Byzantine Resilient Non-Convex SVRG with Distributed Batch Gradient Computations

Prashant Khanduri PKHANDUR@SYR.EDU
Saikiran Bulusu SABULUSU@SYR.EDU
Pranay Sharma PSHARM04@SYR.EDU
Pramod K. Varshney VARSHNEY@SYR.EDU
Syracuse University, Syracuse, NY 13244

Abstract

In this work, we consider the distributed stochastic optimization problem of minimizing a non-convex function $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} f(x; \xi)$ in an adversarial setting, where the individual functions $f(x; \xi)$ can also be potentially non-convex. We assume that at most α-fraction of a total of K nodes can be Byzantines. We propose a robust stochastic variance-reduced gradient (SVRG) like algorithm for the problem, where the batch gradients are computed at the worker nodes (WNs) and the stochastic gradients are computed at the server node (SN). For the non-convex optimization problem, we show that we need $\tilde{O}\left(\frac{1}{\epsilon^{5/3} K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}}\right)$ gradient computations on average at each node (SN and WNs) to reach an ϵ-stationary point. The proposed algorithm guarantees convergence via the design of a novel Byzantine filtering rule which is independent of the problem dimension. Importantly, we capture the effect of the fraction of Byzantine nodes α present in the network on the convergence performance of the algorithm.

1. Introduction

In the current machine learning landscape, with the data sizes growing exponentially, distributed learning has become an important paradigm. In a distributed learning setup, multiple distributed nodes along with a server node perform the learning task at hand. In distributed settings, the computational load on the server node is relieved as heavy computations are usually distributed across the multiple nodes in the network, see Dekel et al. [3], Ho et al. [4], Recht et al. [9], Zinkevich et al. [17]. With multiple nodes in the network, the problem of robust learning becomes important as some of the nodes in the network can potentially be Byzantine. These Byzantine nodes can adversely affect the convergence performance of the algorithm. Therefore, it is important to design algorithms which are robust to Byzantine actions and at the same time provide sufficient convergence guarantees. In this work, we propose one such algorithm for distributed learning, when a maximum $\alpha \in [0, \frac{1}{2})$ fraction of nodes in the network are Byzantines.

Many works in the past have looked into the problem of learning in the presence of Byzantines, e.g., in Alistarh et al. [1], Blanchard et al. [2], Recht et al. [9], Su and Xu [11], Xie et al. [12, 13], Yin et al. [14, 15]. However, only a handful have considered the problems when the objective functions are non-convex, partially, because of the difficulty in dealing with the non-convex objective functions, see Blanchard et al. [2], Xie et al. [12, 13], Yin et al. [14, 15]. In this work, we deal with a non-convex optimization problem in the presence of Byzantines. To the best of our knowledge,

this is the first work which provides guarantees for an SVRG-like algorithm (Sections 2 and 3) for distributed learning in the presence of Byzantines with non-convex objective functions.

Related Work: In Su and Xu [11], a robust gradient descent (GD) based algorithm was proposed in the presence of Byzantine nodes for learning with strongly convex objective functions. Also, in Yin et al. [14, 15], the authors proposed Byzantine resilient GD algorithms for non-convex objective functions. Note that GD based algorithms require computations of the gradients of complete batches, therefore, in Alistarh et al. [1], Blanchard et al. [2], Li et al. [8], Xie et al. [12, 13], the authors proposed stochastic gradient descent (SGD) based robust learning algorithms in the presence of Byzantine nodes. In particular, Alistarh et al. [1], Li et al. [8], considered the objective functions to be strongly convex and convex, respectively. On the other hand, in Blanchard et al. [2] and Xie et al. [12, 13], the authors proposed robust algorithms in the presence of Byzantine nodes for non-convex objective functions.

In this work, we consider a non-convex learning problem in the presence of at most α-fraction Byzantine nodes. The proposed algorithm in this work is based on the stochastic variance reduced gradient (SVRG) and stochastically controlled stochastic gradient (SCSG) frameworks proposed in Reddi et al. [10] and Lei et al. [7], respectively, for non-convex objective functions. SVRG and SCSG reduce the variance of stochastic gradients in SGD by introducing inner and outer loops and by computing a batch gradient at the beginning of every outer loop. SCSG based algorithms originally proposed in Lei and Jordan [6] have been shown to improve the performance of SVRG by selecting the number of inner iterations in a random fashion, whereas SVRG chooses the inner loop size in a deterministic fashion. Our work, extends the framework of non-convex SCSG to distributed setting, even in the presence of Byzantine nodes. Below, we list our contributions:

Contributions:

- We propose a novel algorithm for distributed non-convex optimization in the presence of Byzantine nodes. The proposed algorithm is a variant of the SVRG algorithm (Section 3), where the computationally demanding batch gradients are computed at the worker nodes (WNs) and the stochastic gradients are computed at the server node (SN).

- We provide convergence guarantees for the proposed algorithm as a function of $\alpha \in [0, \frac{1}{2})$, which is the upper bound on the fraction of Byzantine nodes present in the network. Importantly, we show that as the number of Byzantines reduces to zero ($\alpha = 0$), the algorithm improves the best known convergence rates for distributed non-convex optimization (Section 4) Jiang and Agrawal [5], Yu et al. [16].

- We propose a novel Byzantine filtering rule which is independent of the problem dimension. Specifically, the aggregation rule proposed in this work does not perform coordinate-wise operations and thereby the convergence rates are independent of the problem dimension. Therefore, the proposed algorithm is suitable for high-dimensional learning problems.

2. Model and Assumptions

We consider a model similar to the one in Alistarh et al. [1]. We assume that there are K WNs and a SN in the network. Specifically, we want to solve the following problem in a distributed fashion:

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} f(x; \xi)$$

where the functions $f(\cdot\,;\xi) : \mathbb{R}^d \to \mathbb{R}$, for ξ chosen uniformly randomly from distribution $\mathcal{D}$, and f, can potentially be non-convex. All the nodes including SN have access to the stochastic functions from distribution $\mathcal{D}$. Of all the WNs, we assume that at most α-fraction are Byzantines with $\alpha \in [0, 1/2)$.

The set of honest nodes is denoted by $\mathcal{G}$. For each honest node $k \in \mathcal{G}$, the following assumptions are made.

Assumption 1 (Gradient Lipschitz continuity) *All the functions $f(\cdot\,;\xi)$ for any $\xi \sim \mathcal{D}$ and f are assumed to be L-smooth, i.e., we have $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2$, with $L > 0$.*

Assumption 2 (Bounded Variance) *For any $\xi \sim \mathcal{D}$ we have $\|\nabla f(x;\xi) - \nabla f(x)\| \leq \mathcal{V}$.*

Remark 1 *Assumption 2 is also required in Alistarh et al. [1] to design the Byzantine filtering strategy. Moreover, for $\alpha = 0$, the above assumption can be relaxed to $\mathbb{E}\|\nabla f(x;\xi) - \nabla f(x)\| \leq \mathcal{V}$, which is a standard assumption in stochastic non-convex optimization literature.*

For non-convex problems, it is generally not feasible to measure the suboptimality of the function value, therefore, usually the convergence of non-convex problems is measured in terms of expected gradient norm square, $\mathbb{E}\|f(x)\|^2$. Below we define an ϵ-stationary point for a non-convex optimization problem.

Definition 2.1 (ϵ-Stationary Point) *A point x is called ϵ-stationary if $\|\nabla f(x)\|^2 \leq \epsilon$. Moreover, a stochastic algorithm is said to achieve ϵ-stationarity in t iterations if $\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \epsilon$, where the expectation is over the stochasticity of the algorithm until time instant t.*

Next, we discuss the algorithm.

3. Algorithm

Now, we discuss the steps of Algorithm 1. As mentioned earlier, we consider a SVRG-like algorithm, which can be thought of as a distributed version of SCSG, where the WNs compute the batch gradients and share the computed batch gradients with the SN. Note that the algorithm is similar to SVRG except the fact that the algorithm uses a geometric random variable to decide the number of inner iterations.

The algorithm consists of a total of T epochs. At the start of each epoch $t = 1, 2, \ldots, T$, the SN broadcasts the point x_0^t to the WNs. The WNs are then expected to compute their batch gradients at x_0^t and forward them to the SN. However, a Byzantine node forwards an arbitrary vector to the SN. The honest WNs, on the other hand, compute and forward their batch gradients, $\mu_t^{(k)}$. The set $\mathcal{S}_t^{(k)}$ for $t = 1, 2, \ldots, T$ and $k \in [K]$ consists of B i.i.d. sample functions $f(x_0^t; \xi_{t,i}^{(k)})$ with $\xi_{t,i}^{(k)} \sim \mathcal{D}$ and $\mathcal{S}_t^{(k)} = (\xi_{t,i}^{(k)})_{i=1}^B$. Note that for simplicity, we assume the batch sizes $|\mathcal{S}_t^{(k)}| = B$, for all $t = 1, 2, \ldots, T$ and $k \in \mathcal{G}$. Furthermore, a generalization with variable batch sizes at different nodes is straightforward. To summarize, the vectors sent by node k at the tth epoch to the SN is:

$$\mu_t^{(k)} = \begin{cases} \frac{1}{B} \sum_{i=1}^B \nabla f(x_0^t; \xi_{t,i}^{(k)}) & \text{if } k \in \mathcal{G} \\ * & \text{if } k \notin \mathcal{G} \end{cases} \tag{1}$$

where $*$ indicates an arbitrary vector sent by the Byzantine node.

Algorithm 1 Byzantine SVRG with Distributed Batch Gradient Computations

---

Input: $\tilde{x}_0 \in \mathbb{R}^d$, step sizes $(\eta_t)_{t=1}^T$, batch size B, Variance Bound $\mathcal{V}$ (Assumption 2), $\mathfrak{T}_\mu = 2\mathcal{V}\sqrt{\frac{C}{B}}$

(Lemma 9) where $C = 2\log\left(\frac{2K}{\delta}\right)$ with $\delta \in (0,1)$ (Theorem 2).

for $t = 1,2,\ldots,T$ **do**

$\quad$ $x_0^t \leftarrow \tilde{x}_{t-1} \qquad \rightarrow$ **Push to WNs;**

$\quad$ **for** $k = 1,2,\ldots,K$ **do**

$\quad\quad$ $\mu_t^{(k)} = \begin{cases} \frac{1}{B}\sum_{i=1}^B \nabla f(x_0^t; \xi_{t,i}^{(k)}) & \text{for } k \in \mathcal{G} \\ * & \text{for } k \notin \mathcal{G} \end{cases} \qquad \rightarrow$ **Push to SN**

$\quad$ **end**

$\quad$ $\mu_t^{\text{med}} \leftarrow \mu_t^{(k)}$ where $k \in [K]$ is any WN such that $|\{k' \in [K] : \|\mu_t^{(k')} - \mu_t^{(k)}\| \leq \mathfrak{T}_\mu\}| > K/2$;

$\quad$ $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 2\mathfrak{T}_\mu\}$;

$\quad$ **if** $|\mathcal{G}_t| < (1-\alpha)K$ **then**

$\quad\quad$ $\mu_t^{\text{med}} \leftarrow \mu_t^{(k)}$ where $k \in [K]$ is any WN s.t. $|\{k' \in [K] : \|\mu_t^{(k')} - \mu_t^{(k)}\| \leq 2\mathcal{V}\}| > K/2$;

$\quad\quad$ $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 4\mathcal{V}\}$;

$\quad$ **end**

$\quad$ $\mu_t = \frac{1}{|\mathcal{G}_t|}\sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$;

$\quad$ **for** $n = 1,2,\ldots,N_t$, $N_t \sim Geom(\frac{B}{B+1})$ **do**

$\quad\quad$ $v_{n-1}^t = \nabla f(x_{n-1}^t; \xi_{n-1}^t) - \nabla f(x_0^t; \xi_{n-1}^t) + \mu_t$;

$\quad\quad$ $x_n^t = x_{n-1}^t - \eta_t v_{n-1}^t$;

$\quad$ **end**

$\quad$ $\tilde{x}_t \leftarrow x_{N_t}^t$;

end

Output: $\tilde{x}_a$ chosen uniformly randomly from $(\tilde{x}_t)_{t=1}^T$.

After receiving the batch gradients from the WNs, the SN performs a Byzantine Filtering Step. In this step, the SN computes its estimate of the good set $\mathcal{G}_t$. The server node then aggregates the gradients received from the WNs in the set $\mathcal{G}_t$ and forms an estimate of the batch gradient, μ_t. This is followed by the SVRG-like inner loop ($n = 1, 2, \ldots, N_t$), where the inner loop size N_t is chosen randomly using a geometric random variable with parameter $\frac{B}{B+1}$, i.e., $N_t \sim \text{Geom}\left(\frac{B}{B+1}\right)$. The SN then performs the SVRG update step using the estimated batch gradient, μ_t, and stochastic gradients computed at x_{n-1}^t and x_0^t. Next, we discuss the proposed Byzantine filtering strategy.

Byzantine Filtering Step: To design the Byzantine filtering rule, the SN uses the computed batch gradients from all the worker nodes, $(\mu_t^{(k)})_{k \in [K]}$. Note that some of the received batch gradients may be arbitrary sent by the Byzantines (see (1)). Below we define the Byzantine filtering rule and then motivate its construction.

First, the SN computes the vector median of the received vectors $(\mu_t^{(k)})_{k \in [K]}$ using $\mathfrak{T}_\mu$ as defined in Algorithm 1. Then the SN filters the nodes it *believes* to be Byzantines and constructs $\mathcal{G}_t$ using the rule: $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 2\mathfrak{T}_\mu\}$. Note that the above rule is motivated by the fact that for good nodes $k \in \mathcal{G}$, we expect with high probability the batch gradients to be concentrated around the true gradient, $\nabla f(\cdot)$. However, with non-zero probability, the above set

can be empty or can have $|\mathcal{G}_t| < (1 - \alpha)K$. As we know that $|\mathcal{G}| \geq (1 - \alpha)K$, we would want to avoid such scenarios. For this purpose, if we have $|\mathcal{G}_t| < (1 - \alpha)K$ using the first rule (see Algorithm 1), we update the definition of the median using $2\mathcal{V}$ to define the new median and construct: $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 4\mathcal{V}\}$. This condition ensures that we always have $|\mathcal{G}_t| \geq (1 - \alpha)K$. Note that the Byzantine filtering rule is similar to the one used in Alistarh et al. [1]. However, in contrast to Alistarh et al. [1], we do not maintain a running sum of any statistic to filter Byzantines. Instead, we control the impact of Byzantine nodes by selecting a batch size appropriately. Importantly, Alistarh et al. [1] considered only convex objectives. Next, we provide the guarantees for the algorithm.

4. Convergence Guarantees

Let us denote by $\mathbb{E}G_{\text{comp, SN}}(\epsilon)$ the expected number of total gradient computations required at the SN (same number of computations are required at individual WNs) to reach an ϵ-stationary point. By Algorithm 1 we have:

$$\mathbb{E}G_{\text{comp, SN}}(\epsilon) = \sum_{t=1}^{T}(B + \mathbb{E}[N_t]) = 2TB \ \text{ with } \ N_t \sim \text{Geom}\left(\frac{B}{B+1}\right).$$

where T is the number of iterations required to reach an ϵ-stationary point (see Output of Algorithm 1). Next, we state the convergence result.

Theorem 2 *If Assumptions 1 and 2 are satisfied, for step size $\eta_t = \eta = \frac{1}{3LB^{2/3}}$ and $B \geq 16$ such that we have $\delta \in (0, 1)$ such that: $\mathrm{e}^{\frac{\delta B}{2(1-2\delta)}} \leq \frac{2K}{\delta} \leq \mathrm{e}^{\frac{B}{2}}$ and $\delta \leq \frac{1}{25KB}$ then we have:*

$$\mathbb{E}\|\nabla f(\tilde{x}_a)\|^2 \leq \underbrace{\frac{12L\mathbb{E}[f(\tilde{x}_0) - f(\tilde{x}^*)]}{TB^{1/3}}}_{T=O\left(\frac{1}{\epsilon B^{1/3}}\right)} + \underbrace{\frac{32\mathcal{V}^2}{(1-\alpha)^2KB}}_{B=O\left(\frac{1}{\epsilon K}\right)} + \underbrace{\frac{2176\alpha^2\mathcal{V}^2C}{(1-\alpha)^2B}}_{B=O\left(\frac{\alpha^2}{\epsilon}\right)}.$$

Using the above, $\mathbb{E}G_{\text{comp, SN}}(\epsilon)$ can be computed as:

Corollary 3 *Under the assumptions stated in Theorem 2:*
(i): We have: $\mathbb{E}G_{comp, SN}(\epsilon) \leq \tilde{O}\left(\frac{1}{\epsilon^{5/3}K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}}\right)$ where $\tilde{O}(\cdot)$ hides the $\log$ factors.
(ii) Moreover, when $\alpha = 0$ we have: $\mathbb{E}G_{comp, SN}(\epsilon) \leq O\left(\frac{1}{\epsilon^{5/3}K^{2/3}}\right)$.

Note that for the case when $\alpha = 0$ and $k \leq 1/\epsilon$, our algorithm improves upon the best known convergence rates of $O\left(\frac{1}{\epsilon^2 K}\right)$ for distributed stochastic non-convex optimization given in Jiang and Agrawal [5], Yu et al. [16]. Moreover, if Algorithm 1 is run only at the SN, we achieve $\mathbb{E}G_{\text{comp, SN}}(\epsilon) \leq O\left(\frac{1}{\epsilon^{5/3}}\right)$ which is the same as is computed in Lei et al. [7].

5. Conclusion

In this work, we proposed the first non-convex SVRG like algorithm which considers distributed optimization in the presence of Byzantine nodes. We proposed a novel aggregation rule which is independent of the problem dimension, d. Importantly, we captured the effect of Byzantine nodes in the network and showed that the proposed algorithm outperforms the best known convergence rates known in the literature in the presence of Byzantine nodes.

References

[1] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems 31*, pages 4613–4623. Curran Associates, Inc., 2018.

[2] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems 30*, pages 119–129. Curran Associates, Inc., 2017.

[3] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[4] Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in neural information processing systems*, pages 1223–1231, 2013.

[5] Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Advances in Neural Information Processing Systems*, pages 2525–2536, 2018.

[6] Lihua Lei and Michael Jordan. Less than a Single Pass: Stochastically Controlled Stochastic Gradient. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 148–156, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[7] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

[8] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[9] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.

[10] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 314–323, New York, New York, USA, 20–22 Jun 2016. PMLR.

[11] Lili Su and Jiaming Xu. Securing distributed machine learning in high dimensions. *arXiv preprint arXiv:1804.10140*, 2018.

[12] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Phocas: dimensional byzantine-resilient stochastic gradient descent. *arXiv preprint arXiv:1805.09682*, 2018.

[13] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. *arXiv preprint arXiv:1805.10032*, 2018.

[14] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5650–5659, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[15] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Defending against saddle point attack in Byzantine-robust distributed learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7074–7084, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[16] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193, 2019.

[17] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Appendix A.

For the purpose of the proof, we consider a more general model where the nodes can choose different batch sizes, B_t, for $t = 1, 2, \ldots, T$, across epochs. For varying batch sizes in Algorithm 1 the Byzantine filtering constant $\mathfrak{T}_\mu$ will be evaluated inside the epochs and will be a function of B_t. Rest of the algorithm will stay the same. Our proof follows the structure similar to the one in Lei et al. [7] with a few major differences. The problem considered in Lei et al. [7] was a finite sum problem with all the gradient computations designated at the central node. The key novelty in our proof lies in proving the boundedness of the norm square of the error term e_t (see Lemma 7).

Proof of Theorem 2: Using the smoothness of function f we have:

$$
\begin{aligned}
\mathbb{E}_{\xi_n^t} f(x_{n+1}^t) &\leq f(x_n^t) - \eta_t \langle \mathbb{E}_{\xi_n^t} v_n^t, \nabla f(x_n^t) \rangle + \frac{L\eta_t^2}{2} \mathbb{E}_{\xi_n^t} \|v_n^t\|^2 \\
&\overset{(a)}{\leq} f(x_n^t) - \eta_t(1 - L\eta_t)\|\nabla f(x_n^t)\|^2 \\
&\quad - \eta_t \langle e_t, \nabla f(x_n^t) \rangle + \frac{L^3\eta_t^2}{2}\|x_n^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2
\end{aligned}
\tag{2}
$$

where (a) follows from Lemma 4. Denoting by $\mathbb{E}_t$ the expectation w.r.t. all $\xi_1^t, \xi_2^t, \ldots$ given N_t. Since $\xi_1^t, \xi_2^t, \ldots$ are independent of N_t, $\mathbb{E}_t$ is equivalent to expectation w.r.t. $\xi_1^t, \xi_2^t, \ldots$. We have from above

$$
\begin{aligned}
\mathbb{E}_t f(x_{n+1}^t) &\leq \mathbb{E}_t f(x_n^t) - \eta_t(1 - L\eta_t)\mathbb{E}_t\|\nabla f(x_n^t)\|^2 \\
&\quad - \eta_t \mathbb{E}_t \langle e_t, \nabla f(x_n^t) \rangle + \frac{L^3\eta_t^2}{2}\mathbb{E}_t\|x_n^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2
\end{aligned}
$$

Now taking $n = N_t$ and denoting by $\mathbb{E}_{N_t}$ expectation w.r.t. N_t we get:

$$
\begin{aligned}
\mathbb{E}_{N_t}\mathbb{E}_t f(x_{N_t+1}^t) &\leq \mathbb{E}_{N_t}\mathbb{E}_t f(x_{N_t}^t) - \eta_t(1 - L\eta_t)\mathbb{E}_{N_t}\mathbb{E}_t\|\nabla f(x_{N_t}^t)\|^2 \\
&\quad - \eta_t\mathbb{E}_{N_t}\mathbb{E}_t \langle e_t, \nabla f(x_{N_t}^t) \rangle + \frac{L^3\eta_t^2}{2}\mathbb{E}_{N_t}\mathbb{E}_t\|x_{N_t}^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2
\end{aligned}
$$

Using Fubini's theorem and rearranging the terms we have

$$
\begin{aligned}
\eta_t(1 - L\eta_t)\mathbb{E}_{N_t}\mathbb{E}_t\|\nabla f(x_{N_t}^t)\|^2 &\leq \mathbb{E}_{N_t}\mathbb{E}_t f(x_{N_t}^t) - \mathbb{E}_{N_t}\mathbb{E}_t f(x_{N_t+1}^t) \\
&\quad - \eta_t\mathbb{E}_{N_t}\mathbb{E}_t \langle e_t, \nabla f(x_{N_t}^t) \rangle + \frac{L^3\eta_t^2}{2}\mathbb{E}_{N_t}\mathbb{E}_t\|x_{N_t}^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2 \\
&\overset{(a)}{=} \frac{1}{B_t}[f(x_0^t) - \mathbb{E}_t\mathbb{E}_{N_t}f(x_{N_t}^t)] \\
&\quad - \eta_t\mathbb{E}_{N_t}\mathbb{E}_t \langle e_t, \nabla f(x_{N_t}^t) \rangle + \frac{L^3\eta_t^2}{2}\mathbb{E}_{N_t}\mathbb{E}_t\|x_{N_t}^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2
\end{aligned}
$$

where (a) follows from Lemma 11, Lemma 12 and Fubini's theorem. Now taking expectation over all the randomness and using the fact that $x_{N_t}^t = \tilde{x}_t$ and $\tilde{x}_{t-1} = x_0^t$ we get

$$\eta_t(1 - L\eta_t)\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 \leq \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] - \eta_t\mathbb{E}\langle e_t, \nabla f(\tilde{x}_t)\rangle$$

$$+ \frac{L^3\eta_t^2}{2}\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2 + L\eta_t^2\mathbb{E}\|e_t\|^2$$

$$\stackrel{(a)}{=} \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + \frac{1}{B_t}\mathbb{E}\langle e_t, \tilde{x}_t - \tilde{x}_{t-1}\rangle$$

$$+ \frac{L^3\eta_t^2}{2}\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2 + \eta_t(1 + L\eta_t)\mathbb{E}\|e_t\|^2$$

$$\stackrel{(b)}{\leq} \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + \left(\frac{1}{2\eta_t B_t}\left(-\frac{1}{B_t} + \eta_t^2 L^2\right)\right)\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2$$

$$- \frac{1}{B_t}\mathbb{E}\langle\nabla f(\tilde{x}_t), \tilde{x}_t - \tilde{x}_{t-1}\rangle + \frac{\eta_t}{B_t}\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 + \frac{\eta_t}{B_t}\mathbb{E}\|e_t\|^2$$

$$+ \frac{L^3\eta_t^2}{2}\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2 + \eta_t(1 + L\eta_t)\mathbb{E}\|e_t\|^2$$

where (a) follows from Lemma 5, (b) follows from Lemma 6. Rearranging the terms to get:

$$\eta_t\left(1 - L\eta_t - \frac{1}{B_t}\right)\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 + \left(\frac{1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2}{2\eta_t B_t^2}\right)\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2$$

$$\leq \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + \frac{1}{B_t}\mathbb{E}\langle\nabla f(\tilde{x}_t), \tilde{x}_{t-1} - \tilde{x}_t\rangle + \eta_t\left(1 + L\eta_t + \frac{1}{B_t}\right)\mathbb{E}\|e_t\|^2.$$

Now using the Young's inequality $\mathbb{E}\langle a, b\rangle \leq \mathbb{E}\left[\frac{\beta}{2}\|a\|^2 + \frac{1}{2\beta}\|b\|^2\right]$ on $\mathbb{E}\langle\nabla f(\tilde{x}_t), \tilde{x}_{t-1} - \tilde{x}_t\rangle$ with

$$\beta = \frac{1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2}{\eta_t B_t}, \quad a = \tilde{x}_{t-1} - \tilde{x}_t \quad \text{and} \quad b = \nabla f(\tilde{x}_t).$$

we get

$$\eta_t\left(1 - L\eta_t - \frac{1}{B_t} - \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)}\right)\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2$$

$$\leq \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + \eta_t\left(1 + L\eta_t + \frac{1}{B_t}\right)\mathbb{E}\|e_t\|^2$$

$$\eta_t\left(1 - L\eta_t - \frac{1}{B_t} - \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)}\right)\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2$$

$$\stackrel{(a)}{\leq} \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + \eta_t\left(1 + L\eta_t + \frac{1}{B_t}\right)\left(\frac{4\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2\mathcal{V}^2 C}{(1-\alpha)^2 B_t}\right).$$

$$\tag{3}$$

where (a) follows from Lemma 7.

Choosing η_t such that we have $1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2 > 0$, this implies that we have

$$\frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)} > \frac{1}{2}.$$

Further we choose η_t such that we have:

$$1 - L\eta_t - \frac{1}{B_t} - \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)} \geq \frac{1}{4}$$

$$L\eta_t + \frac{1}{B_t} + \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)} \leq \frac{3}{4}.$$

Choosing η_t such that we can ensure:

$$(i): \frac{1}{2} < \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)} \leq \frac{5}{8}, \quad (ii): L\eta_t \leq \frac{1}{16} \quad \text{and} \quad (iii): \frac{1}{B_t} \leq \frac{1}{16}.$$

Condition *(i)* above implies:

$$\eta_t^2 L^2 B_t + \eta_t^3 L^3 B_t^2 \leq \frac{1}{5}.$$

Further ensuring η_t such that

$$\eta_t^2 L^2 B_t \leq \frac{1}{10} \quad \text{and} \quad \eta_t^3 L^3 B_t^2 \leq \frac{1}{10}.$$

This implies that

$$\eta_t \leq \frac{1}{10^{1/2} L B_t^{1/2}} \quad \text{and} \quad \eta_t \leq \frac{1}{10^{1/3} L B_t^{2/3}}.$$

Furthermore from conditions *(ii)* and *(iii)* above we get:

$$\eta_t \leq \frac{1}{16L} \quad \text{and} \quad B_t \geq 16.$$

The above discussion implies that we must have $B_t \geq 16$ and we can choose $\eta_t \leq \frac{1}{3LB_t^{2/3}}$ as we have

$$\frac{1}{3LB_t^{2/3}} \leq \min\left\{\frac{1}{16L}, \frac{1}{10^{1/3} L B_t^{2/3}}, \frac{1}{10^{1/2} L B_t^{1/2}}\right\} \quad \text{for } B_t \geq 16.$$

This choice of η_t ensures that the term:

$$1 - L\eta_t - \frac{1}{B_t} - \frac{1}{2(1 - \eta_t^2 L^2 B_t - \eta_t^3 L^3 B_t^2)} \geq \frac{1}{4}. \tag{4}$$

Now replacing η_t and B_t in the term:

$$1 + \eta_t L + \frac{1}{B_t} \leq 1 + \frac{1}{3B_t^{2/3}} + \frac{1}{B_t} \leq 2. \tag{5}$$

Now replacing (4) and (5) in (3), and replacing $\eta_t = \frac{1}{3LB_t^{2/3}}$ we get

$$\frac{\eta_t}{4}\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 \leq \frac{1}{B_t}\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)] + 2\eta_t\left(\frac{4\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t}\right)$$

$$\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 \leq \frac{12L\mathbb{E}[f(\tilde{x}_{t-1}) - f(\tilde{x}_t)]}{B_t^{1/3}} + \frac{32\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{2176\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t}.$$

Constant batch size: For $B_t = B$ and summing over $t = 1, 2, \ldots, T$ and choosing x_a using Algorithm 1 we get:

$$\mathbb{E}\|\nabla f(\tilde{x}_a)\|^2$$
$$\leq \frac{12L\mathbb{E}[f(\tilde{x}_0) - f(\tilde{x}^*)] + 32\mathcal{V}^2(1-\alpha)^{-2}K^{-1}\sum_{t=1}^{T} B_t^{-2/3} + 2176\alpha^2\mathcal{V}^2C(1-\alpha)^{-2}\sum_{t=1}^{T} B_t^{-2/3}}{\sum_{t=1}^{T} B_t^{1/3}}.$$

Replace $B_t = B$ we get:

$$\mathbb{E}\|\nabla f(\tilde{x}_a)\|^2 \leq \underbrace{\frac{12L\mathbb{E}[f(\tilde{x}_0) - f(\tilde{x}^*)]}{TB^{1/3}}}_{T=O\left(\frac{1}{\epsilon B^{1/3}}\right)} + \underbrace{\frac{32\mathcal{V}^2}{(1-\alpha)^2KB}}_{B=O\left(\frac{1}{\epsilon K}\right)} + \underbrace{\frac{2176\alpha^2\mathcal{V}^2C}{(1-\alpha)^2B}}_{B=O\left(\frac{\alpha^2}{\epsilon}\right)}.$$

Now to guarantee that we get an ϵ-accurate solution we need: $T = O\left(\frac{1}{\epsilon B^{1/3}}\right)$ number of iterations for the first term. For the second term, we need $B_K = O\left(\frac{1}{\epsilon K}\right)$ batch size and for the third term, we need $B_\alpha = O\left(\frac{\alpha^2}{\epsilon}\right)$ batch size to account for the Byzantine workers. This implies that the total number of gradient computations, $\mathbb{E}G_{\text{comp, SN}}$ required at the SN (and at the individual WNs) on an average are of the order of:

$$\mathbb{E}G_{\text{comp, SN}}(\epsilon) \leq TB_K + TB_\alpha$$
$$\mathbb{E}G_{\text{comp, SN}}(\epsilon) \leq O\left(\frac{1}{\epsilon^{5/3}K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}}\right).$$

And the expected number of gradient computations across the network denoted by, $\mathbb{E}G_{\text{comp, NW}}$, are of the order of:

$$\mathbb{E}G_{\text{comp, NW}}(\epsilon) \leq O\left(\frac{K^{1/3}}{\epsilon^{5/3}} + \frac{K\alpha^{4/3}}{\epsilon^{5/3}}\right).$$

Moreover, note that if we have $\alpha = 0$ we get the expected computational complexity and the expected number of gradient computations across the network are of the order of:

$$\mathbb{E}G_{\text{comp, SN}}(\epsilon) \leq O\left(\frac{1}{\epsilon^{5/3}K^{2/3}}\right) \quad \text{and} \quad \mathbb{E}G_{\text{comp, NW}}(\epsilon) \leq O\left(\frac{K^{1/3}}{\epsilon^{5/3}}\right).$$

∎

Lemma 4 *We have*

$$\mathbb{E}_{\xi_n^t}\|v_n^t\|^2 \leq L^2\|x_n^t - x_0^t\|^2 + 2\|\nabla f(x_n^t)\|^2 + 2\|e_t\|^2.$$

Proof From the definition of v_n^t we have:

$$v_n^t = \nabla f(x_n^t; \xi_n^t) - \nabla f(x_0^t; \xi_n^t) + \mu_t.$$

where $\mu_t = \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$. We define

$$e_t = \mu_t - \nabla f(x_0^t) = \mu_t - \nabla f(\tilde{x}_{t-1}).$$

This implies that we have

$$\mathbb{E}_{\xi_n^t} v_n^t = \nabla f(x_n^t) + e_t.$$

Now taking $\mathbb{E}_{\xi_n^t} \|v_n^t\|^2$ and using $\mathbb{E}\|Z\|^2 = \mathbb{E}\|Z - \mathbb{E}Z\|^2 + \|\mathbb{E}Z\|^2$, we have

$$
\begin{aligned}
\mathbb{E}_{\xi_n^t} \|v_n^t\|^2 &= \mathbb{E}_{\xi_n^t} \|v_n^t - \mathbb{E}_{\xi_n^t} v_n^t\|^2 + \|\mathbb{E}_{\xi_n^t} v_n^t\|^2 \\
&= \mathbb{E}_{\xi_n^t} \|\nabla f(x_n^t; \xi_n^t) - \nabla f(x_0^t; \xi_n^t) - (\nabla f(x_n^t) - \nabla f(x_0^t))\|^2 + \|\nabla f(x_n^t) + e_t\|^2 \\
&\overset{(a)}{\leq} \mathbb{E}_{\xi_n^t} \|\nabla f(x_n^t; \xi_n^t) - \nabla f(x_0^t; \xi_n^t) - (\nabla f(x_n^t) - \nabla f(x_0^t))\|^2 + 2\|\nabla f(x_n^t)\|^2 + 2\|e_t\|^2 \\
&\overset{(b)}{\leq} \mathbb{E}_{\xi_n^t} \|\nabla f(x_n^t; \xi_n^t) - \nabla f(x_0^t; \xi_n^t)\|^2 + 2\|\nabla f(x_n^t)\|^2 + 2\|e_t\|^2 \\
&\overset{(c)}{\leq} L^2 \|x_n^t - x_0^t\|^2 + 2\|\nabla f(x_n^t)\|^2 + 2\|e_t\|^2.
\end{aligned}
$$

where (a) follows from the definition of v_n^t and Lemma 14, (b) follows from variance inequality and (c) follows from the Gradient Lipschitz continuity of $f(\cdot; \xi_n^t)$. ∎

Lemma 5 *We have*

$$\eta_t \mathbb{E}\langle e_t, \nabla f(\tilde{x}_t)\rangle = \frac{1}{B_t} \mathbb{E}\langle e_t, \tilde{x}_{t-1} - \tilde{x}_t\rangle - \eta_t \mathbb{E}\|e_t\|^2.$$

Proof Consider the term $M_n^t = \langle e_t, x_n^t - x_0^t\rangle$. From the definition of M_n^t we have

$$M_{n+1}^t - M_n^t = \langle e_t, x_{n+1}^t - x_n^t\rangle = -\eta_t \langle e_t, v_n^t\rangle.$$

Taking expectation w.r.t. ξ_n^t, we have

$$
\begin{aligned}
\mathbb{E}_{\xi_n^t}(M_{n+1}^t - M_n^t) &= -\eta_t \langle e_t, \mathbb{E}_{\xi_n^t} v_n^t\rangle \\
&\overset{(a)}{=} -\eta_t \langle e_t, \nabla f(x_n^t)\rangle - \eta_t \|e_t\|^2,
\end{aligned}
$$

where (a) follows from the definition of v_n^t. Denoting by $\mathbb{E}_t$ the expectation w.r.t. all $\xi_1^t, \xi_2^t, \ldots$ given N_t. Since $\xi_1^t, \xi_2^t, \ldots$ are independent of N_t, $\mathbb{E}_t$ is equivalent to expectation w.r.t. $\xi_1^t, \xi_2^t, \ldots$. We have

$$\mathbb{E}_t(M_{n+1}^t - M_n^t) = -\eta_t \langle e_t, \mathbb{E}_t \nabla f(x_n^t)\rangle - \eta_t \|e_t\|^2.$$

Taking $n = N_t$ and expectation w.r.t. N_t as $\mathbb{E}_{N_t}$ we have

$$\mathbb{E}_{N_t}\mathbb{E}_t(M_{N_t+1}^t - M_{N_t}^t) = -\eta_t \langle e_t, \mathbb{E}_{N_t}\mathbb{E}_t \nabla f(x_{N_t}^t)\rangle - \eta_t \|e_t\|^2.$$

Using Fubini's theorem, Lemma 11, Lemma 12 and using the fact $x_{N_t}^t = \tilde{x}_t$ and $\tilde{x}_{t-1} = x_0^t$, we have

$$\frac{1}{B_t}\mathbb{E}_{N_t}\mathbb{E}_t\langle e_t, \tilde{x}_t - \tilde{x}_{t-1}\rangle = -\eta_t \langle e_t, \mathbb{E}_{N_t}\mathbb{E}_t \nabla f(\tilde{x}_t)\rangle - \eta_t \|e_t\|^2.$$

Taking expectation w.r.t. the whole past yields the statement of the lemma. ∎

Lemma 6 *We have*

$$
2\eta_t \mathbb{E}\langle e_t, \tilde{x}_t - \tilde{x}_{t-1} \rangle \leq \left(-\frac{1}{B_t} + \eta_t^2 L^2 \right) \mathbb{E} \|\tilde{x}_t - \tilde{x}_{t-1}\|^2
$$
$$
- 2\eta_t \mathbb{E}\langle \nabla f(\tilde{x}_t), \tilde{x}_t - \tilde{x}_{t-1} \rangle + 2\eta_t^2 \mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 + 2\eta_t^2 \mathbb{E}\|e_t\|^2.
$$

Proof We have from the update equation $x_{n+1}^t = x_n^t - \eta_t v_n^t$, we have

$$
\mathbb{E}_{\xi_n^t} \|x_{n+1}^t - x_0^t\|^2 = \mathbb{E}_{\xi_n^t} \|x_n^t - \eta_t v_n^t - x_0^t\|^2
$$
$$
= \|x_n^t - x_0^t\|^2 + \eta_t^2 \mathbb{E}_{\xi_n^t} \|v_n^t\|^2 - 2\eta_t \langle \mathbb{E}_{\xi_n^t} v_n^t, x_n^t - x_0^t \rangle
$$
$$
\overset{(a)}{\leq} (1 + \eta_t^2 L^2)\|x_n^t - x_0^t\|^2 - 2\eta_t \langle \nabla f(x_n^t), x_n^t - x_0^t \rangle
$$
$$
- 2\eta_t \langle e_t, x_n^t - x_0^t \rangle + 2\eta_t^2 \|\nabla f(x_n^t)\|^2 + 2\eta_t^2 \|e_t\|^2 \qquad (6)
$$

where (a) follows from Lemma 4 and the definition of v_n^t. Denoting by $\mathbb{E}_t$ the expectation w.r.t. all $\xi_1^t, \xi_2^t, \dots$ given N_t. Since $\xi_1^t, \xi_2^t, \dots$ are independent of N_t, $\mathbb{E}_t$ is equivalent to expectation w.r.t. $\xi_1^t, \xi_2^t, \dots$. We have

$$
\mathbb{E}_t \|x_{n+1}^t - x_0^t\|^2 \leq (1 + \eta_t^2 L^2)\mathbb{E}_t \|x_n^t - x_0^t\|^2 - 2\eta_t \mathbb{E}_t \langle \nabla f(x_n^t), x_n^t - x_0^t \rangle
$$
$$
- 2\eta_t \mathbb{E}_t \langle e_t, x_n^t - x_0^t \rangle + 2\eta_t^2 \mathbb{E}_t \|\nabla f(x_n^t)\|^2 + 2\eta_t^2 \|e_t\|^2.
$$

Now taking $n = N_t$ and taking expectation $\mathbb{E}_{N_t}$ w.r.t. N_t we have

$$
2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \tilde{x}_t - \tilde{x}_{t-1} \rangle \leq (1 + \eta_t^2 L^2)\mathbb{E}_{N_t} \mathbb{E}_t \|x_{N_t}^t - x_0^t\|^2 - \mathbb{E}_{N_t} \mathbb{E}_t \|x_{N_t+1}^t - x_0^t\|^2
$$
$$
- 2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle \nabla f(\tilde{x}_t), \tilde{x}_t - \tilde{x}_{t-1} \rangle + 2\eta_t^2 \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla f(\tilde{x}_t)\|^2 + 2\eta_t^2 \|e_t\|^2
$$
$$
\overset{(a)}{=} \left(-\frac{1}{B_t} + \eta_t^2 L^2 \right) \mathbb{E}_{N_t} \mathbb{E}_t \|\tilde{x}_t - \tilde{x}_{t-1}\|^2
$$
$$
- 2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle \nabla f(\tilde{x}_t), \tilde{x}_t - \tilde{x}_{t-1} \rangle + 2\eta_t^2 \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla f(\tilde{x}_t)\|^2 + 2\eta_t^2 \|e_t\|^2.
$$

where (a) follows from Lemma 11, Lemma 12 and Fubini's theorem. Finally, rearranging the terms and taking expectation w.r.t. the whole past yields the lemma. ∎

Lemma 7 *Choosing δ and B_t in Algorithm 1 such that the following are satisfied:*

(i) $e^{\frac{\delta B_t}{2(1-2\delta)}} \leq \frac{2K}{\delta} \leq e^{\frac{B_t}{2}}.$

(ii) $\delta \leq \frac{1}{25KB_t}.$

then $\mathbb{E}\|e_t\|^2$ is bounded as:

$$
\mathbb{E}\|e_t\|^2 \leq \frac{4\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t}.
$$

Proof From the definition $e_t = \mu_t - \nabla f(\tilde{x}_{t-1})$, where $\mu_t = \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$. Therefore, we have

$$
\mathbb{E}\|e_t\|^2 = \mathbb{E}\left\| \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right) \right\|^2
$$

Now let us define three types of events and their complements:

Definition 8 *We define three events and their complements which will be used to bound $\mathbb{E}\|e_t\|^2$:*

1. **Event A (Event $A!$):** *We denote the event of Lemma 9 as Event A, we define it again here for convenience. For all good nodes $k \in \mathcal{G}$ we have:*

   (a) $\|\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\| \leq \mathcal{V}\sqrt{\frac{C}{B_t}}$.

   (b) $\|\mu_t^{(k)} - \mu_t^{med}\| \leq 4\mathcal{V}\sqrt{\frac{C}{B_t}}$ *and* $\|\mu_t^{med} - \nabla f(\tilde{x}_{t-1})\| \leq 3\mathcal{V}\sqrt{\frac{C}{B_t}}$.

   Note that we have $\mathbb{P}[\text{Event } A] \geq 1 - \delta$. The complement of Event A is denoted as Event A! and we have $\mathbb{P}[\text{Event } A!] \leq \delta$.

2. **Event $\bar{A}$ (Event $\bar{A}!$):** *We define Event $\bar{A}$ as $\mathcal{G} \subset \mathcal{G}_t$. Consequently, Event $\bar{A}!$ is defined as $\mathcal{G} \not\subset \mathcal{G}_t$.*

3. **Event $R1$ (Event $R2$):** *We define Event R1 as the event that Rule 1 is executed. And Event R2 is the complement of Rule R1 and indicates that Rule 2 is executed. Rule 1 and Rule 2 are defined below for convenience.*

Rule 1: $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 2\mathfrak{T}_\mu\}$; for median evaluated as:
$\mu_t^{\text{med}} \leftarrow \mu_t^{(k)}$ where $k \in [K]$ is any WN such that $|\{k' \in [K] : \|\mu_t^{(k')} - \mu_t^{(k)}\| \leq \mathfrak{T}_\mu\}| > K/2$;

In case $|\mathcal{G}_t| < (1 - \alpha)K$ we use:

Rule 2: $\mathcal{G}_t = \{k \in [K] : \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 4\mathcal{V}\}$ for median evaluated as:
$\mu_t^{\text{med}} \leftarrow \mu_t^{(k)}$ where $k \in [K]$ is any WN s.t. $|\{k' \in [K] : \|\mu_t^{(k')} - \mu_t^{(k)}\| \leq 2\mathcal{V}\}| > K/2$;

Relationship between events:

- Note from Lemma 9 and from Definition 8 of Event A we have: Event $A \subset$ Event $\bar{A}$. This and Lemma 9 imply that we have:

$$\mathbb{P}[\text{Event } \bar{A}] \geq \mathbb{P}[\text{Event } A] \geq 1 - \delta. \tag{7}$$

- Consider Event $\bar{A}!$ as given in Definition 8. From above we have Event $A! \supset$ Event $\bar{A}!$. This along with Lemma 9 imply that we have:

$$\mathbb{P}[\text{Event } \bar{A}!] \leq \mathbb{P}[\text{Event } A!] \leq \delta. \tag{8}$$

- From the definition of Event $R1$ and Event A in Definition 8 we have: Event $A \subset$ Event $R1$. This further implies from Lemma 9 that we have:

$$\mathbb{P}[\text{Event } R1] \geq \mathbb{P}[\text{Event } A] \geq 1 - \delta \tag{9}$$

- Event $R2$ is complement of Event $R1$. This and the above implies Event $R2 \subset$ Event $A!$. Thus we have from Lemma 9:

$$\mathbb{P}[\text{Event } R2] \leq \mathbb{P}[\text{Event } A!] \leq \delta \tag{10}$$

Now we can write $\mathbb{E}\|e_t\|^2$ as:

$$\mathbb{E}\|e_t\|^2 \overset{(a)}{=} \mathbb{P}[\text{Event } \bar{A}] \, \mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}\right] + \mathbb{P}[\text{Event } \bar{A}!] \, \mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}!\right]$$

$$\overset{(b)}{\leq} \mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}\right] + \delta \, \mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}!\right] \tag{11}$$

where (a) follows from the law of total expectation and (b) follows from (7) and (8) above. Now let us first consider the first term $\mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}\right]$ under Event $\bar{A}$:

$$\mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}\right] = \mathbb{E}\left[\left\|\frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2 \middle| \text{Event } \bar{A}\right]$$

$$\overset{(a)}{\leq} \frac{1}{(1-\alpha)^2 K^2} \mathbb{E}\left[\left\|\sum_{k \in \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right) + \sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2\right] \quad \text{under Event } \bar{A}$$

$$\overset{(b)}{\leq} \frac{2}{(1-\alpha)^2 K^2} \left(\mathbb{E}\left[\left\|\sum_{k \in \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2\right]\right.$$

$$\left. + \underbrace{\mathbb{E}\left[\left\|\sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2\right]}_{I_{\mathcal{G}_t \setminus \mathcal{G}}}\right) \quad \text{under Event } \bar{A}$$

$$\tag{12}$$

where (a) follows from the fact that $|\mathcal{G}_t| \geq (1-\alpha)K$ and Event $\bar{A}$ implies $\mathcal{G} \subset \mathcal{G}$. (b) follows from Lemma 14. Now considering the two terms separately under Event $\bar{A}$, first consider the terms for $k \in \mathcal{G}$:

$$\mathbb{E}\left\|\sum_{k \in \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2 = \mathbb{E}\left\|\sum_{k \in \mathcal{G}} \left(\frac{1}{B_t} \sum_{i=1}^{B_t} \nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})\right)\right\|^2$$

$$\overset{(a)}{=} \frac{1}{B_t^2} \mathbb{E}\left[\sum_{k \in \mathcal{G}} \sum_{i=1}^{B_t} \left\|\nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})\right\|^2\right]$$

$$+ \frac{1}{B_t^2} \underbrace{\mathbb{E}\left[\sum_{(k,i) \neq (k',i'), k,k' \in \mathcal{G}} \langle \nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1}), \nabla f(\tilde{x}_{t-1}; \xi_{t,i'}^{(k')}) - \nabla f(\tilde{x}_{t-1}) \rangle\right]}_{=0}$$

$$= \frac{1}{B_t^2} \mathbb{E}\left[\sum_{k \in \mathcal{G}} \sum_{i=1}^{B_t} \left\|\nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})\right\|^2\right]$$

$$\overset{(b)}{\leq} \frac{K \mathcal{V}^2}{B_t} \tag{13}$$

where (a) follows from the fact that $f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)})$ are chosen uniformly independently across i and k. (b) follows from Assumption 2 and the fact that $|\mathcal{G}| \leq K$. Now consider the second term in (12)

15

defined as:

$$\mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}] = \mathbb{E}\left[\left\| \sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right) \right\|^2\right] \text{ under Event } \bar{A}$$

Note that all the nodes $k \in \mathcal{G}_t \setminus \mathcal{G}$ under Event $\bar{A}$ can come either from Rule 1 (Event $R1$ is true) or from Rule 2 (Event $R2$ is true). Again using the law of total expectation we can write $\mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}]$ as:

$$\mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}] \overset{(a)}{=} \mathbb{P}[\text{Event } R1 | \text{Event } \bar{A}] \, \mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}, \text{Event } R1]$$
$$+ \mathbb{P}[\text{Event } R2 | \text{Event } \bar{A}] \, \mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}, \text{Event } R2]$$
$$\overset{(b)}{\leq} \mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}, \text{Event } R1] + \left(\frac{\delta}{1 - \delta} \right) \mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}, \text{Event } R2]. \quad (14)$$

where (a) follows from the fact that Event $R2$ is the complement of Event $R1$ and the application of the law of total expectation. (b) follows form the following:

$$\mathbb{P}[\text{Event } R2 | \text{Event } \bar{A}] \overset{(c)}{=} \frac{\mathbb{P}[\text{Event } R2 \cap \text{Event } \bar{A}]}{\mathbb{P}[\text{Event } \bar{A}]} \overset{(d)}{\leq} \frac{\mathbb{P}[\text{Event } R2]}{\mathbb{P}[\text{Event } \bar{A}]} \overset{(e)}{\leq} \frac{\delta}{1 - \delta}.$$

where (c) follows since $\mathbb{P}[\text{Event } \bar{A}] \neq 0$. (d) follows since $[\text{Event } R2 \cap \text{Event } \bar{A}] \subset \text{Event } R2$ and finally, (e) follow from (10) and (7).

Now let us consider the first term of (14), we have:

$$\mathbb{E}[I_{\mathcal{G}_t \setminus \mathcal{G}} | \text{Event } \bar{A}, \text{Event } R1] = \mathbb{E}\left[\left\| \sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right) \right\|^2\right] \text{ under Event } \bar{A}, \text{Event } R1$$

$$\overset{(a)}{\leq} \alpha K \, \mathbb{E}\left[\sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left\| \mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right\|^2 \right] \text{ under Event } \bar{A}, \text{Event } R1$$

$$\overset{(b)}{\leq} 2\alpha K \, \mathbb{E}\left[\sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left\| \mu_t^{(k)} - \mu_t^{\text{med}} \right\|^2 \right]$$

$$+ 2\alpha K \, \mathbb{E}\Big[\underbrace{\sum_{k \in \mathcal{G}_t \setminus \mathcal{G}} \left\| \mu_t^{\text{med}} - \nabla f(\tilde{x}_{t-1}) \right\|^2}_{J_{\mathcal{G}_t \setminus \mathcal{G}}} \Big] \text{ under Event } \bar{A}, \text{Event } R1$$

$$\overset{(c)}{\leq} 2\alpha K \left[\alpha K \frac{16 \mathcal{V}^2 C}{B_t} \right] + 2\alpha K \left[\alpha K \frac{18 \mathcal{V}^2 C}{B_t} \right]$$

$$= \frac{68 \alpha^2 K^2 \mathcal{V}^2 C}{B_t}. \quad (15)$$

where (a) follows from Lemma 14 and the fact that $|\mathcal{G}_t \backslash \mathcal{G}| \leq \alpha K$, (b) follows by adding and subtracting μ_t^{med} and applying Lemma 14 and (c) follows from:

$$
\mathbb{E}[J_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]
$$
$$
= \mathbb{P}[\mathrm{Event}\ A | \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]\ \mathbb{E}[J_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ A, \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]
$$
$$
+ \mathbb{P}[\mathrm{Event}\ A! | \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]\ \mathbb{E}[J_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ A!, \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]
$$
$$
\stackrel{(d)}{\leq} \mathbb{E}[J_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ A, \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1] + \left(\frac{\delta}{1 - 2\delta} \right) \mathbb{E}[J_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ A!, \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]
$$
$$
\stackrel{(e)}{\leq} \alpha K \frac{9 \mathcal{V}^2 C}{B_t} + \left(\frac{\delta}{1 - 2\delta} \right) \alpha K 9 \mathcal{V}^2
$$
$$
\stackrel{(f)}{\leq} \alpha K \frac{18 \mathcal{V}^2 C}{B_t},
$$

where (d) follows from:

$$
\mathbb{P}[\mathrm{Event}\ A! | \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1] = \frac{\mathbb{P}[\mathrm{Event}\ A!, \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]}{\mathbb{P}[\mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]}
$$
$$
\leq \frac{\mathbb{P}[\mathrm{Event}\ A!]}{\mathbb{P}[\mathrm{Event}\ \bar{A}, \mathrm{Event}\ R1]}
$$
$$
\stackrel{(g)}{\leq} \frac{\delta}{1 - 2\delta},
$$

where (g) follows from the fact that we have:

$$
\mathbb{P}[\mathrm{Event}\ \bar{A} \cap \mathrm{Event}\ R1] = \mathbb{P}[\mathrm{Event}\ \bar{A}] + \mathbb{P}[\mathrm{Event}\ R1] - \mathbb{P}[\mathrm{Event}\ \bar{A} \cup \mathrm{Event}\ R1] \geq 1 - \delta + 1 - \delta - 1 = 1 - 2\delta.
$$

Moreover, (e) follows from the application of Lemma 9 and Lemma 10. Finally, (f) follows from choosing δ such that:

$$
e^{\frac{\delta B_t}{2(1 - 2\delta)}} \leq \frac{2K}{\delta}.
$$

Now let us consider the second term of (14), we have:

$$
\mathbb{E}[I_{\mathcal{G}_t \backslash \mathcal{G}} | \mathrm{Event}\ \bar{A}, \mathrm{Event}\ R2] = \mathbb{E}\left[\left\| \sum_{k \in \mathcal{G}_t \backslash \mathcal{G}} \left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right) \right\|^2 \right] \mathrm{under\ Event}\ \bar{A}, \mathrm{Event}\ R2
$$
$$
\stackrel{(a)}{\leq} \alpha K\ \mathbb{E}\left[\sum_{k \in \mathcal{G}_t \backslash \mathcal{G}} \left\| \mu_t^{(k)} - \nabla f(\tilde{x}_{t-1}) \right\|^2 \right] \mathrm{under\ Event}\ \bar{A}, \mathrm{Event}\ R2
$$
$$
\stackrel{(b)}{\leq} 2\alpha K\ \mathbb{E}\left[\sum_{k \in \mathcal{G}_t \backslash \mathcal{G}} \left\| \mu_t^{(k)} - \mu_t^{\mathrm{med}} \right\|^2 \right]
$$
$$
+ 2\alpha K\ \mathbb{E}\left[\sum_{k \in \mathcal{G}_t \backslash \mathcal{G}} \left\| \mu_t^{\mathrm{med}} - \nabla f(\tilde{x}_{t-1}) \right\|^2 \right] \mathrm{under\ Event}\ \bar{A}, \mathrm{Event}\ R2
$$
$$
\stackrel{(c)}{\leq} 2\alpha K \left[16 \alpha K \mathcal{V}^2 \right] + 2\alpha K \left[9 \alpha K \mathcal{V}^2 \right]
$$
$$
= 50 \alpha^2 K^2 \mathcal{V}^2 < 68 \alpha^2 K^2 \mathcal{V}^2. \tag{16}
$$

where (a) follows from Lemma 14 and the fact that $|\mathcal{G}_t \backslash \mathcal{G}| \leq \alpha K$, (b) follows by adding and subtracting μ_t^{med} and applying Lemma 14 and (c) follows from the fact that under Event $R2$ all nodes $k \in \mathcal{G}_t \backslash \mathcal{G}$ must satisfy Lemma 10 statement (b). Now replacing (15) and (16) in (14) we get:

$$\mathbb{E}[I_{\mathcal{G}_t \backslash \mathcal{G}} | \text{Event } \bar{A}] \leq \frac{68\alpha^2 K^2 \mathcal{V}^2 C}{B_t} + \left(\frac{\delta}{1-\delta}\right) 68\alpha^2 K^2 \mathcal{V}^2 \overset{(a)}{\leq} \frac{136\alpha^2 K^2 \mathcal{V}^2 C}{B_t}. \qquad (17)$$

where (a) follows from choosing δ such that:

$$e^{\frac{\delta B_t}{2(1-\delta)}} \leq \frac{2K}{\delta}.$$

Replacing (13) and (17) in (12), we have the bound on the first term of (11) as:

$$\mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}\right] \leq \frac{2}{(1-\alpha)^2 K^2}\left(\frac{K\mathcal{V}^2}{B_t} + \frac{136\alpha^2 K^2 \mathcal{V}^2 C}{B_t}\right)$$

$$= \frac{2\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t}. \qquad (18)$$

We have bounded the first term of (11). Now let us consider the second term of (11):

$$\delta \, \mathbb{E}\left[\|e_t\|^2 | \text{Event } \bar{A}!\right] = \delta \, \mathbb{E}\left[\left\|\frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t}\left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2\right] \text{ under Event } \bar{A}!$$

$$\overset{(a)}{\leq} \frac{\delta}{(1-\alpha)^2 K^2} \, \mathbb{E}\left[\left\|\sum_{k \in \mathcal{G}_t}\left(\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right)\right\|^2\right] \text{ under Event } \bar{A}!$$

$$\overset{(b)}{\leq} \frac{\delta}{(1-\alpha)^2 K} \, \mathbb{E}\left[\sum_{k \in \mathcal{G}_t}\left\|\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\right\|^2\right] \text{ under Event } \bar{A}!$$

$$\overset{(c)}{\leq} \frac{2\delta}{(1-\alpha)^2 K}\left(\mathbb{E}\left[\sum_{k \in \mathcal{G}_t}\left\|\mu_t^{(k)} - \mu_t^{\mathrm{med}}\right\|^2\right] + \mathbb{E}\left[\sum_{k \in \mathcal{G}_t}\left\|\mu_t^{\mathrm{med}} - \nabla f(\tilde{x}_{t-1})\right\|^2\right]\right) \text{ under Event } \bar{A}!$$

$$\overset{(d)}{\leq} \frac{2\delta}{(1-\alpha)^2 K}\left(16K\mathcal{V}^2 + 9K\mathcal{V}^2\right) = \delta \frac{50\mathcal{V}^2}{(1-\alpha)^2}. \qquad (19)$$

where (a) follows from the fact that $|\mathcal{G}_t| \geq (1-\alpha)K$. (b) follows from Lemma 14 and the fact that $|\mathcal{G}_t| \leq K$. (c) follows from adding and subtracting μ_t^{med} and applying Lemma 14. Finally, (d) follows by assuimg $\frac{2K}{\delta} \leq e^{\frac{B_t}{2}}$ and $|\mathcal{G}_t| \leq K$. The assumption $\frac{2K}{\delta} \leq e^{\frac{B_t}{2}}$ ensures that in the worst case the terms in inequality (c) are bounded.

Finally, replacing (18) and (19) in (11) we get:

$$\mathbb{E}\|e_t\|^2 \leq \frac{2\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t} + \delta \frac{50\mathcal{V}^2}{(1-\alpha)^2}$$

$$\overset{(a)}{\leq} \frac{4\mathcal{V}^2}{(1-\alpha)^2 K B_t} + \frac{272\alpha^2 \mathcal{V}^2 C}{(1-\alpha)^2 B_t}.$$

where (a) follows by choosing $\delta \leq \frac{1}{25K B_t}$. Therefore, we have the bound. ∎

18

Lemma 9 *For any $t \in [T]$ and for all $k \in \mathcal{G}$ with probability at least $1 - \delta$ (we call this event as Event A) we have:*

(a): $\|\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\| \leq \mathcal{V}\sqrt{\frac{C}{B_t}}.$

(b): This further implies that we have $\|\mu_t^{(k)} - \mu_t^{med}\| \leq 4\mathcal{V}\sqrt{\frac{C}{B_t}}$ *and* $\|\mu_t^{med} - \nabla f(\tilde{x}_{t-1})\| \leq 3\mathcal{V}\sqrt{\frac{C}{B_t}}.$

where C is defined as: $C = 2\log\left(\frac{2K}{\delta}\right).$

Let us denote this event as Event A.

Proof *(a)* The proof follows by considering the random variable, $\nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})$ and from Assumption 2 we have $\|\nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})\| \leq \mathcal{V}$. Now applying Lemma 13 on the summation with $\|\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\| = \|\frac{1}{B_t}\sum_{i=1}^{B_t}\left(\nabla f(\tilde{x}_{t-1}; \xi_{t,i}^{(k)}) - \nabla f(\tilde{x}_{t-1})\right)\|$, we get the result.

(b) follows from the straightforward application of the above result. ∎

Note that we call the above event by Event A and we have that the probability of Event A being true as: $\mathbb{P}[\text{Event A}] \geq 1 - \delta$ from Lemma 9 and the discussion above. Now let us consider the case when Event A! (Complement of Event A) is true. In that case, the discussion above implies that we have $\mathbb{P}[\text{Event A!}] \leq \delta$.

For the case when the set $|\mathcal{G}_t| < (1 - \alpha)K$ (please see Algorithm 1), we will make use of the following lemma.

Lemma 10 *For any $t \in [T]$ and for all $k \in \mathcal{G}$ we have:*

(a): $\|\mu_t^{(k)} - \nabla f(\tilde{x}_{t-1})\| \leq \mathcal{V}.$

(b): This further implies that we have $\|\mu_t^{(k)} - \mu_t^{med}\| \leq 4\mathcal{V}$ *and* $\|\mu_t^{med} - \nabla f(\tilde{x}_{t-1})\| \leq 3\mathcal{V}.$

Proof *(a)* The proof follows from the definition of $\mu_t^{(k)}$ and the application of the triangle inequality along with Assumption 2.

(b) follows from the straightforward application of the above. ∎

Lemma 11 *If $N \sim Geom(\Gamma)$ for $\Gamma > 0$. Then for any sequence $D_0, D_1, \ldots$ with $\mathbb{E}|D_N| < \infty$, we have*

$$\mathbb{E}[D_N - D_{N+1}] = \left(\frac{1}{\Gamma} - 1\right)(D_0 - \mathbb{E}D_N)$$

Proof Proof follows from Lei et al. [7]. ∎

Lemma 12 *For step size $\eta_t \leq \frac{1}{3LB_t^{2/3}}$, we have:*

(i) $\mathbb{E}\|\tilde{x}_t - \tilde{x}_{t-1}\|^2 < \infty.$

(ii) $\mathbb{E}(f(\tilde{x}_t) - f(\tilde{x}^*)) < \infty.$

(iii) $\mathbb{E}\|\nabla f(\tilde{x}_t)\|^2 < \infty.$

(iv) $\mathbb{E}|\langle e_t, \tilde{x}_t - \tilde{x}_{t-1} \rangle| < \infty.$

(v) $\mathbb{E}|\langle e_t, \nabla f(\tilde{x}_t) \rangle| < \infty.$

Proof The lemma is proven using induction and follows the same structure as the proof in Lei et al. [7]. The second inequality (2) in the proof of theorem yields

$$\mathbb{E}_{\xi_n^t} f(x_{n+1}^t) \leq f(x_n^t) - \eta_t(1 - L\eta_t)\|\nabla f(x_n^t)\|^2$$
$$- \eta_t\langle e_t, \nabla f(x_n^t)\rangle + \frac{L^3\eta_t^2}{2}\|x_n^t - x_0^t\|^2 + L\eta_t^2\|e_t\|^2, \qquad (20)$$

using Young's inequality $\langle a, b \rangle \leq \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ for any $\beta > 0$, on $-\eta_t\langle e_t, \nabla f(x_n^t)\rangle$ with $\beta = \frac{1}{2}$ we get:

$$-\eta_t\langle e_t, \nabla f(x_n^t)\rangle \leq \eta_t\|e_t\|^2 + \frac{\eta_t}{4}\|\nabla f(x_n^t)\|^2.$$

Moreover, using the fact that $\eta_t \leq \frac{1}{3LB_t^{2/3}} \leq \frac{1}{4L}$ since $B_t \geq 16$ and rearranging the terms in (20) we have

$$\eta_t\left(1 - L\eta_t - \frac{1}{4}\right)\|\nabla f(x_n^t)\|^2$$
$$\leq f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t) + \frac{L^3\eta_t^2}{2}\|x_n^t - x_0^t\|^2 + \eta_t(1 + L\eta_t)\|e_t\|^2$$
$$\eta_t\left(1 - \frac{1}{4} - \frac{1}{4}\right)\|\nabla f(x_n^t)\|^2$$
$$\leq f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t) + \frac{L^3\eta_t^2}{2}\|x_n^t - x_0^t\|^2 + \eta_t\left(1 + \frac{1}{4}\right)\|e_t\|^2$$
$$\eta_t\|\nabla f(x_n^t)\|^2 \leq 2\left(f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t)\right) + L^3\eta_t^2\|x_n^t - x_0^t\|^2 + \frac{5\eta_t}{2}\|e_t\|^2. \qquad (21)$$

Now using the first inequality (6) in Proof of Lemma 6, we have

$$\mathbb{E}_{\xi_n^t}\|x_{n+1}^t - x_0^t\|^2 \leq (1 + \eta_t^2 L^2)\|x_n^t - x_0^t\|^2 - 2\eta_t\langle\nabla f(x_n^t), x_n^t - x_0^t\rangle$$
$$- 2\eta_t\langle e_t, x_n^t - x_0^t\rangle + 2\eta_t^2\|\nabla f(x_n^t)\|^2 + 2\eta_t^2\|e_t\|^2,$$

using Young's inequality $\langle a, b \rangle \leq \frac{\beta}{2}\|a\|^2 + \frac{1}{2\beta}\|b\|^2$ for any $\beta > 0$, on $-2\eta_t\langle\nabla f(x_n^t), x_n^t - x_0^t\rangle$ and $-2\eta_t\langle e_t, x_n^t - x_0^t\rangle$ with $\beta = 8\eta_t B_t$ we get:

$$-2\eta_t\langle\nabla f(x_n^t), x_n^t - x_0^t\rangle \leq 8\eta_t^2 B_t\|\nabla f(x_n^t)\|^2 + \frac{1}{8B_t}\|x_n^t - x_0^t\|^2$$
$$-2\eta_t\langle e_t, x_n^t - x_0^t\rangle \leq 8\eta_t^2 B_t\|e_t\|^2 + \frac{1}{8B_t}\|x_n^t - x_0^t\|^2,$$

Therefore, we get:

$$\mathbb{E}_{\xi_n^t}\|x_{n+1}^t - x_0^t\|^2$$
$$\leq \left(1 + \eta_t^2 L^2 + \frac{1}{4B_t}\right)\|x_n^t - x_0^t\|^2 + (2\eta_t^2 + 8\eta_t^2 B_t)\|\nabla f(x_n^t)\|^2 + (2\eta_t^2 + 8\eta_t^2 B_t)\|e_t\|^2$$
$$\overset{(a)}{\leq} \left(1 + \frac{13}{36B_t}\right)\|x_n^t - x_0^t\|^2 + 10\eta_t^2 B_t\|\nabla f(x_n^t)\|^2 + 10\eta_t^2 B_t\|e_t\|^2 \qquad (22)$$

where (a) used the fact that we $\eta_t L \leq \frac{1}{3B_t^{2/3}}$. Now plugging (21) into (22) we get:

$$
\begin{aligned}
\mathbb{E}_{\xi_n^t}\|x_{n+1}^t - x_0^t\|^2 &\leq \left(1 + \frac{13}{36B_t}\right)\|x_n^t - x_0^t\|^2 + 10\eta_t^2 B_t\|e_t\|^2 \\
&\quad + 20\eta_t B_t\big(f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t)\big) + 10\eta_t^3 L^3 B_t\|x_n^t - x_0^t\|^2 + 25\eta_t^2 B_t\|e_t\|^2 \\
&\leq \left(1 + \frac{13}{36B_t} + 10\eta_t^3 L^3 B_t\right)\|x_n^t - x_0^t\|^2 + 20\eta_t B_t\big(f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t)\big) + 35\eta_t^2 B_t\|e_t\|^2 \\
&\overset{(a)}{\leq} \left(1 + \frac{711}{972B_t}\right)\|x_n^t - x_0^t\|^2 + 20\eta_t B_t\big(f(x_n^t) - \mathbb{E}_{\xi_n^t} f(x_{n+1}^t)\big) + 35\eta_t^2 B_t\|e_t\|^2 \qquad (23)
\end{aligned}
$$

where (a) follows from using $\eta_t L \leq \frac{1}{3B_t^{2/3}}$. Let us assume

$$
L_n^t = 20\eta_t B_t \mathbb{E}\big(f(x_n^t) - f(\tilde{x}^*)\big) + \mathbb{E}\|x_n^t - x_0^t\|^2
$$

Taking expectation over (23) we get:

$$
L_{n+1}^t \leq \left(1 + \frac{711}{972B_t}\right)L_n^t + 35\eta_t^2 B_t \mathbb{E}\|e_t\|^2
$$

Denoting $\gamma = \frac{711}{972} < 1$ we have:

$$
L_{n+1}^t \leq \left(1 + \frac{\gamma}{B_t}\right)L_n^t + 35\eta_t^2 B_t \mathbb{E}\|e_t\|^2
$$

$$
L_{n+1}^t + \frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma} \overset{(a)}{\leq} \left(1 + \frac{\gamma}{B_t}\right)L_n^t + \left(1 + \frac{\gamma}{B_t}\right)\frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma}
$$

$$
L_{n+1}^t + \frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma} \leq \left(1 + \frac{\gamma}{B_t}\right)\left(L_n^t + \frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma}\right)
$$

where (a) follows from adding and subtracting $\frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma}$ on both sides. this implies that we have:

$$
L_n^t \leq \left(1 + \frac{\gamma}{B_t}\right)^n\left(L_0^t + \frac{35\eta_t^2 B_t^2 \mathbb{E}\|e_t\|^2}{\gamma}\right)
$$

Since we have: $N_t \sim \text{Geom}\left(\frac{B_t}{B_t+1}\right)$, and assuming N_t can be 0 we have

$$
\mathbb{P}[N_t = n] = \frac{1}{B_t+1}\left(\frac{B_t}{B_t+1}\right)^n \leq \left(\frac{B_t}{B_t+1}\right)^n.
$$

Now the term:

$$
\mathbb{E}\left[\left(1 + \frac{\gamma}{B_t}\right)^{N_t}\right] \leq \sum_{n\geq 0}\left(\frac{B_t+\gamma}{B_t} \times \frac{B_t}{B_t+1}\right)^n = \sum_{n\geq 0}\left(\frac{B_t+\gamma}{B_t+1}\right)^n \overset{(a)}{=} \frac{B_t+1}{1-\gamma}.
$$

(a) follows since $\gamma = \frac{711}{972} < 1$. This implies that:

$$
\mathbb{E}L_{N_t}^t \leq \frac{B_t+1}{1-\gamma}\big(L_0^t + 70\eta_t^2 B_t \mathbb{E}\|e_t\|^2\big).
$$

which is finite since $\mathbb{E}\|e_t\| < \infty$ is finite by Lemma 7 as well as the filtering rule of Algorithm 1. The induction hypothesis implies that $\mathbb{E}L_{N_t} < \infty$.

All the claims follow. ∎

Lemma 13 (Alistarh et al. [1] Lemma 2.4) *Let the sequence of random variables $X_1, X_2, \ldots, X_N \in \mathbb{R}^d$ represent a random process such that we have $\mathbb{E}[X_n|X_1, \ldots, X_{n-1}] = 0$ and $\|X_n\| \leq M$. Then,*

$$\mathbb{P}[\|X_1 + \ldots + X_N\|^2 \leq 2\log(2/\delta)M^2 N] \geq 1 - \delta.$$

Lemma 14 *For $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, we have*

$$\|X_1 + X_2 + \ldots + X_n\|^2 \leq n\|X_1\|^2 + n\|X_2\|^2 + \ldots + n\|X_n\|^2.$$