Sparsified Distributed Adaptive Learning with Error Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed...

1 Introduction

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 $\mathbb{R}^d$.

Some related work:

[18] develops variant of signSGD (as a biased compression schemes) for distributed optimization. Contributions are mainly on this error feedback variant. In [26], the authors provide theoretical results on the convergence of sparse Gradient SGD for distributed optimization (we want that for AMS here). [27] develops a variant of distributed SGD with sparse gradients too. Contributions include a memory term used while compressing the gradient (using top k for instance). Speeding up the convergence in $\frac{1}{T^3}$.

2 Preliminaries

Sparse Optimization Methods.

Distributed Learning. When a large number of compute engines is available, being able to train global machine learning models while mutualizing the available and *decentralized* source of computation has been a growing focus for the community.

Decentralized optimization methods include methods such as ADMM [6], Distributed Subgradient Descent [24], Dual Averaging [11], Prox-PDA [14], GNSD [21], and Choco-SGD [20].

A recent work [7], which focuses on adaptive gradient methods, namely the Adam [19] annd the AMSGrad [25] optimization methods, develops a decentralized variant of gradient based and adaptive methods in the context of gossip protocols. To date, very few contributions provided attempt to efficiently run adaptive gradient method is such a distributed setting. Apart from [7], **(author?)** [23] proposes a decentralized version of AMSGrad [25] which provably satisfies some non-standard regret. Though, no sparsified variants of them have been proposed for practical purposes nor been studied in the literature.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Compression-Based Distributed Optimization. While the capabilities of the compute powers is exploding, the communication complexity between either the central server and the decentralized workers or among workers is becoming ineffectively large [9, 22]. Gradient sparsification constitutes one popular method to induce sparsity through the optimization procedure and reduce the number of bits transmitted at each iteration. Extensive works have studied this technique to improve the communication efficiency of SGD-based methods such as distributed SGD. This large class of sparsification techniques include gradient quantization leveraging quantized vector of gradients in the communication phase [2, 29, 16, 28, 13, 8, 15], gradient sparsificaiton generally selection top k components of the vector to be communicated, see [27, 1], or variants of the particular SGD algorithm such as low-precision SGD [4, 18] proposing a trade-off between communication cost and precision, and signSGD [10, 30] where only the signs of the gradient vectors are communicated. Most of these works apply to the SGD method [5] as a prototype where a novel method and some convergence results are presented with a rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ where T denotes the total number of iterations, see [3], thus achieving the same rate as plain SGD, see [12, 17].

Yet these communication reduction techniques, still presents a negative dependence on the number of workers, typically a linear dependence. Hence the need for even more efficient techniques which constitutes the object of our paper.

3 Method

Consider standard synchronous distributed optimization setting. AMSGrad is used as the prototype, and the local workers is only in charge of gradient computation.

3.1 TopK AMSGrad with Error Feedback

The key difference (and interesting part) of our TopK AMSGrad compared with the following arxiv paper "Quantized Adam"https://arxiv.org/pdf/2004.14180.pdf is that, in our model only gradients are transmitted. In "QAdam", each local worker keeps a local copy of moment estimator m and v, and compresses and transmits m/v as a whole. Thus, that method is very much like the sparsified distributed SGD, except that g is changed into m/v. In our model, the moment estimates m and v are computed only at the central server, with the compressed gradients instead of the full gradient. This would be the key (and difficulty) in convergence analysis.

Algorithm 1 SPARS-AMS for Distributed Learning

1: **Input**: parameter β_1, β_2, learning rate η_t.
2: Initialize: central server parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$; $e_{0,i} = 0$ the error accumulator for each worker; sparsity parameter k; n local workers; $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$
3: **for** $t = 1$ to T **do**
4:   **parallel for worker** $i \in [n]$ **do**:
5:     Receive model parameter θ_t from central server
6:     Compute stochastic gradient $g_{t,i}$ at θ_t
7:     Compute $\tilde{g}_{t,i} = TopK(g_{t,i} + e_{t,i}, k)$
8:     Update the error $e_{t+1,i} = e_{t,i} + g_{t,i} - \tilde{g}_{t,i}$
9:     Send $\tilde{g}_{t,i}$ back to central server
10:   **end parallel**
11:   **Central server do:**
12:   $\bar{g}_t = \frac{1}{n}\sum_{i=1}^{N} \tilde{g}_{t,i}$
13:   $m_t = \beta_1 m_{t-1} + (1-\beta_1)\bar{g}_t$
14:   $v_t = \beta_2 v_{t-1} + (1-\beta_2)\bar{g}_t^2$
15:   $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
16:   Update global model $\theta_t = \theta_{t-1} - \eta_t \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
17: **end for**

3.2 Convergence Analysis

Several mild assumptions to make: Nonconvex and smooth loss function, unbiased stochastic gradient, bounded variance of the gradient, bounded norm of the gradient, control of the distance between the true gradient and its sparse variant.

Check [7] starting with single machine and extending to distributed settings (several machines).

Under the distributed setting, the goal is to derive an upper bound to the second order moment of the gradient of the objective function at some iteration $T_f \in [1, T]$.

3.3 Mild Assumptions

We begin by making the following assumptions.

A 1. *(Smoothness) For $i \in [\![n]\!]$, f_i is L-smooth: $\|\nabla f_i(\theta) - \nabla f_i(\vartheta)\| \leq L \|\theta - \vartheta\|$.*

A 2. *(Unbiased and Bounded gradient **per worker**) For any iteration index $t > 0$ and worker index $i \in [\![n]\!]$, the stochastic gradient is unbiased and bounded from above: $\mathbb{E}[g_{t,i}] = \nabla f_i(\theta_t)$ and $\|g_{t,i}\| \leq G_i$.*

A 3. *(Bounded variance **per worker**) For any iteration index $t > 0$ and worker index $i \in [\![n]\!]$, the variance of the noisy gradient is bounded: $\mathbb{E}[|g_{t,i} - \nabla f_i(\theta_t)|^2] < \sigma_i^2$.*

Denote by $Q(\cdot)$ the quantization operator Line 7 of Algorithm 1, which takes as input a gradient vector and returns a quantized version of it, and note $\tilde{g} := Q(g)$. Assume that

A 4. *(Bounded Quantization) For any iteration $t > 0$, there exists a constant $q > 0$ such that $\|g_{t,i} - \tilde{g}_{t,i}\| \leq q \|g_{t,i}\|$, where $g_{t,i}$ is the stochastic gradient computed at iteration t for worker i.*

Denote for all $\theta \in \Theta$:

$$f(\theta) := \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \, , \tag{2}$$

where n denotes the number of workers.

3.4 Intermediary Lemmas

Lemma 1. *Under Assumption 2 and Assumption 4 we have for any iteration $t > 0$:*

$$\|m_t\|^2 \leq (q^2 + 1) G^2 \quad \textit{and} \quad \hat{v}_t \leq (q^2 + 1) G^2 \tag{3}$$

where m_t and $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$ are defined Line 15 of Algorithm 1 and $G^2 = \frac{1}{n}\sum_{i=1}^{N} G_i^2$.

Lemma 2. *Under A1 to A4, with a decreasing sequence of stepsize $\{\eta_t\}_{t>0}$, we have:*

$$-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I}_\mathsf{d})^{-1/2} \bar{g}_t \right\rangle] \leq -\frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2 \frac{G^2 \eta_{t+1}}{\epsilon 2 n^2} \tag{4}$$

*where I_d is the identity matrix, $\hat{V}_t$ the diagonal matrix which diagonal entries are $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$ defined Line 15 of Algorithm 1 and $\bar{g}_t$ is the aggregation of all **quantized** gradients from the workers.*

Lemma 3. *Under A1 to A4, with a decreasing sequence of stepsize $\{\eta_t\}_{t>0}$, we have:*

$$\begin{aligned}
\mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] \leq & -\frac{\eta_{t+1}(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2 \frac{G^2 \eta_{t+1}}{\epsilon 2 n^2} \\
& - \eta_{t+1}\beta_1 \mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I}_\mathsf{d})^{-1/2} m_t \right\rangle] \\
& + \left(\frac{L}{2} + \beta_1 L\right) \|\theta_t - \theta_{t-1}\|^2 \\
& + \eta_{t+1} G^2 \mathbb{E}[\sum_{j=1}^{d} \left[(\hat{v}_{t+1}^j + \epsilon)^{-1/2} - (\hat{v}_t^j + \epsilon)^{-1/2} \right]]
\end{aligned} \tag{5}$$

where d denotes the dimension of the parameter vector

The main theorem in the decentralized setting reads:

Theorem 1. *Under A1 to A4, with a constant stepsize $\eta_t = \eta = \frac{L}{\sqrt{T_m}}$, we have:*

$$\frac{1}{T_m}\sum_{t=0}^{T_m-1}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq \frac{\mathbb{E}[f(\theta_0)-f(\theta_{T_m})]}{L\Delta_1\sqrt{T_m}} + d\frac{L\Delta_3}{\Delta_1\sqrt{T_m}} + \frac{\Delta_2}{\eta\Delta_1 T_m} + \frac{1-\beta_1}{\Delta_1}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \tag{6}$$

where

$$\begin{aligned}\Delta_1 &:= \frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \quad , \quad \Delta_2 := q^2 + \sum_{k=t+1}^{\infty}\beta_1^{k-t+2}\frac{G^2}{\epsilon 2n^2} \\ \Delta_3 &:= \left(\frac{L}{2} + 1 + \frac{\beta_1 L}{1-\beta_1}\right)(1-\beta_2)^{-1}(1-\frac{\beta_1^2}{\beta_2})^{-1}\end{aligned} \tag{7}$$

4 Sequential Model

Single machine method

Algorithm 2 SPARS-AMS : Single machine setting

1: **Input**: parameter β_1, β_2, learning rate η_t.
2: Initialize: central server parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$; $e_0 = 0$ the error accumulator; sparsity parameter k; $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$
3: **for** $t = 1$ to T **do**
4: Compute stochastic gradient $g_t = g_{t,i_t}$ at θ_t for randomly sampled index i_t
5: Compute $\tilde{g}_t = TopK(g_t + e_t, k)$
6: Update the error $e_{t+1} = e_t + g_t - \tilde{g}_t$
7: $m_t = \beta_1 m_{t-1} + (1-\beta_1)\tilde{g}_t$
8: $v_t = \beta_2 v_{t-1} + (1-\beta_2)\tilde{g}_t^2$
9: $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
10: Update global model $\theta_t = \theta_{t-1} - \eta_t \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
11: **end for**

Let m_t' and $\hat{v}_t'$ be the first and second moment moving average of standard AMSGrad using full gradients. Denote

$$a_t = \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}, \quad a_t' = \frac{m_t'}{\sqrt{\hat{v}_t' + \epsilon}}.$$

Define the sequence

$$\mathcal{E}_{t+1} = \mathcal{E}_t + a_t' - a_t,$$

such that the auxiliary model

$$\begin{aligned}\theta_{t+1}' &:= \theta_{t+1} - \eta\mathcal{E}_{t+1} \\ &= \theta_t - \eta a_t - \eta\mathcal{E}_{t+1} \\ &= \theta_t - \eta a_t - \eta(\mathcal{E}_t + a_t' - a_t) \\ &= \theta_t' - \eta a_t'\end{aligned}$$

follows the update of full-gradient AMSGrad. By smoothness assumption we have

$$f(\theta_{t+1}') \leq f(\theta_t') - \eta\langle\nabla f(\theta_t'), a_t'\rangle + \frac{L}{2}\|\theta_{t+1}' - \theta_t'\|^2.$$

Thus,

$$
\begin{aligned}
\mathbb{E}[f(\theta'_{t+1}) - f(\theta'_t)] &\leq -\eta\mathbb{E}[\langle \nabla f(\theta'_t), a'_t\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|a'_t\|^2] \\
&= -\eta\mathbb{E}[\langle \nabla f(\theta'_t), a'_t\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|a'_t\|^2] + \eta\mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(\theta'_t), a'_t\rangle] \\
&\leq -\eta\mathbb{E}[\langle \nabla f(\theta'_t), a'_t\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\|a'_t\|^2] + \eta\mathbb{E}[\frac{\eta^2\rho}{2}\|\mathcal{E}_t\|^2 + \frac{1}{2\rho}\|a'_t\|^2] \\
&\leq -\eta\frac{\mathbb{E}\|\nabla f(\theta_t)\|^2}{\sqrt{G^2+\epsilon}} + \frac{\eta}{2\rho}\frac{\mathbb{E}\|\nabla f(\theta_t)\|^2}{\epsilon} + \frac{\eta^2 L}{2}\mathbb{E}[\|a'_t\|^2] + \frac{\eta^3\rho}{2}\mathbb{E}\|\mathcal{E}_t\|^2,
\end{aligned}
$$

when $\beta_1 = 0$ for example. We may discard this assumption and use more complicated bound on the first two terms. The third term can be bounded by constant yielding $O(1/\sqrt{T})$ rate eventually when taking decreasing learning rate. The key is to get a good bound on the cumulative error sequence, $\mathcal{E}_t$. We have the following:

$$
\begin{aligned}
\mathbb{E}\|\mathcal{E}_{t+1}\|^2 &= \mathbb{E}\|\mathcal{E}_t + a'_t - a_t + TopK(\mathcal{E}_t + a'_t) - TopK(\mathcal{E}_t + a'_t)\|^2 \\
&\leq 2\mathbb{E}\|\mathcal{E}_t + a'_t - TopK(\mathcal{E}_t + a'_t)\|^2 + 2\mathbb{E}\|a_t - TopK(\mathcal{E}_t + a'_t)\|^2 \\
&\leq 2q\mathbb{E}\|\mathcal{E}_t + a'_t\| + 2\mathbb{E}\|a_t - TopK(\mathcal{E}_t + a'_t)\|^2 \\
&\leq 2q[(1+r)\mathbb{E}\|\mathcal{E}_t\|^2 + (1+\frac{1}{r})\mathbb{E}\|a'_t\|^2] + 2\mathbb{E}\|a_t - TopK(\mathcal{E}_t + a'_t)\|^2.
\end{aligned}
$$

Current try: If we can bound the last term in the same form as the first two terms, then we can use recursion to get the desired result. We can have

$$
\mathbb{E}\|a_t - TopK(\mathcal{E}_t + a'_t)\|^2 = \mathbb{E}\|\frac{\tilde{m}_t}{\sqrt{\hat{v}_t+\epsilon}} - \|^2
$$

5 Experiments

Our proposed TopK-EF with AMSGrad matches that of full AMSGrad, in distributed learning. Number of local workers is 20. Error feedback fixes the convergence issue of using solely the TopK gradient.

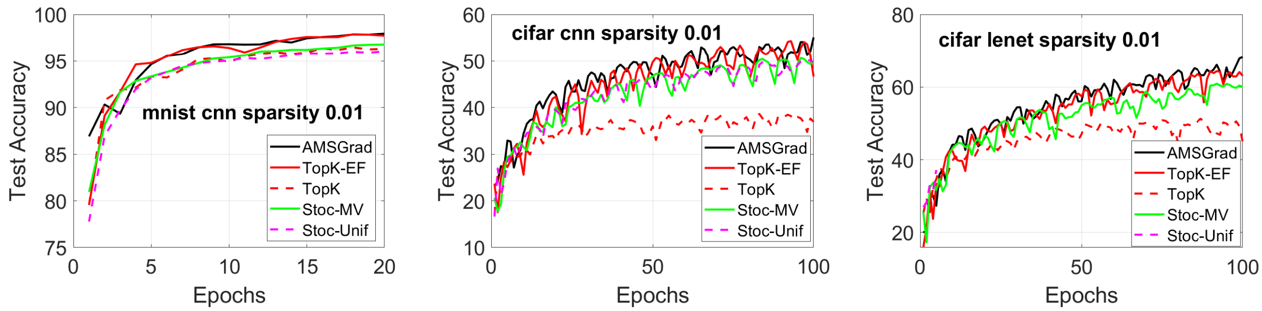

Figure 1: Test accuracy.

6 Conclusion

References

[1] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

[2] 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.

[3] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. *arXiv preprint arXiv:1809.10505*, 2018.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[5] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[7] Congliang Chen, Li Shen, Haozhi Huang, Qi Wu, and Wei Liu. Quantized adam with error feedback. *arXiv preprint arXiv:2004.14180*, 2020.

[8] Yongjian Chen, Tao Guan, and Cheng Wang. Approximate nearest neighbor search by residual vector quantization. *Sensors*, 10(12):11259–11273, 2010.

[9] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *Symposium on Operating Systems Design and Implementation*, pages 571–582, 2014.

[10] Christopher De Sa, Matthew Feldman, Christopher Ré, and Kunle Olukotun. Understanding and optimizing asynchronous low-precision stochastic gradient descent. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 561–574, 2017.

[11] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[12] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[13] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Trading redundancy for communication: Speeding up distributed sgd for non-convex optimization. In *International Conference on Machine Learning*, pages 2545–2554. PMLR, 2019.

[14] Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pages 1529–1538, 2017.

[15] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

[16] Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2530–2541, 2018.

[17] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

[18] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487, 2019.

[21] Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321, 2019.

[22] 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.

[23] Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *arXiv preprint arXiv:1901.09109*, 2019.

[24] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48, 2009.

[25] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[26] Shaohuai Shi, Kaiyong Zhao, Qiang Wang, Zhenheng Tang, and Xiaowen Chu. A convergence analysis of distributed sgd with communication-efficient gradient sparsification. In *IJCAI*, pages 3411–3417, 2019.

[27] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[28] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. *arXiv preprint arXiv:1710.09854*, 2017.

[29] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *arXiv preprint arXiv:1705.07878*, 2017.

[30] Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. Swalp: Stochastic weight averaging in low precision training. In *International Conference on Machine Learning*, pages 7015–7024. PMLR, 2019.

A Appendix

B Proofs

B.1 Proof of Lemmas

Lemma. *Under Assumption 2 and Assumption 4 we have for any iteration $t > 0$:*

$$\|m_t\|^2 \leq (q^2+1)G^2 \quad \textit{and} \quad \hat{v}_t \leq (q^2+1)G^2 \tag{8}$$

where m_t and $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$ are defined Line 15 of Algorithm 1 and $G^2 = \frac{1}{n}\sum_{i=1}^N G_i^2$.

Proof. We start by writing

$$\|\bar{g}_t\|^2 = \left\| \frac{1}{n}\sum_{i=1}^N \tilde{g}_{t,i} \right\|^2 \leq \frac{1}{n}\sum_{i=1}^N \|\tilde{g}_{t,i}\|^2 \tag{9}$$

Though, using Assumption 2 and Assumption 4 we have:

$$\|\tilde{g}_{t,i}\|^2 = \|g_{t,i} + \tilde{g}_{t,i} - g_{t,i}\|^2 \leq \|g_{t,i}\|^2 + \|\tilde{g}_{t,i} - g_{t,i}\|^2 \leq (q^2+1)G_i^2 \tag{10}$$

Hence

$$\|\bar{g}_t\|^2 \leq (q^2+1)G^2 \tag{11}$$

where $G^2 = \frac{1}{n}\sum_{i=1}^N G_i^2$. Then, by construction in Algorithm 1:

$$\|m_t\|^2 \leq \beta_1^2 \|m_{t-1}\|^2 + (1-\beta_1)^2 \|\bar{g}_t\|^2 \leq \beta_1^2 \|m_{t-1}\|^2 + (1-\beta_1)^2 (q^2+1)G^2 \tag{12}$$

Since we have by initialization that $\|m_0\|^2 \leq G^2$, then we prove by induction that $\|m_t\|^2 \leq (q^2+1)G^2$.

Similarly

$$\hat{v}_t = \max(v_t, \hat{v}_{t-1}) = \max(\hat{v}_{t-1}, \beta_2 v_{t-1} + (1-\beta_2)\bar{g}_t^2) \leq max(\hat{v}_{t-1}, \beta_2 v_{t-1} + (1-\beta_2)(q^2+1)G^2) \tag{13}$$

□

Lemma. *Under A1 to A4, with a decreasing sequence of stepsize $\{\eta_t\}_{t>0}$, we have:*

$$-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I}_\mathsf{d})^{-1/2}\bar{g}_t \right\rangle] \leq -\frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \tag{14}$$

where I_d is the identity matrix, $\hat{V}_t$ the diagonal matrix which diagonal entries are $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$ defined Line 15 of Algorithm 1 and $\bar{g}_t$ is the aggregation of all ***quantized*** *gradients from the workers.*

Proof. We first decompose $\bar{g}_t$ as the sum of the unbiased stochastic gradients and its quantized versions as computed Line 7 of Algorithm 1:

$$\bar{g}_t = \frac{1}{n}\sum_{i=1}^N \tilde{g}_{t,i} = \frac{1}{n}\sum_{i=1}^N [g_{t,i} + \tilde{g}_{t,i} - g_{t,i}] \tag{15}$$

Hence,

$$\begin{aligned} T_1 :=& -\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I}_\mathsf{d})^{-1/2}\bar{g}_t \right\rangle] \\ =& \underbrace{-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I}_\mathsf{d})^{-1/2}\frac{1}{n}\sum_{i=1}^N g_{t,i} \right\rangle]}_{t1} \underbrace{-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I}_\mathsf{d})^{-1/2}\frac{1}{n}\sum_{i=1}^N \tilde{g}_{t,i} - g_{t,i} \right\rangle]}_{t_2} \end{aligned} \tag{16}$$

Bounding t_1: Using the Tower rule, we have:

$$\begin{aligned} t_1 :=& -\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\frac{1}{n}\sum_{i=1}^{N} g_{t,i} \right\rangle] \\ =& -\eta_{t+1}\mathbb{E}[\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\frac{1}{n}\sum_{i=1}^{N} g_{t,i} \right\rangle |\mathcal{F}_t]] \\ =& -\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\mathbb{E}[\frac{1}{n}\sum_{i=1}^{N} g_{t,i} \right\rangle |\mathcal{F}_t]] \end{aligned} \tag{17}$$

Using Assumption 2 and Lemma 1, we have that

$$\begin{aligned} t_1 :=& -\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\frac{1}{n}\sum_{i=1}^{N} g_{t,i} \right\rangle] \\ \leq& -\eta_{t+1}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \end{aligned} \tag{18}$$

Bounding t_2:

We first recall Young's inequality with a constant $\delta \in (0,1)$ as follows:

$$\langle X \,|\, Y \rangle \leq \frac{1}{\delta}\|X\|^2 + \delta\|Y\|^2 \, . \tag{19}$$

Using Young's inequality (19) with parameter equal to 1:

$$\begin{aligned} t_2 \leq& \frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta_{t+1}}{2n^2}\mathbb{E}[\|(\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\sum_{i=1}^{N}\{\tilde{g}_{t,i} - g_{t,i}\}\|^2] \\ \overset{(a)}{\leq}& \frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta_{t+1}}{2n^2}\mathbb{E}[\|(\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\|^2\sum_{i=1}^{N}\{\tilde{g}_{t,i} - g_{t,i}\}\|^2] \\ \overset{(b)}{\leq}& \frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta_{t+1}}{2n^2}\mathbb{E}[\|(\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\|^2]\mathbb{E}[\|\sum_{i=1}^{N}\{\tilde{g}_{t,i} - g_{t,i}\}\|^2] \\ \overset{(c)}{\leq}& \frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta_{t+1}}{\epsilon 2n^2}\mathbb{E}[\|\sum_{i=1}^{N}\tilde{g}_{t,i} - g_{t,i}\|^2] \\ \overset{(d)}{\leq}& \frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \end{aligned} \tag{20}$$

where (a) uses the Cauchy-Schwartz inequality, (b) is due to the non-negativeness of both $\hat{V}_{t+1}$ and $\|\sum_{i=1}^{N}\{g_{t,i} + \tilde{g}_{t,i} - g_{t,i}\}\|^2$ and (c) uses the Triangle inequality. We use Assumption 3 and Assumption 4 in (d).

Finally, combining (18) and (20) yields

$$-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I}_\mathsf{d})^{-1/2}\bar{g}_t \right\rangle] \leq -\frac{\eta_{t+1}}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \tag{21}$$

$\square$

Lemma. *Under A1 to A4, with a decreasing sequence of stepsize $\{\eta_t\}_{t>0}$, we have:*

$$
\begin{aligned}
\mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] \leq & -\frac{\eta_{t+1}(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \\
& - \eta_{t+1}\beta_1\mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \\
& + \left(\frac{L}{2} + \beta_1 L\right)\|\theta_t - \theta_{t-1}\|^2 \\
& + \eta_{t+1} G^2 \mathbb{E}[\sum_{j=1}^{d}\left[(\hat{v}_{t+1}^j + \epsilon)^{-1/2} - (\hat{v}_t^j + \epsilon)^{-1/2}\right]]
\end{aligned} \tag{22}
$$

where d denotes the dimension of the parameter vector

Proof. Denote the following auxiliary variables at iteration $t+1$

$$
z_{t+1} = \theta_{t+1} + \frac{\beta_1}{1-\beta_1}(\theta_{t+1} - \theta_t) \tag{23}
$$

By assumption Assumption 1, we can write the smoothness condition on the overall objective (2), between iteration t and $t+1$:

$$
f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t) \,|\, \theta_{t+1} - \theta_t \rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \tag{24}
$$

Denote by $\hat{V}_t$ the diagonal matrix which diagonal entries are $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$ defined Line 15 of Algorithm 1. Hence, we obtain,

$$
f(\theta_{t+1}) \leq f(\theta_t) - \eta_{t+1}\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \tag{25}
$$

where $\mathsf{I_d}$ denotes the identity matrix.

We now take the expectation of those various terms conditioned on the filtration $\mathcal{F}_t$ of the total randomness up to iteration t.

$$
\mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] \leq -\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] + \frac{L}{2}\mathbb{E}[\|\theta_{t+1} - \theta_t\|^2] \tag{26}
$$

We now focus on the computation of the inner product obtained in the equation above. We have

$$
\begin{aligned}
& \eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] \qquad (27) \\
= & \eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} + (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} - (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] \\
= & \eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] + \eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, \left[(\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} - (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2}\right] m_{t+1} \right\rangle] \\
= & \eta_{t+1}\beta_1\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] + \eta_{t+1}(1-\beta_1)\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} \bar{g}_t \right\rangle] \\
& + \eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, \left[(\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} - (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2}\right] m_{t+1} \right\rangle] \qquad (28)
\end{aligned}
$$

where $\bar{g}_t$ is the aggregated gradients from all workers.

Plugging the above in (26) yields:

$$
\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] \\
\leq & \underbrace{-\beta_1 \mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle]}_{A_t} \eta_{t+1} \\
& \underbrace{-\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, \left[(\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} - (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} \right] m_{t+1} \right\rangle]}_{B_t} \eta_{t+1} \\
& \underbrace{-(1-\beta_1)\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} \bar{g}_t \right\rangle]}_{C_t} \eta_{t+1} + \frac{L}{2}\mathbb{E}[\|\theta_{t+1} - \theta_t\|^2]
\end{aligned}
\tag{29}
$$

To begin with, by the tower rule, we have that

$$
A_t = -\beta_1 \mathbb{E}[\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle |\mathcal{F}_t]] \tag{30}
$$

$$
= -\beta_1 \left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] - \beta_1 \left\langle \nabla f(\theta_t) - \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \tag{31}
$$

$$
\tag{32}
$$

where we recognize the first term as the term in (27), at iteration $t-1$ and hence apply the same decomposition as in (28). Coupling with the smoothness of f, which gives that

$$
-\beta_1 \left\langle \nabla f(\theta_t) - \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \leq \frac{\beta_1 L}{\eta_{t-1}} \|\theta_t - \theta_{t-1}\|^2
$$

we obtain,

$$
\begin{aligned}
A_t &= -\beta_1 \mathbb{E}[\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle |\mathcal{F}_t]] \\
&\leq \eta_{t+1}\beta_1 (A_{t-1} + B_{t-1} + C_{t-1}) + \eta_{t+1} \frac{\beta_1 L}{\eta_{t-1}} \|\theta_t - \theta_{t-1}\|^2
\end{aligned}
\tag{33}
$$

Then,

$$
\begin{aligned}
B_t &= -\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, \left[(\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} - (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} \right] m_{t+1} \right\rangle] \\
&= \mathbb{E}[\sum_{j=1}^{d} \nabla^j f(\theta_t) m^j_{t=1} \left[(\hat{v}^j_{t+1} + \epsilon)^{-1/2} - (\hat{v}^j_t + \epsilon)^{-1/2} \right]] \\
&\overset{(a)}{\leq} \mathbb{E}[\|\nabla f(\theta_t)\| \, \|m_{t=1}\| \sum_{j=1}^{d} \left[(\hat{v}^j_{t+1} + \epsilon)^{-1/2} - (\hat{v}^j_t + \epsilon)^{-1/2} \right]] \\
&\overset{(b)}{\leq} G^2 \mathbb{E}[\sum_{j=1}^{d} \left[(\hat{v}^j_{t+1} + \epsilon)^{-1/2} - (\hat{v}^j_t + \epsilon)^{-1/2} \right]]
\end{aligned}
\tag{34}
$$

where $\nabla^j f(\theta_t)$ denotes the j-th component of the gradient vector $\nabla f(\theta_t)$, (a) uses of the Cauchy-Schwartz inequality and (b) boils down from the norm of the gradient vector boundedness assumption 2, denoting $G := \frac{1}{n}\sum_{i=1}^{n} G_i$.

Plugging the above into (29) yields

$$
\begin{aligned}
\mathbb{E}[f(\theta_{t+1})-f(\theta_t)] \leq & \eta_{t+1}(A_t+B_t+C_t)+\frac{L}{2}\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2] \\
\leq & -\eta_{t+1}\beta_1\mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t+\epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \\
& +\eta_{t+1}G^2\mathbb{E}[\sum_{j=1}^{d}\left[(\hat{v}_{t+1}^j+\epsilon)^{-1/2}-(\hat{v}_t^j+\epsilon)^{-1/2}\right]] \\
& +\left(\frac{L}{2}+\eta_{t+1}\frac{\beta_1 L}{\eta_{t-1}}\right)\|\theta_t-\theta_{t-1}\|^2 \\
& -\eta_{t+1}(1-\beta_1)\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I_d})^{-1/2}\bar{g}_t \right\rangle]
\end{aligned}
\tag{35}
$$

We bound the last term on the RHS, $-\eta_{t+1}\mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1}+\epsilon \mathsf{I_d})^{-1/2}\bar{g}_t \right\rangle]$ with Lemma 2

Under the assumption that we use a decreasing stepsize such that $\eta_{t+1} \leq \eta_t$, and given that according to Line 15 we have that $\hat{v}_{t+1} \geq \hat{v}_t$ by construction, we obtain

$$
\begin{aligned}
\mathbb{E}[f(\theta_{t+1})-f(\theta_t)] \leq & -\frac{\eta_{t+1}(1-\beta_1)}{2}(\epsilon+\frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2]+q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \\
& -\eta_{t+1}\beta_1\mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t+\epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \\
& +\left(\frac{L}{2}+\beta_1 L\right)\|\theta_t-\theta_{t-1}\|^2 \\
& +\eta_{t+1}G^2\mathbb{E}[\sum_{j=1}^{d}\left[(\hat{v}_{t+1}^j+\epsilon)^{-1/2}-(\hat{v}_t^j+\epsilon)^{-1/2}\right]]
\end{aligned}
\tag{36}
$$

Finally, using Lemma 2, we obtain the desired result. □

B.2 Proof of Theorem 1

Theorem. *Under A1 to A4, with a constant stepsize $\eta_t = \eta = \frac{L}{\sqrt{T_m}}$, we have:*

$$
\frac{1}{T_m}\sum_{t=0}^{T_m-1}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq \frac{\mathbb{E}[f(\theta_0)-f(\theta_{T_m})]}{L\Delta_1\sqrt{T_m}}+d\frac{L\Delta_3}{\Delta_1\sqrt{T_m}}+\frac{\Delta_2}{\eta\Delta_1 T_m}+\frac{1-\beta_1}{\Delta_1}(\epsilon+\frac{G^2}{1-\beta_2})^{-\frac{1}{2}}
\tag{37}
$$

where

$$
\begin{aligned}
&\Delta_1 := \frac{(1-\beta_1)}{2}(\epsilon+\frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \quad , \quad \Delta_2 := q^2+\sum_{k=t+1}^{\infty}\beta_1^{k-t+2}\frac{G^2}{\epsilon 2n^2} \\
&\Delta_3 := \left(\frac{L}{2}+1+\frac{\beta_1 L}{1-\beta_1}\right)(1-\beta_2)^{-1}(1-\frac{\beta_1^2}{\beta_2})^{-1}
\end{aligned}
\tag{38}
$$

Proof. By Lemma 3 we have

$$
\begin{aligned}
\mathbb{E}[f(\theta_{t+1})-f(\theta_t)] \leq & -\frac{\eta_{t+1}(1-\beta_1)}{2}(\epsilon+\frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2]+q^2\frac{G^2\eta_{t+1}}{\epsilon 2n^2} \\
& -\eta_{t+1}\beta_1\mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t+\epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \\
& +\left(\frac{L}{2}+\beta_1 L\right)\|\theta_t-\theta_{t-1}\|^2 \\
& +\eta_{t+1}G^2\mathbb{E}[\sum_{j=1}^{d}\left[(\hat{v}_{t+1}^j+\epsilon)^{-1/2}-(\hat{v}_t^j+\epsilon)^{-1/2}\right]]
\end{aligned}
\tag{39}
$$

Let us consider the following sequence, defined for all $t > 0$:

$$R_t := f(\theta_t) - \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \tag{40}$$

We compute the following expectation:

$$\begin{aligned}\mathbb{E}[R_{t+1}] - \mathbb{E}[R_t] =& \mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] - \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2} \mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] \\ &+ \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle]\end{aligned} \tag{41}$$

Using the Assumption 1, we note that:

$$\mathbb{E}[f(\theta_{t+1}) - f(\theta_t)] \leq -\eta_{t+1} \mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] + \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \tag{42}$$

which yields

$$\begin{aligned}\mathbb{E}[R_{t+1}] - \mathbb{E}[R_t] =& -(\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) \mathbb{E}[\left\langle \nabla f(\theta_t) \,|\, (\hat{V}_{t+1} + \epsilon \mathsf{I_d})^{-1/2} m_{t+1} \right\rangle] \\ &+ \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] \\ &+ \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \\ \leq& (\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) \mathbb{E}[A_t + B_t + C_t] \\ &- \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \mathbb{E}[A_{t-1} + B_{t-1} + C_{t-1}] \\ &+ \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2\end{aligned} \tag{43}$$

where A_t, B_t, C_t are defined in (29).

We use (33) and (34) to bound A_t and B_t, and Lemma 2 to bound C_t where we precise that the learning rate η_{t+1} becomes $\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}$. Hence

$$\begin{aligned}\mathbb{E}[R_{t+1}] - \mathbb{E}[R_t] \leq& \left((\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) \beta_1 - \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \right) \mathbb{E}[A_{t-1} + B_{t-1} + C_{t-1}] \\ &+ (\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) G^2 \mathbb{E}[\sum_{j=1}^{d} \left[(\hat{v}_{t+1}^j + \epsilon)^{-1/2} - (\hat{v}_t^j + \epsilon)^{-1/2} \right]] \\ &+ \left(\frac{L}{2} + (\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) \frac{\beta_1 L}{\eta_{t-1}} \right) \|\theta_{t+1} - \theta_t\|^2 \\ &- (\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}) \frac{(1-\beta_1)}{2} (\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\ &+ q^2 \eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2} \frac{G^2}{\epsilon 2 n^2}\end{aligned} \tag{44}$$

where the last term in the LHS is due to Lemma 3.

By assumption, we have that for all $t > 0$, $\eta_{t=1} \leq \eta_t$. Also, set the tuning parameters such that

$$\eta_t + \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \leq \frac{\eta_t}{1-\beta_1} \tag{45}$$

so that

$$\begin{aligned} &(\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2})\beta_1 - \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} = 0 \\ \iff &(\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2})\beta_1 = \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1} \end{aligned} \tag{46}$$

Note that $-(\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2})\frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \leq -\eta_{t+1}\frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}$ since $\sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2} \geq 0$.

The above coupled with (44) yields

$$\begin{aligned} \mathbb{E}[R_{t+1}] - \mathbb{E}[R_t] \leq &- \eta_{t+1}\frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + q^2\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2}\frac{G^2}{\epsilon 2n^2} \\ &- (\eta_{t+1} + \sum_{k=t+1}^{\infty} \eta_k \beta_1^{k-t+2})G^2\mathbb{E}[\sum_{j=1}^{d}\left[(\hat{v}_t^j + \epsilon)^{-1/2} - (\hat{v}_{t+1}^j + \epsilon)^{-1/2}\right]] \\ &+ \left(\frac{L}{2} + 1 + \frac{\beta_1 L}{1-\beta_1}\right)\mathbb{E}[\|\theta_{t+1} - \theta_t\|^2] \end{aligned} \tag{47}$$

We now sum from $t = 0$ to $t = T_{\text{m}} - 1$ the inequality in (47), and divide it by T_{m}:

$$\begin{aligned} &\eta\frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\frac{1}{T_{\text{m}}}\sum_{t=0}^{T_{\text{m}}-1}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \\ \leq &\frac{\mathbb{E}[R_0] - \mathbb{E}[R_{T_{\text{m}}}]}{T_{\text{m}}} + \frac{q^2\eta + \sum_{k=t+1}^{\infty} \eta \beta_1^{k-t+2}\frac{G^2}{\epsilon 2n^2}}{T_{\text{m}}} \\ &+ \left(\frac{L}{2} + 1 + \frac{\beta_1 L}{1-\beta_1}\right)\frac{1}{T_{\text{m}}}\sum_{t=0}^{T_{\text{m}}-1}\mathbb{E}[\|\theta_{t+1} - \theta_t\|^2] \end{aligned} \tag{48}$$

where we have used the fact that $(\hat{v}_t^j + \epsilon)^{-1/2} - (\hat{v}_{t+1}^j + \epsilon)^{-1/2} \geq 0$ for all dimension $j \in [d]$ by construction of $\hat{v}_{t+1}^j$.

We now bound the two remaining terms:

Bounding $-\mathbb{E}[R_{T_{\text{m}}}]$:

By definition (40) of R_t we have, using Lemma 1:

$$\begin{aligned} -\mathbb{E}[R_{T_{\text{m}}}] &\leq \sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1}\mathbb{E}[\left\langle \nabla f(\theta_{t-1}) \,|\, (\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t \right\rangle] - f(\theta_{T_{\text{m}}}) \\ &\leq \|\sum_{k=t}^{\infty} \eta_k \beta_1^{k-t+1}\|\|\nabla f(\theta_{t-1})\|\|(\hat{V}_t + \epsilon \mathsf{I_d})^{-1/2} m_t\| \\ &\leq \eta_{t+1}(1-\beta_1)(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} - f(\theta_{T_{\text{m}}}) \end{aligned} \tag{49}$$

Bounding $\sum_{t=0}^{T_{\mathsf{m}}-1} \mathbb{E}[\|\theta_{t+1}-\theta_t\|^2]$:

By definition in Algorithm 1:

$$\|\theta_{t+1}-\theta_t\|^2 = \eta_{t+1}^2 \left[(\hat{V}_{t+1}+\epsilon \mathsf{I}_{\mathsf{d}})^{-\frac{1}{2}} m_{t+1}\right]^2 = \eta_{t+1}^2 \sum_{j=1}^{d} \frac{|m_{t+1}^j|^2}{\hat{v}_{t+1}^j+\epsilon} \tag{50}$$

For any dimension $j \in [d]$,

$$\begin{aligned} |m_{t+1}^j|^2 &= |\beta_1 m_t^j + (1-\beta_1)\bar{g}_t^j|^2 \\ &\leq \beta_1(\beta_1^2 |m_{t-1}^j|^2 + (1-\beta_1)^2 |\bar{g}_{t-1}^j|^2) + |\bar{g}_t^j|^2 \\ &\leq \sum_{k=0}^{t} \beta_1^{2(t-k)} |\bar{g}_k^j|^2 \\ &\leq \sum_{k=0}^{t} \frac{\beta_1^{2(t-k)}}{\beta_2^{t-k}} \beta_2^{t-k} |\bar{g}_k^j|^2 \end{aligned} \tag{51}$$

Using Cauchy-Schwartz inequality we obtain

$$\begin{aligned} |m_{t+1}^j|^2 \leq \sum_{k=0}^{t} \frac{\beta_1^{2(t-k)}}{\beta_2^{t-k}} \beta_2^{t-k} |\bar{g}_k^j|^2 &\leq \sum_{k=0}^{t} \left(\frac{\beta_1^2}{\beta_2}\right)^{t-k} \sum_{k=0}^{t} \beta_2^{t-k} |\bar{g}_k^j|^2 \\ &\leq \frac{1}{1-\frac{\beta_1^2}{\beta_2}} \sum_{k=0}^{t} \beta_2^{t-k} |\bar{g}_k^j|^2 \end{aligned} \tag{52}$$

On the other hand we have

$$\hat{v}_{t+1}^j \geq \beta_2 \hat{v}_t^j + (1-\beta_2)(\bar{g}_t^j)^2 \tag{53}$$

and since it is also true for iteration $t=1$, we have by induction replacing v_t^j in the above that

$$\hat{v}_{t+1}^j \geq (1-\beta_2) \sum_{k=0}^{t} \beta_2^{t-k} |\bar{g}_k^j|^2 \iff \frac{\sum_{k=0}^{t} \beta_2^{t-k} |\bar{g}_k^j|^2}{\hat{v}_{t+1}^j} \leq (1-\beta_2)^{-1} \tag{54}$$

Hence, we can derive from (50) that

$$\begin{aligned} \|\theta_{t+1}-\theta_t\|^2 = \eta_{t+1}^2 \sum_{j=1}^{d} \frac{|m_{t+1}^j|^2}{\hat{v}_{t+1}^j+\epsilon} &\leq \eta_{t+1}^2 \sum_{j=1}^{d} \frac{|m_{t+1}^j|^2}{\hat{v}_{t+1}^j} \\ &\overset{(a)}{\leq} \eta_{t+1}^2 \sum_{j=1}^{d} \frac{1}{1-\frac{\beta_1^2}{\beta_2}} \frac{\sum_{k=0}^{t} \beta_2^{t-k} |\bar{g}_k^j|^2}{\hat{v}_{t+1}^j} \\ &\overset{(b)}{\leq} \eta_{t+1}^2 d (1-\beta_2)^{-1} (1-\frac{\beta_1^2}{\beta_2})^{-1} \end{aligned} \tag{55}$$

where (a) uses (52) and (b) uses (54).

Plugging the two bounds in (48), we obtain the following bound:

$$\begin{aligned} \frac{1}{T_{\mathsf{m}}} \sum_{t=0}^{T_{\mathsf{m}}-1} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq & \frac{\mathbb{E}[f(\theta_0)-f(\theta_{T_{\mathsf{m}}})]}{\eta \Delta_1 T_{\mathsf{m}}} + \frac{q^2 \eta + \sum_{k=t+1}^{\infty} \eta \beta_1^{k-t+2} \frac{G^2}{\epsilon 2 n^2}}{\eta \Delta_1 T_{\mathsf{m}}} \\ & + \frac{1-\beta_1}{\Delta_1} (\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}} \\ & + \left(\frac{L}{2} + 1 + \frac{\beta_1 L}{1-\beta_1}\right) \frac{1}{\eta \Delta_1} \eta^2 d (1-\beta_2)^{-1} (1-\frac{\beta_1^2}{\beta_2})^{-1} \end{aligned} \tag{56}$$

where $\Delta_1 := \frac{(1-\beta_1)}{2}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}$

With a constant stepsize $\eta = \frac{L}{\sqrt{T_{\text{m}}}}$ we get the final convergence bound as follows:

$$\begin{aligned}\frac{1}{T_{\text{m}}}\sum_{t=0}^{T_{\text{m}}-1}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq & \frac{\mathbb{E}[f(\theta_0) - f(\theta_{T_{\text{m}}})]}{L\Delta_1\sqrt{T_{\text{m}}}} + d\frac{L\Delta_3}{\Delta_1\sqrt{T_{\text{m}}}} \\ & + \frac{\Delta_2}{\eta\Delta_1 T_{\text{m}}} + \frac{1-\beta_1}{\Delta_1}(\epsilon + \frac{G^2}{1-\beta_2})^{-\frac{1}{2}}\end{aligned} \tag{57}$$

where $\Delta_2 := q^2 + \sum_{k=t+1}^{\infty}\beta_1^{k-t+2}\frac{G^2}{\epsilon 2n^2}$ and $\Delta_3 := \left(\frac{L}{2} + 1 + \frac{\beta_1 L}{1-\beta_1}\right)(1-\beta_2)^{-1}(1-\frac{\beta_1^2}{\beta_2})^{-1}$.

□