Layerwise and Dimensionwise Adaptive Local AMSMethod for Federated Learning

Abstract

To be completed...

1   Introduction

A growing and important task while learning models on observed data, is the ability to train the latter over a large number of clients which could either be devices or distinct entities. In the paradigm of Federated Learning (FL) [6, 9], the focus of our paper, a central server orchestrates the optimization over those clients under the constraint that the data can neither be centralized nor shared among the clients. Most modern machine learning tasks can be casted as a large finite-sum optimization problem written as:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \tag{1}$$

where n denotes the number of workers, f_i represents the average loss for worker i and θ the global model parameter taking value in Θ a subset of $\mathbb{R}^d$. While this formulation recalls that of distributed optimization, the core principle of FL is different that standard distributed paradigm.

FL currently suffers from two bottlenecks: communication efficiency and privacy. We focus on the former in this paper. While local updates, updates during which each client learn their local models, can reduce drastically the number of communication rounds between the central server and devices, new techniques must be employed to tackle this challenge. Some quantization [1, 16] or compression [8] methods allow to decrease the number of bits communicated at each round and are efficient method in a distributed setting. The other approach one can take is to accelerate the local training on each device and thus sending a better local model to the server at each round.

Under the important setting of heterogenous data, i.e. the data among each device can be distributed according to different distributions, current local optimization algorithms are perfectible. The most popular method for FL is using multiple local Stochastic Gradient Descent (SGD) steps in each device, sending those local models to the server that computes the average over those received local vector of parameters and broadcasts it back to the devices. This is called FedAvg and has been introduced in [9].

In [2], the authors motivate the usage of adaptive gradient optimization methods as a better alternative to the standard SGD inner loop in FedAvg. They propose an adaptive gradient method, namely Local AMSGrad, with communication cost sublinear in T that is guaranteed to converge to stationary points in $\mathcal{O}(\sqrt{d/Tn})$, where T is the number of iterations.

Based on recent progress in adaptive methods for accelerating the training procedure, see [17], we propose a variant of Local AMSGrad integrating dimensionwise and layerwise adaptive learning rate in each device's local update. Our contributions are as follows:

- We develop a novel optimization algorithm for federated learning, namely Fed-LAMB, following a principled layerwise adaptation strategy to accelerate training of deep neural networks.

- We provide a rigorous theoretical understanding of the non asymptotic convergence rate of Fed-LAMB. Based on the recent progress on nonconvex stochastic optimization, we derive for a any finite number of rounds performed by our method, a characterization of the rate at which the classical

suboptimilality condition, *i.e.*, , the second order moment of the gradient of the objective function, decreases. Our bound in $\mathcal{O}(\sqrt{\frac{pL}{nR}})$ matches state of the art methods in Federated Learning reaching a sublinear convergence in R, the total number of rounds.

- We exhibit the advantages of our method on several benchmarks supervised learning methods on both homogeneous and heterogeneous settings.

The plan of our paper is as follows. After having established a literature review of both realms of federated and adaptive learning in subsection 1.1, we develop in Section 2, our method, namely FED-LAMB, based on the computation per layer and per dimension, of a scaling factor in the traditional stepsize of AMSGrad. Theoretical understanding of our method's behaviour with respect to convergence towards a stationary point is developed in Section **??**. We present numerical illustrations showing the advantages of our method in Section 4.

1.1   Related Work

Adaptive gradient methods.   In classical stochastic nonconvex optimization, adaptive methods have proven to be the spearhead of any practicioner. Those gradient based optimization algorithms alleviate the possibly high nonconvexity of the objective function by adaptively updating each coordinate of their learning rate using past gradients. Most used examples AMSGrad [14], Adam [5], RMSPROP [15], AdADELTA [18], and NADAM [3].

Their popularity and efficiency are due to their great performance at training deep neural networks. They generally combine the idea of adaptivity from AdaGrad [4, 10], as explained above, and the idea of momentum from Nesterov's Method [11] or Heavy ball method [12] using past gradients. AdaGrad displays a great edge when the gradient is sparse compared to other classical methods. Its update has a notable feature: it leverages an anisotropic learning rate depending on the magnitude of the gradient for each dimension which helps in exploiting the geometry of the data.

The anisotropic nature of this update represented a real breakthrough in the training of high dimensional and nonconvex loss functions. This adaptive learning rate helps accelerating the convergence when the gradient vector is sparse [4], yet, when applying AdaGrad to train deep neural networks, it is observed that the learning rate might decay too fast, see [5] for more details. Consequently, [5] develops Adam leveraging a moving average of the gradients divided by the square root of the second moment of this moving average (element-wise multiplication). A variant, called AMSGrad described in [14] ought to fix Adam failures and is presented in Algorithm 1.

Algorithm 1 AMSGrad [14]

1: **Required**: parameter β_1, β_2, and η_t.
2: Init: $w_1 \in \Theta \subseteq \mathbb{R}^d$ and $v_0 = \epsilon \mathbf{1} \in \mathbb{R}^d$.
3: **for** $t = 1$ to T **do**
4:     Get mini-batch stochastic gradient g_t at w_t.
5:     $\theta_t = \beta_1 \theta_{t-1} + (1 - \beta_1)g_t$.
6:     $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$.
7:     $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$.
8:     $w_{t+1} = w_t - \eta_t \frac{\theta_t}{\sqrt{\hat{v}_t}}$.   (element-wise division)
9: **end for**

The difference between Adam and AMSGrad lies in Line 7 of Algorithm 1.

A natural extension of Algorithm 1 has been developed in [17] specifically for multi layered neural network. A principled layerwise adaptation strategy to accelerate training of deep neural networks using large mini-batches is proposed using either a standard stochastic gradient update or a generalized adaptive method under the setting of a classical single server empirical risk minimization problem.

Federated learning.   An extension of the well known parameter server framework, where a model is being trained on several servers in a distributed manner, is called Federated Learning, see [6]. Here, the central

server only plays the role of compute power for aggregation and global update of the model. Compared with the distributed learning paradigm, in Federated Learning, the data stored in each worker must not be seen by the central server – preserving privacy is key – and the nature of those workers, which can be mobile devices, combined with their usually large amount make communication between the devices and the central server less appealing – communication cost needs to be controlled.

Thus, while traditional distributed gradient methods [13, 7, 19] do not respect those constraints, it has been proposed in [9], an algorithm called Federated Averaging – FED-AVG – extending parallel SGD with local updates performed on each device. In FED-AVG, each worker updates their own model parameters locally using SGD, and the local models are synchronized by periodic averaging on the central parameter server.

2 Layerwise and Dimensionwise Adaptive Methods

Beforehand, it is important to provide useful and important notations used throughout our paper.

Notations: We denote by θ the vector of parameters taking values in $\mathbb{R}^d$. For each layer $\ell \in [\![\mathsf{L}]\!]$, where L is the total number of layers of the neural networks, and each coordinate $j \in [\![p_\ell]\!]$ where p_ℓ is the dimension per layer ℓ, we note $\theta^{\ell,j}$ its jth coordinate. The gradient of f with respect to θ^ℓ is denoted by $\nabla_\ell f(\theta)$. The index $i \in [\![n]\!]$ denotes the index of the worker i in our federated framework. r and t are used as the round and local iteration numbers respectively. The smoothness per layer is denoted by L_ℓ for each layer $\ell \in [\![\mathsf{L}]\!]$.

2.1 AMSGrad, Local AMSGrad and Periodic Averaging

We have introduced the vanilla AMSGrad method in Algorithm 1

2.2 Layerwise and Dimensionwise Learning with Periodic Averaging

In this section, we will recall our problem setting and introduce our novel method for performing efficient federated learning among large amount of devices.

We propose a layerwise and dimensionwise local AMS algorithm in the following:

Algorithm 2 FED-LAMB for Federated Learning

1: **Input**: parameter β_1, β_2, and learning rate α_t.
2: Init: $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, as the global model shared by all devices and $v_0 = \epsilon 1 \in \mathbb{R}^d$ and $\bar{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n}\theta_0$.
3: **for** $r = 1$ to R **do**
4:     Set $\theta_{r,i}^0 = \bar{\theta}_{r-1}$
5:     **parallel for device** $d \in D^r$ **do**:
6:     Compute stochastic gradient $g_{r,i}$ at θ_r.
7:     **for** $t = 1$ to T **do**
8:         $m_{r,i}^t = \beta_1 m_{r-1,i}^{t-1} + (1 - \beta_1)g_{r,i}$.
9:         $m_{r,i}^t = m_{r,i}^t / (1 - \beta_1^r)$.
10:       $v_r^{t,i} = \beta_2 v_{r-1,i}^t + (1 - \beta_2)g_{r,i}^2$.
11:       $v_{r,i}^t = v_{r,i}^t / (1 - \beta_2^r)$.
12:       $\hat{v}_r^t = \max(\hat{v}_{r-1}^t, \frac{1}{n}\sum_{i=1}^{n} v_{r,i}^t)$.
13:       Compute ratio $p_{r,i} = \frac{m_{r,i}^t}{\sqrt{v_r^t} + \epsilon}$.
14:       Update local model for each layer ℓ:

$$\theta_{r,i}^{\ell,t} = \theta_{r,i}^{\ell,t-1} - \alpha_r \phi(\|\theta_{r,i}^{\ell,t-1}\|)(p_{r,i}^\ell + \lambda\theta_{r,i}^{\ell,t-1})/\|p_{r,i}^\ell + \lambda\theta_{r,i}^{\ell,t-1}\|$$

15:     **end for**
16:     Devices send local model $\theta_{r,i}^T = [\theta_{r,i}^{\ell,T}]_{\ell=1}^{\mathsf{L}}$ to the server
17:     Server computes the averages of the local models $\bar{\theta}_r^\ell = \frac{1}{n}\sum_{i=1}^{n}\theta_{r,i}^{\ell,T}$ and send it back to the devices.
18: **end for**

---

3   On The Convergence of Fed-LAMB

In the context of nonconvex stochastic optimization for distributed devices, assume the following:

H1. *For $i \in [\![n]\!]$ and $\ell \in [\![L]\!]$, f_i is L-smooth: $\|\nabla f_i(\theta) - \nabla f_i(\vartheta)\| \leq L_\ell \left\|\theta^\ell - \vartheta^\ell\right\|$.*

We add some classical assumption in the unbiased stochastic optimization realm, on the gradient of the objective function:

H2. *The stochastic gradient is unbiased for any iteration $r > 0$: $\mathbb{E}[g_r] = \nabla f(\theta_r)$ and is bounded from above, i.e., $\|g_t\| \leq M$.*

H3. *The variance of the stochastic gradient is bounded for any iteration $r > 0$ and any dimension $j \in [\![d]\!]$: $\mathbb{E}[|g_r^j - \nabla f(\theta_r)^j|^2] < \sigma^2$.*

H4. *For any value $a \in \mathbb{R}_+^*$, there exists strictly positive constants such that $\phi_m \leq \phi(a) \leq \phi_M$.*

We now state our main result regarding the non asymptotic convergence analysis of our Algorithm 2:

Theorem 1. *Consider $\{\overline{\theta}_r\}_{r>0}$, the sequence of parameters obtained running Algorithm 2. Then, if the number of local epochs is set to $T = 1$ and $\epsilon = \lambda = 0$, we have:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\overline{\theta_r})\|^2] \leq dd \tag{2}$$
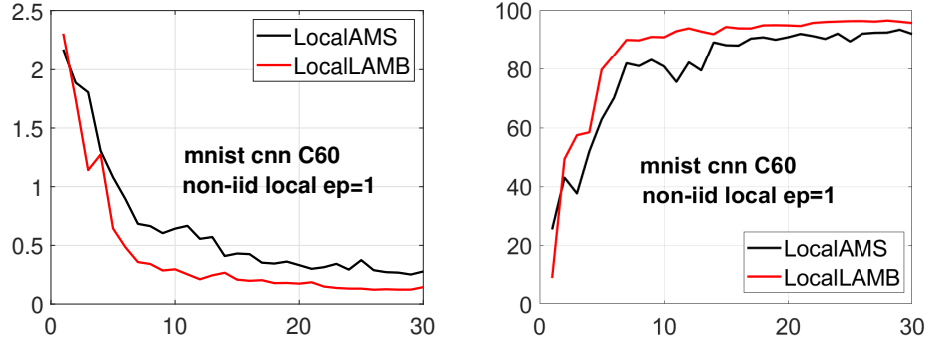
4 Numerical experiments

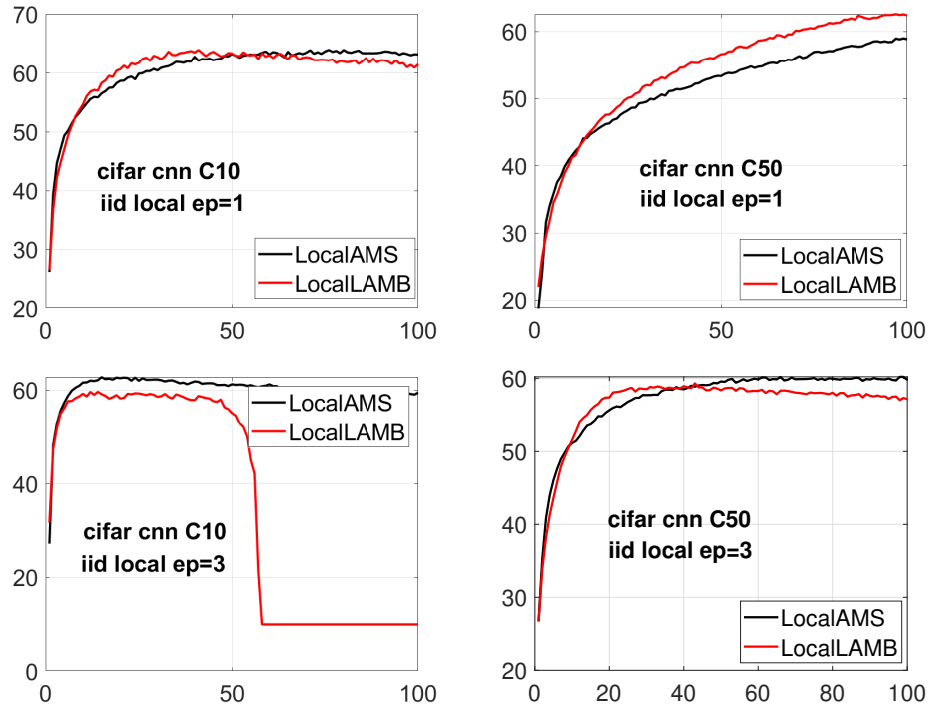Figure 1: Test accuracy on CNN + MNIST. Non-iid data distribution.

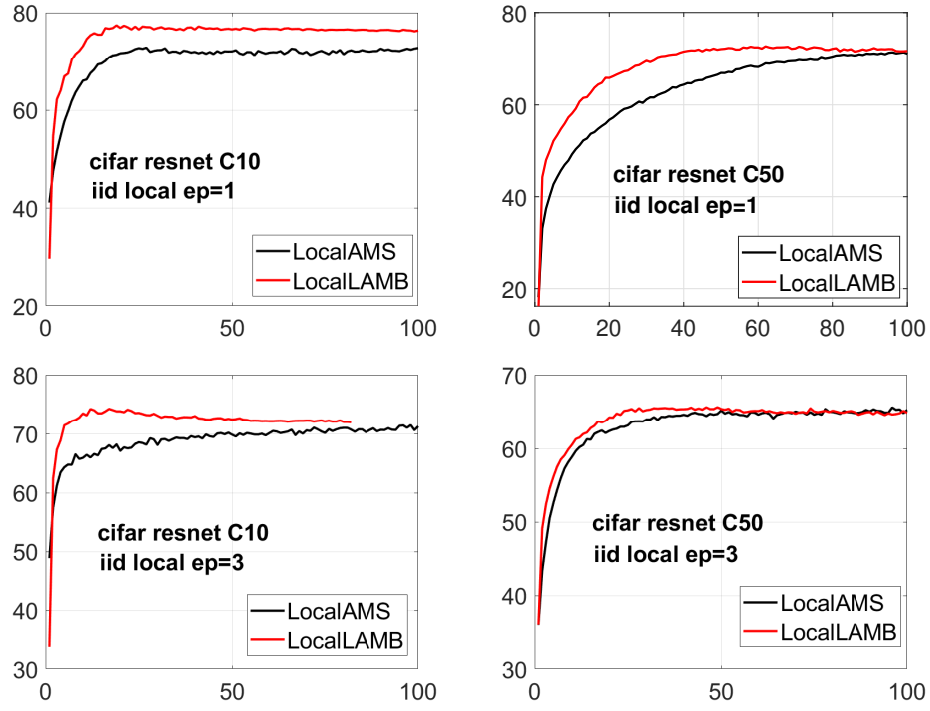Figure 2: Test accuracy on CNN + CIFAR10. iid data distribution.

Figure 3: Test accuracy on ResNet + CIFAR10. iid data distribution.

5 Conclusion

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[2] Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *ACM-IMS Foundations of Data Science Conference (FODS)*, Seattle, WA, 2020.

[3] Timothy Dozat. Incorporating nesterov momentum into adam. *ICLR (Workshop Track)*, 2016.

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 2011.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[6] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[7] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 583–598, 2014.

[8] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

[9] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[10] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. *COLT*, 2010.

[11] Yurii Nesterov. Introductory lectures on convex optimization: A basic course. *Springer*, 2004.

[12] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Mathematics and Mathematical Physics*, 1964.

[13] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24:693–701, 2011.

[14] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *ICLR*, 2018.

[15] T. Tieleman and G. Hinton. Rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

[16] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309, 2018.

[17] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[18] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv:1212.5701*, 2012.

[19] Weijie Zhao, Deping Xie, Ronglai Jia, Yulei Qian, Ruiquan Ding, Mingming Sun, and Ping Li. Distributed hierarchical gpu parameter server for massive scale deep learning ads systems. *arXiv preprint arXiv:2003.05622*, 2020.

A   Appendix

B   Theoretical Analysis

B.1   Intermediary Lemmas

Lemma 1. *Consider* $\{\overline{\theta_r}\}_{r>0}$, *the sequence of parameters obtained running Algorithm 2. Then for $i \in [\![n]\!]$:*

$$\|\overline{\theta_r} - \theta_{r,i}\| \leq \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0} \tag{3}$$

Proof. Assuming the simplest case when $T = 1$, i.e. one local iteration, then by construction of Algorithm 2, we have for all $\ell \in [\![\mathsf{L}]\!]$, $i \in [\![n]\!]$ and $r > 0$:

$$\theta_{r,i}^\ell = \overline{\theta_r}^\ell - \alpha\phi(\|\theta_{r,i}^{\ell,t-1}\|)p_{r,i}^j/\|p_{r,i}^\ell\| = \overline{\theta_r}^\ell - \alpha\phi(\|\theta_{r,i}^{\ell,t-1}\|)\frac{m_{r,i}^t}{\sqrt{v_r^t}}\frac{1}{\|p_{r,i}^\ell\|} \tag{4}$$

leading to

$$\begin{aligned}
\|\overline{\theta_r} - \theta_{r,i}\|^2 &= \left\langle \overline{\theta_r}^\ell - \theta_{r,i}^\ell \,|\, \overline{\theta_r}^\ell - \theta_{r,i}^\ell \right\rangle \\
&\leq \alpha^2 M^2 \phi_M^2 \frac{(1-\beta_2)p}{v_0}
\end{aligned} \tag{5}$$

which concludes the proof. □

B.2   Proof of Theorem 1

Theorem. *Consider* $\{\overline{\theta_r}\}_{r>0}$, *the sequence of parameters obtained running Algorithm 2. Then, if the number of local epochs is set to $T = 1$ and $\epsilon = \lambda = 0$, we have:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\overline{\theta_r})\|^2 \leq dd \tag{6}$$

Case with $T = 1$, $\epsilon = 0$ and $\lambda = 0$: Using H1, we have:

$$\begin{aligned}
f(\bar{\vartheta}_{r+1}) &\leq f(\bar{\vartheta}_r) + \left\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \right\rangle + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2 \\
&\leq f(\bar{\vartheta}_r) + \sum_{\ell=1}^{\mathsf{L}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j(\bar{\vartheta}_{r+1}^{\ell,j} - \bar{\vartheta}_r^{\ell,j}) + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2
\end{aligned} \tag{7}$$

Taking expectations on both sides leads to:

$$-\mathbb{E}[\left\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \right\rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2] \tag{8}$$

Yet, we observe that, using the classical intermediate quantity, used for proving convergence results of adaptive optimization methods, see [], we have:

$$\bar{\vartheta}_r = \overline{\theta}_r + \frac{\beta_1}{1-\beta_1}(\overline{\theta}_r - \overline{\theta}_{r-1}) \tag{9}$$

where $\bar{\theta}_r$ denotes the average of the local models at round r. Then for each layer ℓ,

$$\bar{\vartheta}^\ell_{r+1} - \bar{\vartheta}^\ell_r = \frac{1}{1-\beta_1}(\bar{\theta}^\ell_{r+1} - \bar{\theta}^\ell_r) - \frac{\beta_1}{1-\beta_1}(\bar{\theta}^\ell_r - \bar{\theta}^\ell_{r-1}) \tag{10}$$

$$= \frac{\alpha_r}{1-\beta_1}\frac{1}{n}\sum_{i=1}^n \frac{\phi(\|\theta^\ell_{r,i}\|)}{\|p^\ell_{r,i}\|}p^\ell_{r,i} - \frac{\alpha_{r-1}}{1-\beta_1}\frac{1}{n}\sum_{i=1}^n \frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\|p^\ell_{r-1,i}\|}p^\ell_{r-1,i} \tag{11}$$

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^n \left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|} - \frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|} \right) m^t_{r-1} + \frac{\alpha}{n}\sum_{i=1}^n \frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}g_{r,i} \tag{12}$$

where we have assumed a constant learning rate α.

We note for all $\theta \in \Theta$, the majorant $G > 0$ such that $\phi(\|\theta\|) \leq G$. Then, following (8), we obtain:

$$-\mathbb{E}[\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \rangle] \leq \mathbb{E}[f(\bar{\vartheta}_r) - f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^L \frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\|^2] \tag{13}$$

Developing the LHS of (13) using (10) leads to

$$\langle \nabla f(\bar{\vartheta}_r) \,|\, \bar{\vartheta}_{r+1} - \bar{\vartheta}_r \rangle = \sum_{\ell=1}^L \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j (\bar{\vartheta}^{\ell,j}_{r+1} - \bar{\vartheta}^{\ell,j}_r) \tag{14}$$

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{\ell=1}^L\sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \left[\sum_{i=1}^n \left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|} - \frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|} \right) m^t_{r-1} \right] \tag{15}$$

$$\underbrace{- \frac{\alpha}{n}\sum_{\ell=1}^L\sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \sum_{i=1}^n \frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}g_{r,i}}_{=A_1} \tag{16}$$

Term A_1: Since we have that $\|p^\ell_{r,i}\| \leq \sqrt{\frac{p_\ell}{1-\beta_2}}$ and $1/\sqrt{v^t_r} \leq 1/\sqrt{v_0}$, using H2, we develop the term A_1 as follows:

$$A_1 \leq -\frac{\alpha}{n}\sum_{\ell=1}^L\sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \sum_{i=1}^n \frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}g_{r,i} \tag{17}$$

$$\leq -\frac{\alpha}{n}\sum_{\ell=1}^L \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^n\sum_{j=1}^{p_\ell} \phi(\|\theta^\ell_{r,i}\|)\nabla_\ell f(\bar{\vartheta}_r)^j g^{\ell,j}_{r,i} \tag{18}$$

$$- \frac{\alpha}{n}\sum_{\ell=1}^L\sum_{i=1}^n\sum_{j=1}^{p_\ell} \left(\phi(\|\theta^\ell_{r,i}\|)\nabla_\ell f(\bar{\vartheta}_r)^j \frac{p^\ell_{r,i}}{\|p^\ell_{r,i}\|} \right) \mathbf{1}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(p^\ell_{r,i})\right) \tag{19}$$

Taking the expectations on both sides yields:

9

$$\mathbb{E}[A_1] \le -\alpha \sum_{\ell=1}^{\mathsf{L}} \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \mathbb{E}\left[\phi(\|\theta_{r,i}^\ell\|)\nabla_\ell f(\bar{\vartheta}_r)^j g_{r,i}^{\ell,j}\right] \tag{20}$$

$$-\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \mathbb{E}\left[\phi(\|\theta_{r,i}^\ell\|)\nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|} \mathbf{1}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(p_{r,i}^\ell)\right)\right] \tag{21}$$

$$\le -\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \phi_m \sqrt{\frac{1-\beta_2}{M^2 p_\ell}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} (\nabla_\ell f(\bar{\vartheta}_r)^j)^2 \tag{22}$$

$$-\frac{\alpha}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \phi_M \mathbb{E}\left[\left|\nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|}\right| \mathbf{1}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(p_{r,i}^\ell)\right)\right] \tag{23}$$

$$\tag{24}$$

where we have used assumption H4.

Since for any ℓ, i, j, we have

$$\mathbb{E}\left[\left|\nabla_\ell f(\bar{\vartheta}_r)^j \frac{p_{r,i}^\ell}{\|p_{r,i}^\ell\|}\right| \mathbf{1}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(p_{r,i}^\ell)\right)\right] \le \left|\nabla_\ell f(\bar{\vartheta}_r)^j\right| \mathbb{P}\left(\text{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \text{sign}(p_{r,i}^\ell)\right) \tag{25}$$

Then, we obtain

$$\mathbb{E}[A_1] \le -\alpha \phi_m \sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2 p}} \mathbb{E}[\|\overline{\nabla f}(\bar{\vartheta}_r)\|^2] - \alpha \phi_M \sum_{\ell=1}^{\mathsf{L}} \sum_{i=1}^{n} \sum_{j=1}^{p_\ell} \frac{\sigma_i^{\ell,j}}{\sqrt{n}} \tag{26}$$

where $\overline{\nabla f}(\cdot) = \sum_{i=1}^{n} \nabla f_i(\cdot)$

We now need to bound the following terms:

$$A_r^2 := \mathbb{E}[\|\bar{\vartheta}_{r+1} - \bar{\vartheta}_r\|^2] \tag{27}$$

$$A_r^3 := \frac{\alpha \beta_1}{1-\beta_1} \frac{1}{n} \sum_{\ell=1}^{\mathsf{L}} \sum_{j=1}^{p_\ell} \nabla_\ell f(\bar{\vartheta}_r)^j \left[\sum_{i=1}^{n} \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right] \tag{28}$$

Term A_r^2: According to definition (9), for each layer $\ell \in [\![\mathsf{L}]\!]$, we have, using the Cauchy-Schwartz inequality, that:

$$\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2 = \left\|\frac{\alpha \beta_1}{1-\beta_1} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t + \frac{\alpha}{n} \sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}\right\|^2 \tag{29}$$

$$\le 2\frac{\alpha^2}{n^2} \left\|\frac{\beta_1}{1-\beta_1} \sum_{i=1}^{n} \left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right) m_{r-1}^t\right\|^2 + \frac{1}{n^2}\left\|\sum_{i=1}^{n} \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} g_{r,i}\right\|^2 \tag{30}$$

Taking the expectation on both sides leads to:

$$
\mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2] \leq 2\alpha^2 \mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right)m_{r-1}^t\right\|^2\right] + \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}g_{r,i}\right\|^2\right]
$$

$$
\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right)m_{r-1}^t\right\|^2\right]
$$

$$
+ \frac{1}{n^2}\mathbb{E}\left[\left|\sum_{i=1}^n\sum_{j=1}^p\left\langle\Gamma_{r,i}^j(\nabla f_i(\theta_r)^j + g_{r,i}^j - \nabla f_i(\theta_r)^j)\,|\,\Gamma_{r,i}^j(\nabla f_i(\theta_r)^j + g_{r,i}^j - \nabla f_i(\theta_r)^j)\right\rangle\right|\right]
$$

$$
\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right)m_{r-1}^t\right\|^2\right]
$$

$$
+ \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] + \frac{1}{n}\left\|\sum_{i=1}^n\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right]\right\|^2
$$

$$(31)$$

where the last line uses assumptions H2 and H3 (unbiased gradient and bounded variance of the stochastic gradient) and $\Gamma := \frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}$.

On the other hand, using the bound on the gradient H2,

$$
\sum_{r=1}^R\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{\phi(\|\theta_{r-1,i}^\ell\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right)m_{r-1}^t\right\|^2\right]
$$

$$
\leq \frac{\beta_1^2}{(1-\beta_1)^2}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\frac{1}{\sqrt{v_r^t}\|p_{r,i}^\ell\|} - \frac{1}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^\ell\|}\right)\right\|^2\right]
$$

$$
\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\frac{1}{\sqrt{v_r^t}} - \frac{1}{\sqrt{v_{r-1}^t}}\right)\right\|^2\right]
$$

$$(32)$$

$$
\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left|\sum_{i=1}^n\sum_{j=1}^p\left(\frac{1}{\sqrt{v_r^{t,j}}} - \frac{1}{\sqrt{v_{r-1}^{t,j}}}\right)\right|\right]
$$

$$
\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}
$$

where, in the telescopic sum, we have used the initial value v_0 of the non decreasing sequence $\{v_r^t\}_{r>0}$ by construction (max operator).

Combining (32) into (31) and summing over the total number of rounds R yields

$$
\sum_{r=1}^R A_r^2 := \sum_{r=1}^R \mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell - \bar{\vartheta}_r^\ell\|^2] \leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}
$$

$$
+ \sum_{r=1}^R\left[\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] + \frac{1}{n}\left\|\sum_{i=1}^n\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right]\right\|^2\right]
$$

$$(33)$$

Term A_r^3: According to similar arguments on the non decreasing sequence involved in the algorithm as in the previous series of calculations, observe that

$$\sum_{r=1}^{R} A_r^3 \leq \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}} \tag{34}$$

Plugging (26) into (13) combined with (33) and (34) injected into the original smoothness definition (8) summed over the total number of rounds:

$$-\sum_{r=1}^{R}\mathbb{E}[\langle\nabla f(\bar{\vartheta}_r)\,|\,\bar{\vartheta}_{r+1}-\bar{\vartheta}_r\rangle] \leq \sum_{r=1}^{R}\mathbb{E}[f(\bar{\vartheta}_r)-f(\bar{\vartheta}_{r+1})] + \sum_{r=1}^{R}\sum_{\ell=1}^{L}\frac{L_\ell}{2}\mathbb{E}[\|\bar{\vartheta}_{r+1}^\ell-\bar{\vartheta}_r^\ell\|^2] \tag{35}$$

gives:

$$\sum_{r=1}^{R}\alpha\phi_m\sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2 p}}\mathbb{E}[\|\overline{\nabla f}(\bar{\vartheta}_r)\|^2] - \alpha\phi_M\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\frac{\sigma_i^{\ell,j}}{\sqrt{n}} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}$$

$$\leq \sum_{r=1}^{R}\mathbb{E}[f(\bar{\vartheta}_r)-f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}$$

$$-\sum_{r=1}^{R}\left[\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right]\right] \tag{36}$$

Noting that $\sum_{r=1}^{R}\mathbb{E}[f(\bar{\vartheta}_r)-f(\bar{\vartheta}_{r+1})] = f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})]$, we obtain

$$\sum_{r=1}^{R}\alpha\phi_m\sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2 p}}\mathbb{E}[\|\overline{\nabla f}(\bar{\vartheta}_r)\|^2] + \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right]$$

$$\leq f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right] + \alpha\phi_M\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\frac{\sigma_i^{\ell,j}}{\sqrt{n}} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}$$

$$+\sum_{\ell=1}^{L}\frac{L_\ell}{2}\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0} \tag{37}$$

leading to

$$\sum_{r=1}^{R}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] \leq f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right]$$

$$+\alpha\phi_M\sigma\mathsf{L}p\sqrt{n} + \frac{\overline{L}_\ell\beta_1^2\mathsf{L}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2 v_0} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}} \tag{38}$$

where $\overline{L}_\ell = \sum_{\ell=1}^{\mathsf{L}} L_\ell$ is the sum of all smoothness constants.

Consider the following inequality:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r) \leq \phi_M(1-\beta_2)\frac{\overline{\nabla f}(\theta_r)}{\sqrt{v_r^t}} \tag{39}$$

12

where $\overline{\nabla} f(\theta_r) := \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_r)$. And using the Cauchy-Schwartz inequality we have

$$\left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\| \geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| - \left\| \frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| \tag{40}$$

Using Lemma 1 and the smoothness assumption H1, we have

$$\begin{aligned} \left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\| &\geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| - \left\| \frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| \\ &\geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| - \overline{L} \alpha^2 M^2 \phi_M^2 \frac{(1 - \beta_2) p}{v_0} \end{aligned} \tag{41}$$

Plugging the above inequality into (38) returns:

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^{R} \mathbb{E} \left[\left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\|^2 \right] &\leq \left(f(\bar{\vartheta}_1) - \mathbb{E}[f(\bar{\vartheta}_{R+1})] \right) + \frac{1}{n} \left\| \sum_{i=1}^{n} \sigma_i^2 \mathbb{E} \left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t} \|p_{r,i}^\ell\|} \right] \right\|^2 \right] \\ &\quad + \alpha \phi_M \sigma \mathsf{L} p \sqrt{n} + \frac{\overline{L}_\ell \beta_1^2 \mathsf{L}(1 - \beta_2) M^2 \phi_M^2 n}{2(1 - \beta_1)^2 v_0} + \frac{\alpha \beta_1}{1 - \beta_1} \sqrt{(1 - \beta_2) p} \frac{\mathsf{L} M^2}{\sqrt{v_0}} + \overline{L} \alpha^2 M^2 \phi_M^2 \frac{(1 - \beta_2) p}{R v_0} \end{aligned} \tag{42}$$

concluding the proof of our main convergence result.