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) [3, 5], 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, 6] or compression [4] 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 [5].

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 [7], 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.
- theoretical results
- We exhibit the advantages of our method on several benchmarks supervised learning methods on both homogeneous and heterogeneous settings.

1.1 Related Work

Federated learning.

Adaptive gradient methods.

2 Layerwise and Dimensionwise Adaptive Methods

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 Local AMS with LAMB

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

Algorithm 1 L&D Local AMS for Federated Learning

Input: parameter β_1, β_2, and learning rate α_t.
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 $\bar{\theta}_0 = \frac{1}{n}\sum_{i=1}^n \theta_0$.
for $r = 1$ to R **do**
  Set $\theta_{r,i}^0 = \bar{\theta}_{r-1}$
  parallel for device $d \in D^r$ **do**:
  Compute stochastic gradient $g_{r,i}$ at θ_r.
  for $t = 1$ to T **do**
    $m_{r,i}^t = \beta_1 m_{r-1,i}^{t-1} + (1-\beta_1) g_{r,i}$.
    $m_{r,i}^t = m_{r,i}^t / (1-\beta_1^r)$.
    $v_r^{t,i} = \beta_2 v_{r-1,i}^t + (1-\beta_2) g_{r,i}^2$.
    $v_{r,i}^t = v_{r,i}^t / (1-\beta_2^r)$.
    $\hat{v}_r^t = \max(\hat{v}_{r-1}^t, \frac{1}{n}\sum_{i=1}^n v_{r,i}^t)$.
    Compute ratio $p_{r,i} = \frac{m_{r,i}^t}{\sqrt{v_r^t + \epsilon}}$.
    Update local model for each layer ℓ:

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

  end for
  Devices send local model $\theta_{r,i}^T = [\theta_{r,i}^{\ell,T}]_{\ell=1}^{\mathsf{L}}$ to the server
  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.
end for

2.2 Finite time convergence bounds

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 1:

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

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

3 Numerical experiments

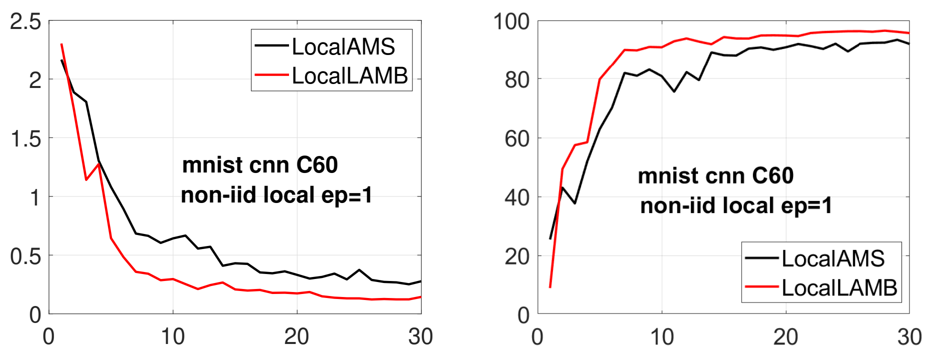

Figure 1: Test accuracy on CNN + MNIST. Non-iid data distribution.

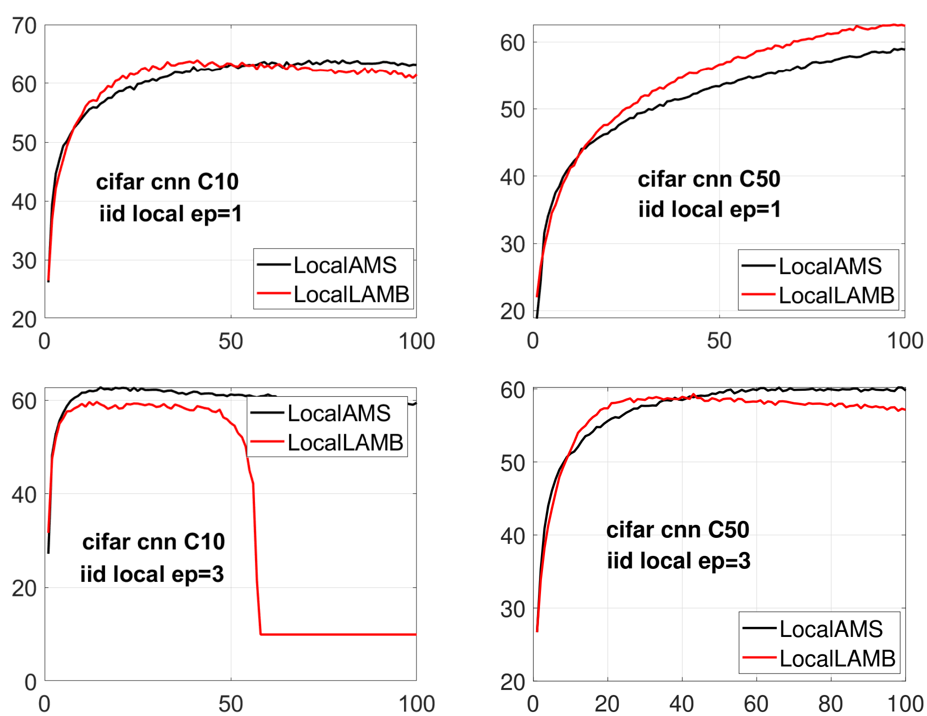

Figure 2: Test accuracy on CNN + CIFAR10. iid data distribution.

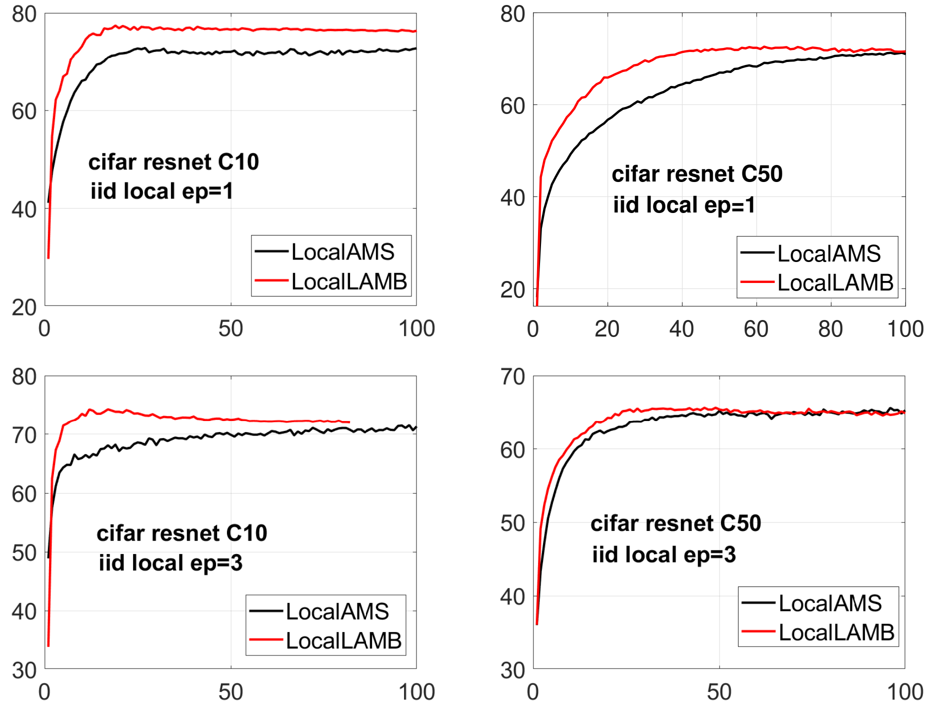

Figure 3: Test accuracy on ResNet + CIFAR10. iid data distribution.

4 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

A Appendix

B Theoretical Analysis

B.1 Intermediary Lemmas

Lemma 1. *Consider $\{\overline{\theta_r}\}_{r>0}$, the sequence of parameters obtained running Algorithm 1. 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 1, 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 1. 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:

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

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

$$\bar{\vartheta}_{r+1}^{\ell} - \bar{\vartheta}_r^{\ell} = \frac{1}{1-\beta_1}(\bar{\theta}_{r+1}^{\ell} - \bar{\theta}_r^{\ell}) - \frac{\beta_1}{1-\beta_1}(\bar{\theta}_r^{\ell} - \bar{\theta}_{r-1}^{\ell}) \tag{10}$$

$$= \frac{\alpha_r}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\|p_{r,i}^{\ell}\|}p_{r,i}^{\ell} - \frac{\alpha_{r-1}}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r-1,i}^{\ell}\|)}{\|p_{r-1,i}^{\ell}\|}p_{r-1,i}^{\ell} \tag{11}$$

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|} - \frac{\phi(\|\theta_{r-1,i}^{\ell}\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^{\ell}\|}\right)m_{r-1}^t + \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}g_{r,i} \tag{12}$$

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

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

$$= \frac{\alpha\beta_1}{1-\beta_1}\frac{1}{n}\sum_{\ell=1}^{\mathsf{L}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\left[\sum_{i=1}^{n}\left(\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|} - \frac{\phi(\|\theta_{r-1,i}^{\ell}\|)}{\sqrt{v_{r-1}^t}\|p_{r-1,i}^{\ell}\|}\right)m_{r-1}^t\right] \tag{15}$$

$$\underbrace{-\frac{\alpha}{n}\sum_{\ell=1}^{\mathsf{L}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}g_{r,i}}_{=A_1} \tag{16}$$

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

$$A_1 \leq -\frac{\alpha}{n}\sum_{\ell=1}^{\mathsf{L}}\sum_{j=1}^{p_\ell}\nabla_\ell f(\bar{\vartheta}_r)^j\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^{\ell}\|)}{\sqrt{v_r^t}\|p_{r,i}^{\ell}\|}g_{r,i} \tag{17}$$

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

$$-\frac{\alpha}{n}\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\left(\phi(\|\theta_{r,i}^{\ell}\|)\nabla_\ell f(\bar{\vartheta}_r)^j\frac{p_{r,i}^{\ell}}{\|p_{r,i}^{\ell}\|}\right)\mathbb{1}\left(\mathrm{sign}(\nabla_\ell f(\bar{\vartheta}_r)^j) \neq \mathrm{sign}(p_{r,i}^{\ell})\right) \tag{19}$$

Taking the expectations on both sides yields:

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

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

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

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

$$\tag{24}$$

where we have used assumption H4.

Since for any ℓ, i, j, we have

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

Then, we obtain

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

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:

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

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

Taking the expectation on both sides leads to:

$$\begin{aligned}
\mathbb{E}[\|\bar{\vartheta}^\ell_{r+1}-\bar{\vartheta}^\ell_r\|^2] &\leq 2\alpha^2\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}-\frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|}\right)m^t_{r-1}\right\|^2\right]+\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}g_{r,i}\right\|^2\right]\\
&\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}-\frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|}\right)m^t_{r-1}\right\|^2\right]\\
&+\frac{1}{n^2}\mathbb{E}\left[\left|\sum_{i=1}^n\sum_{j=1}^p\left\langle\Gamma^j_{r,i}(\nabla f_i(\theta_r)^j+g^j_{r,i}-\nabla f_i(\theta_r)^j)\,|\,\Gamma^j_{r,i}(\nabla f_i(\theta_r)^j+g^j_{r,i}-\nabla f_i(\theta_r)^j)\right\rangle\right|\right]\\
&\leq 2\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}-\frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|}\right)m^t_{r-1}\right\|^2\right]\\
&+\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}\nabla f_i(\theta_r)\right\|^2\right]+\frac{1}{n}\left\|\sum_{i=1}^n\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}\right\|^2\right]
\end{aligned} \tag{31}$$

where the last line uses assumptions H2 and H3 (unbiased gradient and bounded variance of the stochastic gradient) and $\Gamma := \frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}$.

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

$$\begin{aligned}
&\sum_{r=1}^R\mathbb{E}\left[\left\|\frac{\beta_1}{1-\beta_1}\sum_{i=1}^n\left(\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}-\frac{\phi(\|\theta^\ell_{r-1,i}\|)}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|}\right)m^t_{r-1}\right\|^2\right]\\
&\leq\frac{\beta_1^2}{(1-\beta_1)^2}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\frac{1}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}-\frac{1}{\sqrt{v^t_{r-1}}\|p^\ell_{r-1,i}\|}\right)\right\|^2\right]\\
&\leq\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\frac{1}{\sqrt{v^t_r}}-\frac{1}{\sqrt{v^t_{r-1}}}\right)\right\|^2\right]\\
&\leq\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\sum_{r=1}^R\mathbb{E}\left[\left|\sum_{i=1}^n\sum_{j=1}^p\left(\frac{1}{\sqrt{v^{t,j}_r}}-\frac{1}{\sqrt{v^{t,j}_{r-1}}}\right)\right|\right]\\
&\leq\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}
\end{aligned} \tag{32}$$

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

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

$$\begin{aligned}
\sum_{r=1}^R A_r^2 := \sum_{r=1}^R\mathbb{E}[\|\bar{\vartheta}^\ell_{r+1}-\bar{\vartheta}^\ell_r\|^2] &\leq \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}\\
&+\sum_{r=1}^R\left[\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}\nabla f_i(\theta_r)\right\|^2\right]+\frac{1}{n}\left\|\sum_{i=1}^n\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta^\ell_{r,i}\|)}{\sqrt{v^t_r}\|p^\ell_{r,i}\|}\right\|^2\right]\right]
\end{aligned} \tag{33}$$

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:

$$\begin{aligned}
&\sum_{r=1}^{R}\alpha\phi_m\sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2p}}\mathbb{E}[\|\overline{\nabla}f(\vartheta_r)\|^2] - \alpha\phi_M\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\frac{\sigma_i^{\ell,j}}{\sqrt{n}} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}\\
&\leq \sum_{r=1}^{R}\mathbb{E}[f(\bar{\vartheta}_r)-f(\bar{\vartheta}_{r+1})] + \sum_{\ell=1}^{L}\frac{L_\ell}{2}\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}\\
&\quad-\sum_{r=1}^{R}\left[\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right]\right]
\end{aligned} \tag{36}$$

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

$$\begin{aligned}
&\sum_{r=1}^{R}\alpha\phi_m\sqrt{\frac{\mathsf{L}(1-\beta_2)}{M^2p}}\mathbb{E}[\|\overline{\nabla}f(\vartheta_r)\|^2] + \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right]\\
&\leq f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right] + \alpha\phi_M\sum_{\ell=1}^{\mathsf{L}}\sum_{i=1}^{n}\sum_{j=1}^{p_\ell}\frac{\sigma_i^{\ell,j}}{\sqrt{n}} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}\\
&+\sum_{\ell=1}^{L}\frac{L_\ell}{2}\frac{\beta_1^2}{(1-\beta_1)^2}\frac{\mathsf{L}(1-\beta_2)}{p}M^2\phi_M^2\frac{np}{v_0}
\end{aligned} \tag{37}$$

leading to

$$\begin{aligned}
\sum_{r=1}^{R}\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\nabla f_i(\theta_r)\right\|^2\right] &\leq f(\bar{\vartheta}_1)-\mathbb{E}[f(\bar{\vartheta}_{R+1})] + \frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[\frac{\phi(\|\theta_{r,i}^\ell\|)}{\sqrt{v_r^t}\|p_{r,i}^\ell\|}\right\|^2\right]\\
&+\alpha\phi_M\sigma\mathsf{L}p\sqrt{n} + \frac{\overline{L}_\ell\beta_1^2\mathsf{L}(1-\beta_2)M^2\phi_M^2 n}{2(1-\beta_1)^2v_0} + \frac{\alpha\beta_1}{1-\beta_1}\sqrt{(1-\beta_2)p}\frac{\mathsf{L}M^2}{\sqrt{v_0}}
\end{aligned} \tag{38}$$

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

Consider the following inequality:

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

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

$$\left\| \frac{\overline{\nabla} f(\theta_r)}{\sqrt{v_r^t}} \right\| \geq \frac{1}{2} \left\| \frac{\nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| - \left\| \frac{\overline{\nabla} f(\theta_r) - \nabla f(\overline{\theta_r})}{\sqrt{v_r^t}} \right\| \tag{41}$$