\frac{2(1-p_{\mathrm{a}})b^2}{p_{\mathrm{a}}}+(1-b)^2\right)\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\Bigg).
\end{aligned}
$$

669 After rearranging the terms, we get

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right]+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\\
&\quad+\nu\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\rho\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\le \mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&\quad-\Bigg(\frac{1}{2\gamma}-\frac{L}{2}-\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\frac{2(1-b)^2L_\sigma^2}{B}+2\widehat{L}^2\right)\\
&\qquad-\nu\left(\frac{2(1-b)^2L_\sigma^2}{np_{\mathrm{a}}B}+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^2}\right)-\rho\left(\frac{2(1-b)^2L_\sigma^2}{p_{\mathrm{a}}B}+\frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\Bigg)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&\quad+\left(\gamma+\nu\left(1-b\right)^2\right)\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]\\
&\quad+\left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\nu\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b^2}{np_{\mathrm{a}}^2}+\rho\left(\frac{2(1-p_{\mathrm{a}})b^2}{p_{\mathrm{a}}}+(1-b)^2\right)\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\\
&\quad+\left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\nu\frac{2b^2}{np_{\mathrm{a}}}+\rho\frac{2b^2}{p_{\mathrm{a}}}\right)\frac{\sigma^2}{B}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right]+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\\
&\quad+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\rho\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\le \mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&\quad-\Bigg(\frac{1}{2\gamma}-\frac{L}{2}-\frac{4\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\frac{2(1-b)^2L_\sigma^2}{B}+2\widehat{L}^2\right)\\
&\qquad-\frac{\gamma}{b}\left(\frac{2(1-b)^2L_\sigma^2}{np_{\mathrm{a}}B}+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^2}\right)-\rho\left(\frac{2(1-b)^2L_\sigma^2}{p_{\mathrm{a}}B}+\frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\Bigg)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&\quad+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b}{np_{\mathrm{a}}^2}+\rho\left(\frac{2(1-p_{\mathrm{a}})b^2}{p_{\mathrm{a}}}+(1-b)^2\right)\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\\
&+\left(\frac{8b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\gamma b}{np_{\mathrm{a}}}+\rho\frac{2b^2}{p_{\mathrm{a}}}\right)\frac{\sigma^2}{B}.
\end{aligned}
$$

Note that $b \leq \frac{p_{\mathrm{a}}}{2-p_{\mathrm{a}}}$, thus

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

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

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

and

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

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

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

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

therefore

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

Also, we can simplify the last term:

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

thus

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

$$
\begin{aligned}
&+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
\le{}&\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&+\frac{\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\frac{\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{24\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\frac{(1-b)^2L_\sigma^2}{B}+\widehat{L}^2\right)\right.\\
&\qquad\left.-\frac{6\gamma}{np_{\mathrm{a}}b}\left(\frac{(1-b)^2L_\sigma^2}{B}+\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2\right)\right)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t}-\nabla f(x^{t})\right\|^2\right]+\left(\frac{8b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t}-\nabla f_i(x^{t})\right\|^2\right]\\
&+\left(\frac{24b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{6\gamma b}{np_{\mathrm{a}}}\right)\frac{\sigma^2}{B}.
\end{aligned}
$$

Using Lemma 4 and the assumption about γ, we get

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

It is left to apply Lemma 3 with

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

and $C=\left(\frac{24b^2\omega(2\omega+1)}{p_{\mathrm{a}}^2}+\frac{6b}{p_{\mathrm{a}}}\right)\frac{\sigma^2}{nB}$ to conclude the proof. □

Corollary 3. *Suppose that assumptions from Theorem 4 hold, momentum* $b=\Theta\left(\min\left\{\frac{p_{\mathrm{a}}}{\omega}\sqrt{\frac{n\varepsilon B}{\sigma^2}},\frac{p_{\mathrm{a}}n\varepsilon B}{\sigma^2}\right\}\right)$, $\frac{\sigma^2}{n\varepsilon B}\ge 1$, *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* $i\in[n]$, *and batch size* $B_{\text{init}}=\Theta\left(\frac{\sqrt{p_{\mathrm{a}}}B}{b}\right)$, *then Algorithm 1 (DASHA-PP-MVR) needs*

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

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

Proof. Using the result from Theorem 4, we have

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

We choose b to ensure $\left(\frac{48b^2\omega(2\omega+1)}{p_{\mathrm{a}}^2}+\frac{12b}{p_{\mathrm{a}}}\right)\frac{\sigma^2}{nB} = \Theta(\varepsilon)$. Note that $\frac{1}{b} = \Theta\left(\max\left\{\frac{\omega}{p_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{n\varepsilon B}}, \frac{\sigma^2}{p_{\mathrm{a}}n\varepsilon B}\right\}\right) \leq \Theta\left(\max\left\{\frac{\omega^2}{p_{\mathrm{a}}}, \frac{\sigma^2}{p_{\mathrm{a}}n\varepsilon B}\right\}\right)$, thus

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

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

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

Let us bound the norms:

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

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

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

693 Using the choice of B_{init} and b, we obtain

$$T = \mathcal{O}\left(\frac{1}{\varepsilon}\left[\Delta_0\left(L + \frac{\omega}{p_{\mathrm{a}}\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \sqrt{\frac{\sigma^2}{p_{\mathrm{a}}^2 \varepsilon n^2 B}}\left(\mathbb{1}_{p_{\mathrm{a}}}\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)\right)\right.\right.$$
$$\left.\left. + \frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}nB} + \frac{b^2\omega^2\sigma^2}{np_{\mathrm{a}}^{5/2}B} + \frac{b\sigma^2}{p_{\mathrm{a}}^{3/2}nB}\right]\right)$$
$$= \mathcal{O}\left(\frac{1}{\varepsilon}\left[\Delta_0\left(L + \frac{\omega}{p_{\mathrm{a}}\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \sqrt{\frac{\sigma^2}{p_{\mathrm{a}}^2 \varepsilon n^2 B}}\left(\mathbb{1}_{p_{\mathrm{a}}}\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)\right)\right.\right.$$
$$\left.\left. + \frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}nB} + \frac{\varepsilon}{\sqrt{p_{\mathrm{a}}}}\right]\right)$$
$$= \mathcal{O}\left(\frac{\Delta_0}{\varepsilon}\left[L + \frac{\omega}{p_{\mathrm{a}}\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \sqrt{\frac{\sigma^2}{p_{\mathrm{a}}^2 \varepsilon n^2 B}}\left(\mathbb{1}_{p_{\mathrm{a}}}\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)\right] + \frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}n\varepsilon B} + \frac{1}{\sqrt{p_{\mathrm{a}}}}\right).$$

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

696 **Corollary 4.** *Suppose that assumptions of Corollary 3 hold, batch size* $B \leq \min\left\{\frac{\sigma}{p_{\mathrm{a}}\sqrt{\varepsilon}n}, \frac{L_\sigma^2}{\mathbb{1}_{p_{\mathrm{a}}}^2 \widehat{L}^2}\right\}$,
697 *we take RandK compressors with* $K = \Theta\left(\frac{Bd\sqrt{\varepsilon n}}{\sigma}\right)$. *Then the communication complexity equals*
698 $\mathcal{O}\left(\frac{d\sigma}{\sqrt{p_{\mathrm{a}}}\sqrt{n\varepsilon}} + \frac{L_\sigma\Delta_0 d}{p_{\mathrm{a}}\sqrt{n\varepsilon}}\right)$, *and the expected number of stochastic gradient calculations per node equals*
699 $\mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}n\varepsilon} + \frac{L_\sigma\Delta_0\sigma}{p_{\mathrm{a}}\varepsilon^{3/2}n}\right)$.

700 *Proof.* The communication complexity equals

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

Due to $B \leq \frac{L_\sigma^2}{\mathbb{1}_{p_\mathrm{a}}^2 \widehat{L}^2}$, we have $\mathbb{1}_{p_\mathrm{a}} \widehat{L} + \frac{L_\sigma}{\sqrt{B}} \leq \frac{2L_\sigma}{\sqrt{B}}$ and

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

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

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

And the expected number of stochastic gradient calculations per node equals

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

□

F Analysis of DASHA-PP under Polyak-Łojasiewicz Condition

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

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

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*), \quad \forall x \in \mathbb{R}, \tag{27}$$

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

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

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

F.1 Gradient Setting

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

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

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

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

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{p_{\mathrm{a}}} + \frac{L}{\mu} + \frac{\omega\widehat{L}}{p_{\mathrm{a}}\mu\sqrt{n}} + \frac{\widehat{L}}{p_{\mathrm{a}}\mu\sqrt{n}}\right),$$

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

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

communication rounds to get ε-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 $1/p_{\mathrm{a}}$ 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 $a = \frac{p_{\mathrm{a}}}{2\omega+1}$, probability $p_{page} = \frac{B}{m+B}$, $b = \frac{p_{page}p_{\mathrm{a}}}{2-p_{\mathrm{a}}}$,*

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

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

The provided theorem states that to get ε-solution DASHA-PP have to run

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{p_{\mathrm{a}}} + \frac{m}{p_{\mathrm{a}}B} + \frac{L}{\mu} + \frac{\omega}{p_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_{\max}}{\sqrt{B}}\right) + \frac{\sqrt{m}}{p_{\mathrm{a}}\mu\sqrt{nB}}\left(\widehat{L} + \frac{L_{\max}}{\sqrt{B}}\right)\right),$$

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

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

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

F.3 Stochastic Setting

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

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

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

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

The provided theorems states that to get ε-solution DASHA-PP have to run

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{p_{\mathrm{a}}} + \underbrace{\frac{\omega}{p_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{\mu n\varepsilon B}}}_{\mathcal{P}_2} + \frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \underbrace{\frac{\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon B}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)}_{\mathcal{P}_1}\right) \tag{28}$$

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

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

$$\widetilde{\mathcal{O}}\left(\omega + \frac{\sigma^2}{\mu n\varepsilon B} + \frac{L}{\mu} + \frac{\omega}{\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma}{n\mu^{3/2}\sqrt{\varepsilon B}}\left(\frac{L_\sigma}{\sqrt{B}}\right)\right) \tag{29}$$

communication rounds to get ε-solution[8].

In the stochastic setting, the comparison is a little bit more complicated. As in the finite-sum setting, we have to take $B = \mathcal{O}\left(\frac{L_\sigma^2}{\widehat{L}^2}\right)$ to guarantee the degeneration up to $1/p_{\mathrm{a}}$ of the term $\mathcal{P}_1$ from (28). However, DASHA-PP-MVR has also suboptimal term $\mathcal{P}_2$. This suboptimality is tightly connected with the suboptimality of B_{init} in the general nonconvex case, which we discuss in Section 6.3, and it also appears in the analysis of DASHA-MVR (Tyurin and Richtárik, 2023). Let us provide the counterpart of Corollary 4. The corollary reveals that we can escape regimes when $\mathcal{P}_2$ is the bottleneck by choosing the parameters of the compressors.

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

$$\widetilde{\mathcal{O}}\left(\frac{d\sigma}{p_{\mathrm{a}}\sqrt{\mu\varepsilon n}} + \frac{dL_\sigma}{p_{\mathrm{a}}\mu\sqrt{n}}\right),$$

and the expected number of stochastic gradient calculations per node equals

$$\widetilde{\mathcal{O}}\left(\frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon} + \frac{\sigma L_\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon}}\right).$$

Up to Lipschitz constants, DASHA-PP-MVR has the state-of-the-art oracle complexity under PŁ-condition (see (Li et al., 2021a)). Moreover, DASHA-PP-MVR has the state-of-the-art communication complexity of DASHA for a small enough μ.

[8]For simplicity, we omitted $\frac{d}{\zeta_{\mathcal{C}}}$ term from the complexity in the stochastic setting, where $\zeta_{\mathcal{C}}$ is defined in Definition 12. For instance, for the Rand$K$ compressor (see Definition 5 and Theorem 6), $\zeta_{\mathcal{C}} = K$ and $\frac{d}{\zeta_{\mathcal{C}}} = \Theta(\omega)$.

F.4 Proofs of Theorems

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

F.4.1 Standard Lemma under Polyak-Łojasiewicz Condition

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

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

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

$$\mathrm{E}\left[f(x^T) - f^*\right] \le (1-\gamma\mu)^T \left(\left(f(x^0) - f^*\right) + \gamma\Psi^0\right) + \frac{C}{\mu}. \tag{30}$$

Proof. We subtract f^* and use PŁ-condition (27) to get

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

Unrolling the inequality, we have

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

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

F.4.2 Generic Lemma

We now provide the counterpart of Lemma 6.

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

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

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

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

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

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

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

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

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

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

F.4.3 Proof for DASHA-PP under PŁ-condition

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

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

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

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

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

After rearranging the terms, we get

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

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

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

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

Note that $b=\frac{p_{\mathrm{a}}}{2-p_{\mathrm{a}}}$, thus

$$
\begin{aligned}
&\left(\frac{20b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{4\gamma b\,(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}+\rho\left(\frac{2b^2(1-p_{\mathrm{a}})}{p_{\mathrm{a}}}+(1-b)^2\right)\right)\\
&\le\left(\frac{20b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{4\gamma b\,(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}+\rho\,(1-b)\right).
\end{aligned}
$$

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

$$
\left(\frac{20b^2\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{4\gamma b\,(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}+\rho\,(1-b)\right)\le\left(1-\frac{b}{2}\right)\rho,
$$

and

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right]+\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{4\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\\
&+\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{8\gamma\,(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\le \mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&+(1-\gamma\mu)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+(1-\gamma\mu)\frac{4\gamma((2\omega+1)\,p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{20\gamma\omega\,(2\omega+1)\,\widehat{L}^2}{np_{\mathrm{a}}^2}-\frac{4\gamma\,(p_{\mathrm{a}}-p_{\mathrm{aa}})\,\widehat{L}^2}{bnp_{\mathrm{a}}^2}\right.\\
&\qquad\left.-\frac{80b\gamma\omega(2\omega+1)(1-p_{\mathrm{a}})\widehat{L}^2}{np_{\mathrm{a}}^3}-\frac{16\gamma\,(p_{\mathrm{a}}-p_{\mathrm{aa}})\,(1-p_{\mathrm{a}})\widehat{L}^2}{np_{\mathrm{a}}^3}\right)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&+\left(1-\frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(1-\frac{b}{2}\right)\left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}+\frac{8\gamma\,(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right].
\end{aligned}
$$

Due to $\frac{p_{\mathrm{a}}}{2}\le b\le p_{\mathrm{a}}$, we have

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

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

779 Using Lemma 4 and the assumption about γ, we get

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

780 Note that $\gamma \leq \frac{a}{4\mu} \leq \frac{p_{\mathrm{a}}}{4\mu} \leq \frac{b}{2\mu}$, thus $1 - \frac{b}{2} \leq 1 - \gamma\mu$ and

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

781 In the view of Lemma 11 with

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

782 we can conclude the proof of the theorem. □

783 F.4.4 Proof for DASHA-PP-PAGE under PŁ-condition

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

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

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

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

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

After rearranging the terms, we get

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

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

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

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

789 therefore

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

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

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

If we take $\rho=\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}+\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{page}}}$, then

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

therefore

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right] + \frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1} - h^{t+1}\right\|^2\right] + \frac{4\gamma((2\omega+1)\,p_{\mathrm{a}} - p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right] \\
&\quad + \frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] + \left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}} + \frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \\
&\le \mathrm{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad + (1-\gamma\mu)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t - h^t\right\|^2\right] + (1-\gamma\mu)\frac{4\gamma((2\omega+1)\,p_{\mathrm{a}} - p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t - h_i^t\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{10\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(2\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\right. \\
&\qquad - \frac{2\gamma}{bnp_{\mathrm{a}}}\left(2\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\qquad \left. - \left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^3 p_{\mathrm{page}}} + \frac{8\gamma\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\right)\left(2\left(1-p_{\mathrm{a}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right)\right)\mathrm{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \\
&\quad + \left(1 - \frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right] + \left(1 - \frac{b}{2}\right)\left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}} + \frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2\right].
\end{aligned}
$$

Let us simplify the inequality. First, due to $b \ge \frac{p_{\mathrm{page}} p_{\mathrm{a}}}{2}$, we have

$$
\frac{2\gamma}{bnp_{\mathrm{a}}}\left(2\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \le \frac{8\gamma}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\left(\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right).
$$

Second, due to $b \le p_{\mathrm{a}} p_{\mathrm{page}}$ and $p_{\mathrm{aa}} \le p_{\mathrm{a}}^2$, we get

$$
\begin{aligned}
&\left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^3 p_{\mathrm{page}}} + \frac{8\gamma\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\right)\left(2\left(1-p_{\mathrm{a}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\le \left(\frac{40\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2} + \frac{8\gamma\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\right)\left(2\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\le \frac{80\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\quad + \frac{16\gamma\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\left(\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\le \frac{80\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2}\left(\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right) \\
&\quad + \frac{16\gamma}{np_{\mathrm{a}}^2 p_{\mathrm{page}}}\left(\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2 + \frac{(1-p_{\mathrm{page}})L_{\max}^2}{B}\right).
\end{aligned}
$$

Combining all bounds together, we obtain the following inequality:

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

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

Using Lemma 4 and the assumption about γ, we get

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

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

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

It is left to apply Lemma 11 with

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

to conclude the proof. □

F.4.5 Proof for DASHA-PP-MVR under PŁ-condition

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

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

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

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

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

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

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

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

After rearranging the terms, we get

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

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

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

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

Note that $b \leq \frac{p_{\mathrm{a}}}{2-p_{\mathrm{a}}}$, thus

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

And if we take $\rho = \frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^2} + \frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}$, then

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

and

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

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

$$
\left(\frac{40b\gamma\omega(2\omega+1)}{np_{\mathrm{a}}^3} + \frac{2\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{\mathrm{a}}^2}\right)\left(\frac{2(1-b)^2L_\sigma^2}{B} + 8(1-p_{\mathrm{a}})\widehat{L}^2\right)
$$

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

807 therefore

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

808 Also, we can simplify the last term:

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

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

thus

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

Using Lemma 4 and the assumption about γ, we get

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

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

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

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

812 It is left to apply Lemma 11 with

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

813 and $C = \left(\frac{100b^2\omega(2\omega+1)}{p_{\mathrm{a}}^2} + \frac{20b}{p_{\mathrm{a}}}\right)\frac{\sigma^2}{nB}$ to conclude the proof. □

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

$$\widetilde{\mathcal{O}}\left(\frac{d\sigma}{p_{\mathrm{a}}\sqrt{\mu\varepsilon n}} + \frac{dL_\sigma}{p_{\mathrm{a}}\mu\sqrt{n}}\right),$$

816 *and the expected number of stochastic gradient calculations per node equals*

$$\widetilde{\mathcal{O}}\left(\frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon} + \frac{\sigma L_\sigma}{p_{\mathrm{a}} n\mu^{3/2}\sqrt{\varepsilon}}\right).$$

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

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{p_{\mathrm{a}}} + \frac{\omega}{p_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + \frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma}{p_{\mathrm{a}} n\mu^{3/2}\sqrt{\varepsilon B}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)\right)$$

818 communication rounds in the stochastic settings to get ε-solution. Note that $K = \mathcal{O}\left(\frac{d}{p_{\mathrm{a}}\sqrt{n}}\right)$.
819 Moreover, we can skip the initialization procedure and initialize h_i^0 and g_i^0, for instance, with zeros
820 because the initialization error is under a logarithm. Considering Theorem 6, the communication
821 complexity equals

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

822 The expected number of stochastic gradient calculations per node equals

$$\widetilde{\mathcal{O}}\left(B\frac{\omega+1}{p_{\mathrm{a}}} + B\frac{\omega}{p_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + B\frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon B} + B\frac{L}{\mu} + B\frac{\omega}{p_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + B\frac{\sigma}{p_{\mathrm{a}} n\mu^{3/2}\sqrt{\varepsilon B}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right)\right)$$

$$
\begin{aligned}
&= \widetilde{\mathcal{O}}\left(B\frac{\omega+1}{p_{\mathrm{a}}} + B\frac{\omega}{p_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + B\frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon B} + B\frac{L}{\mu} + B\frac{\omega}{p_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + B\frac{\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon B}}\left(\frac{L_\sigma}{\sqrt{B}}\right)\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{Bd}{Kp_{\mathrm{a}}} + \frac{Bd}{Kp_{\mathrm{a}}}\sqrt{\frac{\sigma^2}{\mu n\varepsilon B}} + \frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon} + B\frac{L}{\mu} + \frac{Bd}{Kp_{\mathrm{a}}\mu\sqrt{n}}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma L_\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon}}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\sigma}{p_{\mathrm{a}}\sqrt{\mu\varepsilon n}} + \frac{\sigma^2}{p_{\mathrm{a}}\mu\varepsilon n\sqrt{B}} + \frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon} + \frac{\sigma L}{p_{\mathrm{a}}\mu^{3/2}\sqrt{\varepsilon}n} + \frac{\sigma}{p_{\mathrm{a}}\mu^{3/2}\sqrt{\varepsilon}n}\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma L_\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon}}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{p_{\mathrm{a}}\mu n\varepsilon} + \frac{\sigma L_\sigma}{p_{\mathrm{a}}n\mu^{3/2}\sqrt{\varepsilon}}\right).
\end{aligned}
$$

$\square$

G Description of DASHA-PP-SYNC-MVR

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

Algorithm 8 DASHA-PP-SYNC-MVR

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 $p_{\text{a}} \in (0,1]$ that a node is *participating*[(a)], number of iterations $T \geq 1$.
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
for $t = 0, 1, \ldots, T-1$ **do**
  $x^{t+1} = x^t - \gamma g^t$
  $c^{t+1} = \begin{cases} 1, \text{with probability } p_{\text{mega}}, \\ 0, \text{with probability } 1 - p_{\text{mega}} \end{cases}$
  Broadcast x^{t+1}, x^t to all *participating*[(a)] nodes
  for $i = 1, \ldots, n$ in parallel **do**
    if i^{th} node is *participating*[(a)] **then**
      if $c^{t+1} = 1$ **then**
        Generate i.i.d. samples $\{\xi_{ik}^{t+1}\}_{k=1}^{B'}$ of size B' from $\mathcal{D}_i$.
        $k_i^{t+1} = \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^t;\xi_{ik}^{t+1}) - \frac{b}{p_{\text{mega}}}\left(h_i^t - \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^t;\xi_{ik}^{t+1})\right)$
        $m_i^{t+1} = \frac{1}{p_{\text{a}}} k_i^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)$
      else
        Generate i.i.d. samples $\{\xi_{ij}^{t+1}\}_{j=1}^{B}$ of size B from $\mathcal{D}_i$.
        $k_i^{t+1} = \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t+1};\xi_{ij}^{t+1}) - \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^t;\xi_{ij}^{t+1})$
        $m_i^{t+1} = \mathcal{C}_i\left(\frac{1}{p_{\text{a}}} k_i^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right)$
      end if
      $h_i^{t+1} = h_i^t + \frac{1}{p_{\text{a}}} k_i^{t+1}$
      $g_i^{t+1} = g_i^t + m_i^{t+1}$
      Send m_i^{t+1} to the server
    else
      $h_i^{t+1} = h_i^t$
      $m_i^{t+1} = 0$
      $g_i^{t+1} = g_i^t$
    end if
  end for
  $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n m_i^{t+1}$
end for
Output: $\hat{x}^T$ chosen uniformly at random from $\{x^t\}_{k=0}^{T-1}$
(a): For the formal description see Section 2.2.

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* $a = \frac{p_{\text{a}}}{2\omega+1}$, $b = \frac{p_{mega} p_{\text{a}}}{2 - p_{\text{a}}}$,*probability* $p_{mega} \in (0,1]$, *batch size* $B' \geq B \geq 1$

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

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

$$\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{T}\left[\frac{2\Delta_0}{\gamma} + \frac{4}{p_{mega}p_{\mathrm{a}}}\left\|h^0 - \nabla f(x^0)\right\|^2 + \frac{4\left(1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)}{np_{mega}p_{\mathrm{a}}}\frac{1}{n}\sum_{i=1}^n\left\|h_i^0 - \nabla f_i(x^0)\right\|^2\right] + \frac{12\sigma^2}{nB'}.$$

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

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

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

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

Corollary 6. *Suppose that assumptions from Theorem 11 hold, probability $p_{mega} = \min\left\{\frac{\zeta_{\mathcal{C}}}{d}, \frac{n\varepsilon B}{\sigma^2}\right\}$, 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 $B_{\text{init}} = \Theta\left(\frac{B}{p_{mega}\sqrt{p_{\mathrm{a}}}}\right) = \Theta\left(\max\left\{\frac{Bd}{\sqrt{p_{\mathrm{a}}}\zeta_{\mathcal{C}}}, \frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}n\varepsilon}\right\}\right)$, then DASHA-PP-SYNC-MVR needs*

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

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

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

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

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

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

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

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

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

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

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

communication rounds to get an ε-solution, the expected communication complexity is equal to $\widetilde{\mathcal{O}}\left(\zeta_{\mathcal{C}} T\right)$, and the expected number of stochastic gradient calculations per node equals $\widetilde{\mathcal{O}}(BT)$, where $\zeta_{\mathcal{C}}$ 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 $\frac{d}{\zeta_{\mathcal{C}}} = \Theta\left(\omega\right)$ (holds for the Rand$K$ compressor), then the convergence rate of DASHA-PP-SYNC-MVR is

$$\widetilde{\mathcal{O}}\left(\frac{\omega + 1}{p_{\mathrm{a}}} + \frac{\sigma^2}{p_{\mathrm{a}} \mu n \varepsilon B} + \frac{L}{\mu} + \frac{\omega}{p_{\mathrm{a}} \mu \sqrt{n}}\left(\frac{L_\sigma}{\sqrt{B}} + \widehat{L}\right) + \frac{\sigma}{p_{\mathrm{a}} n \mu^{3/2} \sqrt{\varepsilon B}}\left(\frac{L_\sigma}{\sqrt{B}} + \widehat{L}\right)\right). \tag{31}$$

Comparing (31) with the rate of DASHA-PP-MVR (28), one can see that DASHA-PP-SYNC-MVR improves the suboptimal term $\mathcal{P}_2$ from (28). However, Corollary 5 reveals that we can escape these suboptimal regimes by choosing the parameter K of Rand$K$ 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 $\xi_\cdot^{t+1}$, 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 $\mathrm{E}_k\left[\cdot\right]$, $\mathrm{E}_{\mathcal{C}}\left[\cdot\right]$, $\mathrm{E}_{p_{\mathrm{a}}}\left[\cdot\right]$ and $\mathrm{E}_{p_{\mathrm{mega}}}\left[\cdot\right]$ to be conditional expectations w.r.t. $\xi_\cdot^{t+1}$, $\{\mathcal{C}_i\}_{i=1}^n$, availability, and c^{t+1}, accordingly, conditioned on all previous randomness. Moreover, we define $\mathrm{E}_{t+1}\left[\cdot\right]$ to be a conditional expectation w.r.t. all randomness in iteration $t+1$ conditioned on all previous randomness.

Let us denote

$$k_{i,1}^{t+1} := \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t+1}; \xi_{ik}^{t+1}) - \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t}; \xi_{ik}^{t+1}) - \frac{b}{p_{\mathrm{mega}}}\left(h_i^t - \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t}; \xi_{ik}^{t+1})\right),$$

$$k_{i,2}^{t+1} := \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t+1}; \xi_{ij}^{t+1}) - \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t}; \xi_{ij}^{t+1}),$$

$$h_{i,1}^{t+1} := \begin{cases} h_i^t + \frac{1}{p_{\mathrm{a}}} k_{i,1}^{t+1}, & i^{\text{th}} \text{ node is } \textit{participating}, \\ h_i^t, & \text{otherwise}, \end{cases}$$

$$h_{i,2}^{t+1} := \begin{cases} h_i^t + \frac{1}{p_{\mathrm{a}}} k_{i,2}^{t+1}, & i^{\text{th}} \text{ node is } \textit{participating}, \\ h_i^t, & \text{otherwise}, \end{cases}$$

$$g_{i,1}^{t+1} := \begin{cases} g_i^t + \frac{1}{p_{\mathrm{a}}} k_{i,1}^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right), & i^{\text{th}} \text{ node is } \textit{participating}, \\ g_i^t, & \text{otherwise}, \end{cases}$$

$$g_{i,2}^{t+1} := \begin{cases} g_i^t + \mathcal{C}_i\left(\frac{1}{p_{\mathrm{a}}} k_{i,2}^{t+1} - \frac{a}{p_{\mathrm{a}}}\left(g_i^t - h_i^t\right)\right), & i^{\text{th}} \text{ node is } \textit{participating}, \\ g_i^t, & \text{otherwise}, \end{cases}$$

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

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

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

and

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

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

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

Using

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

and

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

we have

$$\begin{aligned}
&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\mathrm{mega}}}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]\right]\right] \\
&\overset{(13)}{=}p_{\mathrm{mega}}\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_1^{t+1}-h_1^{t+1}-\mathrm{E}_{p_{\mathrm{a}}}\left[g_1^{t+1}-h_1^{t+1}\right]\right\|^2\right] \\
&\quad+\left(1-p_{\mathrm{mega}}\right)\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|g_2^{t+1}-h_2^{t+1}-\mathrm{E}_{p_{\mathrm{a}}}\left[g_2^{t+1}-h_2^{t+1}\right]\right\|^2\right]\right] \\
&\quad+(1-a)^2\left\|g^t-h^t\right\|^2.
\end{aligned}$$

We can use Lemma 1 two times with i) $r_i=g_i^t-h_i^t$ and $s_i=-a\left(g_i^t-h_i^t\right)$ and ii) $r_i=g_i^t-h_i^t$ and $s_i=p_{\mathrm{a}}\mathcal{C}_i\left(\frac{1}{p_{\mathrm{a}}}k_{i,2}^{t+1}-\frac{a}{p_{\mathrm{a}}}\left(g_i^t-h_i^t\right)\right)-k_{i,2}^{t+1}$, to obtain

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

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

In the last inequality, we use Assumption 7. Next, using (12), we have

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

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

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right]\right]\right] \\
={}& p_{\text{mega}}\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,1}^{t+1} - h_{i,1}^{t+1}\right\|^2\right] + \left(1-p_{\text{mega}}\right)\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,2}^{t+1} - h_{i,2}^{t+1}\right\|^2\right]\right] \\
\overset{(13)}{=}{}& p_{\text{mega}}\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,1}^{t+1} - h_{i,1}^{t+1} - (1-a)\left(g_i^t - h_i^t\right)\right\|^2\right] + \left(1-p_{\text{mega}}\right)\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,2}^{t+1} - h_{i,2}^{t+1}\right\|^2\right]\right] \\
&+ p_{\text{mega}}(1-a)^2\left\|g_i^t - h_i^t\right\|^2 \\
={}& \frac{p_{\text{mega}}(1-p_{\text{a}})a^2}{p_{\text{a}}}\left\|g_i^t - h_i^t\right\|^2 + \left(1-p_{\text{mega}}\right)\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,2}^{t+1} - h_{i,2}^{t+1}\right\|^2\right]\right] \\
&+ p_{\text{mega}}(1-a)^2\left\|g_i^t - h_i^t\right\|^2 \\
\overset{(13)}{=}{}& \frac{p_{\text{mega}}(1-p_{\text{a}})a^2}{p_{\text{a}}}\left\|g_i^t - h_i^t\right\|^2 + \left(1-p_{\text{mega}}\right)\mathrm{E}_{\mathcal{C}}\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|g_{i,2}^{t+1} - h_{i,2}^{t+1} - (1-a)\left(g_i^t - h_i^t\right)\right\|^2\right]\right] \\
&+ (1-a)^2\left\|g_i^t - h_i^t\right\|^2 \\
={}& \frac{p_{\text{mega}}(1-p_{\text{a}})a^2}{p_{\text{a}}}\left\|g_i^t - h_i^t\right\|^2 \\
&+ \left(1-p_{\text{mega}}\right)p_{\text{a}}\mathrm{E}_{\mathcal{C}}\left[\left\|g_i^t + \mathcal{C}_i\left(\frac{1}{p_{\text{a}}}k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right) - \left(h_i^t + \frac{1}{p_{\text{a}}}k_{i,2}^{t+1}\right) - (1-a)\left(g_i^t - h_i^t\right)\right\|^2\right] \\
&+ \left(1-p_{\text{mega}}\right)\left(1-p_{\text{a}}\right)\left\|g_i^t - h_i^t - (1-a)\left(g_i^t - h_i^t\right)\right\|^2 \\
&+ (1-a)^2\left\|g_i^t - h_i^t\right\|^2 \\
={}& \frac{p_{\text{mega}}(1-p_{\text{a}})a^2}{p_{\text{a}}}\left\|g_i^t - h_i^t\right\|^2 \\
&+ \left(1-p_{\text{mega}}\right)p_{\text{a}}\mathrm{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(\frac{1}{p_{\text{a}}}k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right) - \left(\frac{1}{p_{\text{a}}}k_{i,2}^{t+1} - a\left(g_i^t - h_i^t\right)\right)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \left(1 - p_{\text{mega}}\right)\left(1 - p_{\text{a}}\right) a^2 \left\|g_i^t - h_i^t\right\|^2 \\
&+ (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
\overset{(13)}{=}& \left(\frac{p_{\text{mega}}(1-p_{\text{a}})a^2}{p_{\text{a}}} + \frac{\left(1 - p_{\text{mega}}\right)\left(1 - p_{\text{a}}\right) a^2}{p_{\text{a}}}\right) \left\|g_i^t - h_i^t\right\|^2 \\
&+ \left(1 - p_{\text{mega}}\right) p_{\text{a}} \mathrm{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(\frac{1}{p_{\text{a}}} k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right) - \left(\frac{1}{p_{\text{a}}} k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right)\right\|^2\right] \\
&+ (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
=& \frac{(1-p_{\text{a}})a^2}{p_{\text{a}}} \left\|g_i^t - h_i^t\right\|^2 \\
&+ \left(1 - p_{\text{mega}}\right) p_{\text{a}} \mathrm{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(\frac{1}{p_{\text{a}}} k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right) - \left(\frac{1}{p_{\text{a}}} k_{i,2}^{t+1} - \frac{a}{p_{\text{a}}}\left(g_i^t - h_i^t\right)\right)\right\|^2\right] \\
&+ (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
\leq& \frac{(1-p_{\text{a}})a^2}{p_{\text{a}}} \left\|g_i^t - h_i^t\right\|^2 \\
&+ \frac{\left(1 - p_{\text{mega}}\right)\omega}{p_{\text{a}}} \left\|k_{i,2}^{t+1} - a\left(g_i^t - h_i^t\right)\right\|^2 \\
&+ (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 \\
\overset{(12)}{\leq}& \frac{2\left(1 - p_{\text{mega}}\right)\omega}{p_{\text{a}}} \left\|k_{i,2}^{t+1}\right\|^2 + \left(\frac{(1-p_{\text{a}})a^2}{p_{\text{a}}} + \frac{2\left(1 - p_{\text{mega}}\right) a^2 \omega}{p_{\text{a}}}\right) \left\|g_i^t - h_i^t\right\|^2 \\
&+ (1-a)^2 \left\|g_i^t - h_i^t\right\|^2 .
\end{aligned}
$$

□

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

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

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

and

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

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

$$
\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[h_{i,1}^{t+1}\right]\right]
$$

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

$$= h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}}\left(h_i^t - \nabla f_i(x^t)\right)$$

and

$$\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[h_{i,2}^{t+1}\right]\right] \\
&= h_i^t + \mathrm{E}_k\left[\frac{1}{B}\sum_{j=1}^{B}\nabla f_i(x^{t+1};\xi_{ij}^{t+1}) - \frac{1}{B}\sum_{j=1}^{B}\nabla f_i(x^t;\xi_{ij}^{t+1})\right] \\
&= h_i^t + \nabla f_i(x^{t+1}) - \nabla f_i(x^t),
\end{aligned}$$

we have

$$\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right]\right] \\
&= p_{\text{mega}}\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|h_1^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right] + (1-p_{\text{mega}})\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|h_2^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right] \\
&\overset{(13)}{=} p_{\text{mega}}\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|h_1^{t+1} - \mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[h_1^{t+1}\right]\right]\right\|^2\right]\right] + (1-p_{\text{mega}})\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\left\|h_2^{t+1} - \mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[h_2^{t+1}\right]\right]\right\|^2\right]\right] \\
&\quad + \left(p_{\text{mega}}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}})\right)\left\|h^t - \nabla f(x^t)\right\|^2.
\end{aligned}$$

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 obtain

$$\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right]\right] \\
&\leq p_{\text{mega}}\left(\frac{1}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|k_{i,1}^{t+1} - \mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right] + \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2}\sum_{i=1}^n\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}}\left(h_i^t - \nabla f_i(x^t)\right)\right\|^2\right) \\
&\quad + (1-p_{\text{mega}})\left(\frac{1}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|k_{i,2}^{t+1} - \mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right] + \frac{p_{\text{a}} - p_{\text{aa}}}{n^2 p_{\text{a}}^2}\sum_{i=1}^n\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right) \\
&\quad + \left(p_{\text{mega}}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}})\right)\left\|h^t - \nabla f(x^t)\right\|^2 \\
&\overset{(12)}{\leq} \frac{p_{\text{mega}}}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|k_{i,1}^{t+1} - \mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right] \\
&\quad + \frac{1-p_{\text{mega}}}{n^2 p_{\text{a}}}\sum_{i=1}^n \mathrm{E}_k\left[\left\|k_{i,2}^{t+1} - \mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right] \\
&\quad + \frac{2\left(p_{\text{a}} - p_{\text{aa}}\right)}{n^2 p_{\text{a}}^2}\sum_{i=1}^n\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 \\
&\quad + \frac{2\left(p_{\text{a}} - p_{\text{aa}}\right)b^2}{n^2 p_{\text{a}}^2 p_{\text{mega}}}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2 + \left(p_{\text{mega}}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 + (1-p_{\text{mega}})\right)\left\|h^t - \nabla f(x^t)\right\|^2.
\end{aligned} \tag{32}$$

Let us consider $\mathrm{E}_k\left[\left\|k_{i,1}^{t+1} - \mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right]$.

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

$$
\left. - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t) - \frac{b}{p_{\text{mega}}}\left(h_i^t - \nabla f_i(x^t)\right)\right)\right\|^2 \Bigg]
$$

$$
= \mathrm{E}_k \left[\left\| \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t};\xi_{ik}^{t+1}) + \frac{b}{p_{\text{mega}}}\left(\frac{1}{B'}\sum_{k=1}^{B'} \nabla f_i(x^{t};\xi_{ik}^{t+1})\right)\right.\right.
$$

$$
\left.\left. - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t) + \frac{b}{p_{\text{mega}}}\left(\nabla f_i(x^t)\right)\right)\right\|^2\right]
$$

$$
= \frac{1}{B'^2}\sum_{k=1}^{B'} \mathrm{E}_k \left[\left\| \frac{b}{p_{\text{mega}}}\left(\nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \nabla f_i(x^{t+1})\right)\right.\right.
$$

$$
\left.\left. + \left(1 - \frac{b}{p_{\text{mega}}}\right)\left(\nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \nabla f_i(x^{t};\xi_{ik}^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right)\right\|^2\right],
$$

893 where we used independence of the mini-batch samples. Using (12), we get

$$
\mathrm{E}_k \left[\left\|k_{i,1}^{t+1} - \mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right]
$$

$$
\leq \frac{2b^2}{B'^2 p_{\text{mega}}^2}\sum_{k=1}^{B'} \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right]
$$

$$
+ \frac{2}{B'^2}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 \sum_{k=1}^{B'} \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ik}^{t+1}) - \nabla f_i(x^{t};\xi_{ik}^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right].
$$

894 Due to Assumptions 5 and 6, we have

$$
\mathrm{E}_k \left[\left\|k_{i,1}^{t+1} - \mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right] \leq \frac{2b^2\sigma^2}{B' p_{\text{mega}}^2} + \frac{2L_\sigma^2}{B'}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 \left\|x^{t+1} - x^t\right\|^2. \tag{33}
$$

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

$$
\mathrm{E}_k \left[\left\|k_{i,2}^{t+1} - \mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right]
$$

$$
= \mathrm{E}_k \left[\left\| \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t+1};\xi_{ij}^{t+1}) - \frac{1}{B}\sum_{j=1}^{B} \nabla f_i(x^{t};\xi_{ij}^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right]
$$

$$
= \frac{1}{B^2}\sum_{j=1}^{B} \mathrm{E}_k\left[\left\|\nabla f_i(x^{t+1};\xi_{ij}^{t+1}) - \nabla f_i(x^{t};\xi_{ij}^{t+1}) - \left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)\right\|^2\right].
$$

896 Due to Assumptions 6, we have

$$
\mathrm{E}_k \left[\left\|k_{i,2}^{t+1} - \mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right] \leq \frac{L_\sigma^2}{B}\left\|x^{t+1} - x^t\right\|^2. \tag{34}
$$

897 Plugging (33) and (34) into (32), we obtain

$$
\mathrm{E}_k \left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right]\right]\right]
$$

$$
\leq \frac{p_{\text{mega}}}{n p_{\text{a}}}\left(\frac{2b^2\sigma^2}{B' p_{\text{mega}}^2} + \frac{2L_\sigma^2}{B'}\left(1 - \frac{b}{p_{\text{mega}}}\right)^2 \left\|x^{t+1} - x^t\right\|^2\right)
$$

$$
+ \frac{\left(1 - p_{\text{mega}}\right) L_\sigma^2}{n p_{\text{a}} B}\left\|x^{t+1} - x^t\right\|^2
$$

$$
+ \frac{2\left(p_{\text{a}} - p_{\text{aa}}\right)}{n^2 p_{\text{a}}^2}\sum_{i=1}^{n}\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2
$$

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

Using Assumption 3, we get

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

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

$$
\begin{aligned}
&\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\right] \\
&= p_{\text{mega}} \mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]+(1-p_{\text{mega}}) \mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right] \\
&\overset{(13)}{=} p_{\text{mega}} \mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,1}^{t+1}-\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,1}^{t+1}\right]\right]\right\|^2\right]\right]+(1-p_{\text{mega}}) \mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[\left\|h_{i,2}^{t+1}-\mathrm{E}_k\left[\mathrm{E}_{p_{\mathrm{a}}}\left[h_{i,2}^{t+1}\right]\right]\right\|^2\right]\right] \\
&\quad+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+\left(1-p_{\text{mega}}\right)\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\
&= p_{\text{mega}} p_{\mathrm{a}} \mathrm{E}_k\left[\left\|h_i^t+\frac{1}{p_{\mathrm{a}}} k_{i,1}^{t+1}-\left(h_i^t+\mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right)\right\|^2\right] \\
&\quad+p_{\text{mega}}\left(1-p_{\mathrm{a}}\right)\left\|h_i^t-\left(h_i^t+\mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right)\right\|^2 \\
&\quad+(1-p_{\text{mega}}) p_{\mathrm{a}} \mathrm{E}_k\left[\left\|h_i^t+\frac{1}{p_{\mathrm{a}}} k_{i,2}^{t+1}-\left(h_i^t+\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right)\right\|^2\right] \\
&\quad+(1-p_{\text{mega}})(1-p_{\mathrm{a}})\left\|h_i^t-\left(h_i^t+\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right)\right\|^2 \\
&\quad+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+\left(1-p_{\text{mega}}\right)\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\
&= p_{\text{mega}} p_{\mathrm{a}} \mathrm{E}_k\left[\left\|\frac{1}{p_{\mathrm{a}}} k_{i,1}^{t+1}-\mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right] \\
&\quad+p_{\text{mega}}\left(1-p_{\mathrm{a}}\right)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-\frac{b}{p_{\text{mega}}}\left(h_i^t-\nabla f_i(x^t)\right)\right\|^2 \\
&\quad+(1-p_{\text{mega}}) p_{\mathrm{a}} \mathrm{E}_k\left[\left\|\frac{1}{p_{\mathrm{a}}} k_{i,2}^{t+1}-\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right] \\
&\quad+(1-p_{\text{mega}})(1-p_{\mathrm{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+\left(1-p_{\text{mega}}\right)\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\
&\overset{(13)}{=} \frac{p_{\text{mega}}}{p_{\mathrm{a}}} \mathrm{E}_k\left[\left\|k_{i,1}^{t+1}-\mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right] \\
&\quad+\frac{(1-p_{\text{mega}})}{p_{\mathrm{a}}} \mathrm{E}_k\left[\left\|k_{i,2}^{t+1}-\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right] \\
&\quad+\frac{p_{\text{mega}}\left(1-p_{\mathrm{a}}\right)}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)-\frac{b}{p_{\text{mega}}}\left(h_i^t-\nabla f_i(x^t)\right)\right\|^2 \\
&\quad+\frac{(1-p_{\text{mega}})(1-p_{\mathrm{a}})}{p_{\mathrm{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2
\end{aligned}
$$

$$+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+(1-p_{\text{mega}})\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2$$

$$\overset{(12)}{\leq}\frac{p_{\text{mega}}}{p_{\text{a}}}\mathrm{E}_k\left[\left\|k_{i,1}^{t+1}-\mathrm{E}_k\left[k_{i,1}^{t+1}\right]\right\|^2\right]$$
$$+\frac{(1-p_{\text{mega}})}{p_{\text{a}}}\mathrm{E}_k\left[\left\|k_{i,2}^{t+1}-\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right]$$
$$+\frac{2(1-p_{\text{a}})}{p_{\text{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2$$
$$+\frac{2\left(1-p_{\text{a}}\right)b^2}{p_{\text{mega}}p_{\text{a}}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+(1-p_{\text{mega}})\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2.$$

900 Using (33) and (34), we get

$$\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\right]$$
$$\leq\frac{2b^2\sigma^2}{p_{\text{a}}p_{\text{mega}}B'}+\frac{2p_{\text{mega}}L_\sigma^2}{p_{\text{a}}B'}\left(1-\frac{b}{p_{\text{mega}}}\right)^2\left\|x^{t+1}-x^t\right\|^2$$
$$+\frac{(1-p_{\text{mega}})L_\sigma^2}{p_{\text{a}}B}\left\|x^{t+1}-x^t\right\|^2$$
$$+\frac{2(1-p_{\text{a}})}{p_{\text{a}}}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2$$
$$+\frac{2\left(1-p_{\text{a}}\right)b^2}{p_{\text{mega}}p_{\text{a}}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+(1-p_{\text{mega}})\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2.$$

901 Next, due to Assumption 3, we obtain

$$\mathrm{E}_k\left[\mathrm{E}_{p_{\text{a}}}\left[\mathrm{E}_{p_{\text{mega}}}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right]\right]$$
$$\leq\frac{2b^2\sigma^2}{p_{\text{a}}p_{\text{mega}}B'}+\left(\frac{2p_{\text{mega}}L_\sigma^2}{p_{\text{a}}B'}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+\frac{(1-p_{\text{mega}})L_\sigma^2}{p_{\text{a}}B}+\frac{2(1-p_{\text{a}})L_i^2}{p_{\text{a}}}\right)\left\|x^{t+1}-x^t\right\|^2$$
$$+\frac{2\left(1-p_{\text{a}}\right)b^2}{p_{\text{mega}}p_{\text{a}}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2+\left(p_{\text{mega}}\left(1-\frac{b}{p_{\text{mega}}}\right)^2+(1-p_{\text{mega}})\right)\left\|h_i^t-\nabla f_i(x^t)\right\|^2.$$

902 The third inequality can be proved with the help of (34) and Assumption 3.

$$\mathrm{E}_k\left[\left\|k_{i,2}^{t+1}\right\|^2\right]$$
$$\overset{(13)}{=}\mathrm{E}_k\left[\left\|k_{i,2}^{t+1}-\mathrm{E}_k\left[k_{i,2}^{t+1}\right]\right\|^2\right]+\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2$$
$$\leq\frac{L_\sigma^2}{B}\left\|x^{t+1}-x^t\right\|^2+\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2$$
$$\leq\left(\frac{L_\sigma^2}{B}+L_i^2\right)\left\|x^{t+1}-x^t\right\|^2.$$

903 □

Theorem 11. *Suppose that Assumptions 1, 2, 3, 5, 6, 7 and 8 hold. Let us take* $a=\frac{p_{\text{a}}}{2\omega+1}$, $b=\frac{p_{mega}p_{\text{a}}}{2-p_{\text{a}}}$*,probability* $p_{mega}\in(0,1]$*, batch size* $B'\geq B\geq1$

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

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

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

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

905 *Proof.* Due to Lemma 2 and the update step from Line 5 in Algorithm 8, we have

$$\begin{aligned}
&\mathrm{E}_{t+1}\left[f(x^{t+1})\right] \\
&\leq \mathrm{E}_{t+1}\left[f(x^t)-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\frac{\gamma}{2}\left\|g^t-\nabla f(x^t)\right\|^2\right] \\
&= \mathrm{E}_{t+1}\left[f(x^t)-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\frac{\gamma}{2}\left\|g^t-h^t+h^t-\nabla f(x^t)\right\|^2\right] \\
&\overset{(13)}{\leq} \mathrm{E}_{t+1}\left[f(x^t)-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\gamma\left(\left\|g^t-h^t\right\|^2+\left\|h^t-\nabla f(x^t)\right\|^2\right)\right].
\end{aligned}$$

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

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

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

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

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

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

and

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

Thus

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

After rearranging the terms, we get

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

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

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

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

Next, since $a = \frac{p_{\mathrm{a}}}{2\omega+1}$, we have $\left(\frac{(2\omega+1-p_{\mathrm{a}})a^2}{p_{\mathrm{a}}} + (1-a)^2\right) \leq 1-a$. We the choice $\eta = \frac{\gamma((2\omega+1)p_{\mathrm{a}} - p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}$, we guarantee $\frac{\gamma((2\omega+1)p_{\mathrm{a}} - p_{\mathrm{aa}})a}{np_{\mathrm{a}}^2} + \eta\left(\frac{(2\omega+1-p_{\mathrm{a}})a^2}{p_{\mathrm{a}}} + (1-a)^2\right) \leq \eta$ and

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

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

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

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

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

914 Next, we take $\rho=\frac{2\gamma(p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^{2}p_{\mathrm{mega}}}$, thus

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

915 Since $\frac{p_{\mathrm{mega}}p_{\mathrm{a}}}{2}\le b\le p_{\mathrm{mega}}p_{\mathrm{a}}$ and $1-p_{\mathrm{a}}\le 1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\le 1$, we get

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

$$
-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{4\gamma\left(2\omega+1\right)\omega}{np_{\mathrm{a}}^2}\left(\frac{L_\sigma^2}{B}+\widehat{L}^2\right)-\left(\frac{8\gamma L_\sigma^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2B}+\frac{8\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2}\right)\right)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]
$$

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

$$
+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]
$$

$$
+\frac{6\gamma\sigma^2}{nB'}.
$$

Using Lemma 4 and the assumption about γ, we get

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

$$
+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]
$$

$$
\leq\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]
$$

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

$$
+\frac{\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\frac{2\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]
$$

$$
+\frac{6\gamma\sigma^2}{nB'}.
$$

It is left to apply Lemma 3 with

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

and $C=\frac{6\sigma^2}{nB'}$ to conclude the proof. □

Corollary 6. *Suppose that assumptions from Theorem 11 hold, probability $p_{mega}=\min\left\{\frac{\zeta_{\mathcal{C}}}{d},\frac{n\varepsilon B}{\sigma^2}\right\}$, 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 $B_{\text{init}}=\Theta\left(\frac{B}{p_{mega}\sqrt{p_{\mathrm{a}}}}\right)=\Theta\left(\max\left\{\frac{Bd}{\sqrt{p_{\mathrm{a}}}\zeta_{\mathcal{C}}},\frac{\sigma^2}{\sqrt{p_{\mathrm{a}}}n\varepsilon}\right\}\right)$, then DASHA-PP-SYNC-MVR needs*

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

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

Proof. Due to the choice of B', we have

$$\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{T}\left[2\Delta_0\left(L + \sqrt{\frac{8\left(2\omega+1\right)\omega}{np_{\text{a}}^2}\left(\widehat{L}^2 + \frac{L_\sigma^2}{B}\right) + \frac{16}{np_{\text{mega}}p_{\text{a}}^2}\left(\left(1 - \frac{p_{\text{aa}}}{p_{\text{a}}}\right)\widehat{L}^2 + \frac{L_\sigma^2}{B}\right)}\right)\right.$$
$$\left. + \frac{4}{p_{\text{mega}}p_{\text{a}}}\left\|h^0 - \nabla f(x^0)\right\|^2 + \frac{4\left(1 - \frac{p_{\text{aa}}}{p_{\text{a}}}\right)}{np_{\text{mega}}p_{\text{a}}}\frac{1}{n}\sum_{i=1}^n\left\|h_i^0 - \nabla f_i(x^0)\right\|^2\right]$$
$$+ \frac{2\varepsilon}{3}.$$

Using

$$\mathrm{E}\left[\left\|h^0 - \nabla f(x^0)\right\|^2\right] = \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\frac{1}{B_{\text{init}}}\sum_{k=1}^{B_{\text{init}}}\nabla f_i(x^0;\xi_{ik}^0) - \nabla f(x^0)\right\|^2\right] \leq \frac{\sigma^2}{nB_{\text{init}}}$$

and

$$\frac{1}{n^2}\sum_{i=1}^n\mathrm{E}\left[\left\|h_i^0 - \nabla f_i(x^0)\right\|^2\right] = \frac{1}{n^2}\sum_{i=1}^n\mathrm{E}\left[\left\|\frac{1}{B_{\text{init}}}\sum_{k=1}^{B_{\text{init}}}\nabla f_i(x^0;\xi_{ik}^0) - \nabla f_i(x^0)\right\|^2\right] \leq \frac{\sigma^2}{nB_{\text{init}}},$$

we have

$$\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{T}\left[2\Delta_0\left(L + \sqrt{\frac{8\left(2\omega+1\right)\omega}{np_{\text{a}}^2}\left(\widehat{L}^2 + \frac{L_\sigma^2}{B}\right) + \frac{16}{np_{\text{mega}}p_{\text{a}}^2}\left(\left(1 - \frac{p_{\text{aa}}}{p_{\text{a}}}\right)\widehat{L}^2 + \frac{L_\sigma^2}{B}\right)}\right)\right.$$
$$\left. + \frac{8\sigma^2}{np_{\text{mega}}p_{\text{a}}B_{\text{init}}}\right]$$
$$+ \frac{2\varepsilon}{3}.$$

Therefore, we can take the following T to get ε–solution.

$$T = \mathcal{O}\left(\frac{1}{\varepsilon}\left[\Delta_0\left(L + \sqrt{\frac{\omega^2}{np_{\text{a}}^2}\left(\widehat{L}^2 + \frac{L_\sigma^2}{B}\right) + \frac{1}{np_{\text{mega}}p_{\text{a}}^2}\left(\widehat{L}^2 + \frac{L_\sigma^2}{B}\right)}\right) + \frac{\sigma^2}{np_{\text{mega}}p_{\text{a}}B_{\text{init}}}\right]\right)$$

Considering the choice of p_{mega} and B_{init}, we obtain

$$T = \mathcal{O}\left(\frac{1}{\varepsilon}\left[\Delta_0\left(L + \left(\frac{\omega}{p_{\text{a}}\sqrt{n}} + \sqrt{\frac{d}{p_{\text{a}}^2\zeta_{\mathcal{C}}n}}\right)\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma}{p_{\text{a}}\sqrt{\varepsilon}n}\left(\frac{\widehat{L}}{\sqrt{B}} + \frac{L_\sigma}{B}\right)\right) + \frac{\sigma^2}{np_{\text{mega}}p_{\text{a}}B_{\text{init}}}\right]\right)$$
$$= \mathcal{O}\left(\frac{\Delta_0}{\varepsilon}\left[L + \left(\frac{\omega}{p_{\text{a}}\sqrt{n}} + \sqrt{\frac{d}{p_{\text{a}}^2\zeta_{\mathcal{C}}n}}\right)\left(\widehat{L} + \frac{L_\sigma}{\sqrt{B}}\right) + \frac{\sigma}{p_{\text{a}}\sqrt{\varepsilon}n}\left(\frac{\widehat{L}}{\sqrt{B}} + \frac{L_\sigma}{B}\right)\right] + \frac{\sigma^2}{\sqrt{p_{\text{a}}}n\varepsilon B}\right).$$

The expected communication complexity equals $\mathcal{O}\left(d + p_{\text{mega}}d + (1 - p_{\text{mega}})\zeta_{\mathcal{C}}\right) = \mathcal{O}\left(d + \zeta_{\mathcal{C}}\right)$ and the expected number of stochastic gradient calculations per node equals $\mathcal{O}\left(B_{\text{init}} + p_{\text{mega}}B' + (1 - p_{\text{mega}})B\right) = \mathcal{O}\left(B_{\text{init}} + B\right).$ □

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

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

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

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

Proof. Let us fix constants $\kappa, \eta, \nu, \rho \in [0, \infty)$ that we will define later. As in the proof of Theorem 11, we can get

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

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

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

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

Next, since $a=\frac{p_{\mathrm{a}}}{2\omega+1}$, we have $\left(\frac{(2\omega+1-p_{\mathrm{a}})a^2}{p_{\mathrm{a}}}+(1-a)^2\right)\leq 1-a$. We the choice $\eta=\frac{2\gamma((2\omega+1)p_{\mathrm{a}}-p_{\mathrm{aa}})}{np_{\mathrm{a}}^2}$, we guarantee $\frac{\gamma((2\omega+1)p_{\mathrm{a}}-p_{\mathrm{aa}})a}{np_{\mathrm{a}}^2}+\eta\left(\frac{(2\omega+1-p_{\mathrm{a}})a^2}{p_{\mathrm{a}}}+(1-a)^2\right)\leq\left(1-\frac{a}{2}\right)\eta$ and

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right]+\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\\
&\quad+\nu\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\rho\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\leq\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad-\Bigg(\frac{1}{2\gamma}-\frac{L}{2}-\frac{8\gamma\left(2\omega+1\right)\omega}{np_{\mathrm{a}}^2}\left(\frac{L_\sigma^2}{B}+\widehat{L}^2\right)\\
&\qquad-\nu\left(\frac{2L_\sigma^2}{np_{\mathrm{a}}B}+\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{a}}^2}\right)-\rho\left(\frac{2L_\sigma^2}{p_{\mathrm{a}}B}+\frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\Bigg)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&\quad+\left(1-\frac{a}{2}\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]\\
&\quad+\left(1-\frac{a}{2}\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]\\
&\quad+\left(\gamma+\nu\left(1-b\right)\right)\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]\\
&\quad+\left(\nu\frac{2\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)b^2}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}+\rho(1-b)\right)\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\\
&\quad+\left(\frac{2\nu b^2}{np_{\mathrm{mega}}p_{\mathrm{a}}}+\frac{2\rho b^2}{p_{\mathrm{a}}p_{\mathrm{mega}}}\right)\frac{\sigma^2}{B'},
\end{aligned}
$$

where simplified the term using $p_{\mathrm{aa}}\geq 0$. Let us take $\nu=\frac{2\gamma}{b}$ to obtain

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

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

938 Next, we take $\rho = \frac{8\gamma(p_{\mathrm{a}} - p_{\mathrm{aa}})}{np_{\mathrm{a}}^2 p_{\mathrm{mega}}}$, thus

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right] + \frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1} - h^{t+1}\right\|^2\right] + \frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right] \\
&+ \frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] + \frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \\
&\le \mathrm{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&- \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\left(2\omega+1\right)\omega}{np_{\mathrm{a}}^2}\left(\frac{L_{\sigma}^2}{B} + \widehat{L}^2\right)\right. \\
&\qquad \left. - \left(\frac{4\gamma L_{\sigma}^2}{bnp_{\mathrm{a}}B} + \frac{4\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)\widehat{L}^2}{bnp_{\mathrm{a}}^2}\right) - \left(\frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{mega}}}\right)\left(\frac{2L_{\sigma}^2}{p_{\mathrm{a}}B} + \frac{2(1-p_{\mathrm{a}})\widehat{L}^2}{p_{\mathrm{a}}}\right)\right)\mathrm{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \\
&+ \left(1 - \frac{a}{2}\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t - h^t\right\|^2\right] + \left(1 - \frac{a}{2}\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t - h_i^t\right\|^2\right] \\
&+ \left(1 - \frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right] + \left(1 - \frac{b}{2}\right)\frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2\right] \\
&+ \left(\frac{4\gamma b}{np_{\mathrm{mega}}p_{\mathrm{a}}} + \frac{16\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)b^2}{np_{\mathrm{a}}^3 p_{\mathrm{mega}}^2}\right)\frac{\sigma^2}{B'},
\end{aligned}
$$

939 Since $\frac{p_{\mathrm{mega}}p_{\mathrm{a}}}{2} \le b \le p_{\mathrm{mega}}p_{\mathrm{a}}$ and $1 - p_{\mathrm{a}} \le 1 - \frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}} \le 1$, we get

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})\right] + \frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1} - h^{t+1}\right\|^2\right] + \frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1} - h_i^{t+1}\right\|^2\right] \\
&+ \frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1} - \nabla f(x^{t+1})\right\|^2\right] + \frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2 p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \\
&\le \mathrm{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&- \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\left(2\omega+1\right)\omega}{np_{\mathrm{a}}^2}\left(\frac{L_{\sigma}^2}{B} + \widehat{L}^2\right)\right. \\
&\qquad \left. - \left(\frac{8\gamma L_{\sigma}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2 B} + \frac{8\gamma\left(p_{\mathrm{a}} - p_{\mathrm{aa}}\right)\widehat{L}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^3}\right) - \left(\frac{16\gamma L_{\sigma}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2 B} + \frac{16\gamma(1-p_{\mathrm{a}})\widehat{L}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2}\right)\right)\mathrm{E}\left[\left\|x^{t+1} - x^t\right\|^2\right]
\end{aligned}
$$

$$+\left(1-\frac{a}{2}\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\left(1-\frac{a}{2}\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]$$
$$+\left(1-\frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(1-\frac{b}{2}\right)\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+\frac{20\gamma\sigma^2}{nB'}$$
$$\leq\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$$
$$-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{8\gamma\left(2\omega+1\right)\omega}{np_{\mathrm{a}}^2}\left(\frac{L_\sigma^2}{B}+\widehat{L}^2\right)-\left(\frac{24\gamma L_\sigma^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2B}+\frac{24\gamma\left(1-\frac{p_{\mathrm{aa}}}{p_{\mathrm{a}}}\right)\widehat{L}^2}{np_{\mathrm{mega}}p_{\mathrm{a}}^2}\right)\right)\mathrm{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]$$
$$+\left(1-\frac{a}{2}\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\left(1-\frac{a}{2}\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]$$
$$+\left(1-\frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(1-\frac{b}{2}\right)\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+\frac{20\gamma\sigma^2}{nB'}.$$

Using Lemma 4 and the assumption about γ, we get

$$\mathrm{E}\left[f(x^{t+1})\right]+\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]$$
$$+\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]$$
$$\leq\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$$
$$+\left(1-\frac{a}{2}\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\left(1-\frac{a}{2}\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]$$
$$+\left(1-\frac{b}{2}\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(1-\frac{b}{2}\right)\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+\frac{20\gamma\sigma^2}{nB'}.$$

Due to $\gamma\leq\frac{a}{2\mu}$ and $\gamma\leq\frac{b}{2\mu}$, we have

$$\mathrm{E}\left[f(x^{t+1})\right]+\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^{t+1}-h^{t+1}\right\|^2\right]+\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]$$
$$+\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]+\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]$$
$$\leq\mathrm{E}\left[f(x^t)\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$$
$$+\left(1-\gamma\mu\right)\frac{2\gamma(2\omega+1)}{p_{\mathrm{a}}}\mathrm{E}\left[\left\|g^t-h^t\right\|^2\right]+\left(1-\gamma\mu\right)\frac{2\gamma\left(\left(2\omega+1\right)p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t-h_i^t\right\|^2\right]$$
$$+\left(1-\gamma\mu\right)\frac{2\gamma}{b}\mathrm{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\left(1-\gamma\mu\right)\frac{8\gamma\left(p_{\mathrm{a}}-p_{\mathrm{aa}}\right)}{np_{\mathrm{a}}^2p_{\mathrm{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+\frac{20\gamma\sigma^2}{nB'}.$$

942 It is left to apply Lemma 11 with

$$\Psi^t = \frac{2(2\omega+1)}{p_{\text{a}}}\mathrm{E}\left[\left\|g^t - h^t\right\|^2\right] + \frac{2\left(\left(2\omega+1\right)p_{\text{a}} - p_{\text{aa}}\right)}{np_{\text{a}}^2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|g_i^t - h_i^t\right\|^2\right]$$
$$+ \frac{2}{b}\mathrm{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right] + \frac{8\left(p_{\text{a}} - p_{\text{aa}}\right)}{np_{\text{a}}^2 p_{\text{mega}}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|h_i^t - \nabla f_i(x^t)\right\|^2\right]$$

943 and $C = \frac{20\sigma^2}{nB'}$ to conclude the proof. □