Communication-Compressed Adaptive Gradient Method for Distributed Nonconvex Optimization

Yujia Wang  
Pennsylvania State University

Lu Lin  
University of Virginia

Jinghui Chen  
Pennsylvania State University

Abstract

Due to the explosion in the size of the training datasets, distributed learning has received growing interest in recent years. One of the major bottlenecks is the large communication cost between the central server and the local workers. While error feedback compression has been proven to be successful in reducing communication costs with stochastic gradient descent (SGD), there are much fewer attempts in building communication-efficient adaptive gradient methods with provable guarantees, which are widely used in training large-scale machine learning models. In this paper, we propose a new communication-compressed AMSGrad for distributed nonconvex optimization problem, which is provably efficient. Our proposed distributed learning framework features an effective gradient compression strategy and a worker-side model update design. We prove that the proposed communication-efficient distributed adaptive gradient method converges to the first-order stationary point with the same iteration complexity as uncompressed vanilla AMSGrad in the stochastic nonconvex optimization setting. Experiments on various benchmarks back up our theory.

1 Introduction

The recent success of deep learning and large-scale training made it possible to use machine learning to solve complicated real-world tasks, such as computer vision (He et al., 2016), speech recognition (Hinton et al., 2012), natural language processing (Devlin et al., 2018), etc. Due to the explosion in the size of training datasets in recent years, single GPU training on such models can easily take up to weeks or even months to finish. As a consequence, distributed training algorithms have attracted growing interest over the years. In standard distributed settings with one central server and $n$ workers, local workers parallelly compute local gradients, while the server aggregates the gradients from the workers, updates the model, and sends the new model back to the workers. However, data transmissions across the machines can be quite expansive in terms of both communication costs (especially for cellular networks) and latency. Therefore, various methods were studied in order to achieve communication-efficient distributed learning by either reducing communication bits (Stich et al., 2018; Basu et al., 2019; Karimireddy et al., 2019) or saving the number of communication rounds (Zheng et al., 2017; Chen et al., 2021b). One of the principled approaches for effectively reducing the communication cost is to perform gradient compression before transmissions. It directly compresses the local fresh gradient on each worker before uploading the gradient to the server. However, the compression would slow down the convergence or even diverge (Beznosikov et al., 2020) due to the loss of information at each compression step. Later on, error feedback (Stich et al., 2018; Karimireddy et al., 2019) was proposed to alleviate this problem and reduce the information loss by proposing a compensating error sequence. Each worker compresses and uploads the combination of the last step compensating error and the local fresh gradient instead of compressing the gradient directly. Recent studies (Seide et al., 2014; Karimireddy et al., 2019; Stich et al., 2018) show that error feedback has been widely used in distributed SGD to communicate efficiently and ensure the same convergence rate as vanilla SGD.

While communication-efficient distributed SGD has been widely studied, there are much fewer attempts in building communication-efficient distributed adaptive gradient methods (Kingma and Ba, 2014; Reddi et al., 2018). It has been shown that SGD works less
efficiently compared with adaptive gradient methods when training large-scale models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020) and GAN (Goodfellow et al., 2014). One of the major challenges is that the traditional error feedback mechanism is not compatible with adaptive gradient methods since the variance term (i.e., second-order moments of the historical gradients) in the adapted gradient method can be unstable due to the accumulated compression error (Tang et al., 2021). 1-bit Adam (Tang et al., 2021) partially solve this problem by first using vanilla Adam at the beginning of training to get an estimate of variance term, then freezing this variance term and performing distributed Adam with the fixed variance. While this solution indeed gets rid of the unstable variance issue, the adaptivity is no longer changing and the major part of the algorithm is actually more similar to SGD with momentum. Due to the same reason, the convergence guarantee provided in Tang et al. (2021) is not related to distributed Adam but is an extension to distributed SGD with momentum.

In this paper, we develop a new compression method which is provably efficient, namely Communication-compressed Distributed Adaptive gradient Method (CD-Adam). We revisit the unstable variance issue in distributed AMSGrad (Reddi et al., 2018) and identify the need for an improved gradient compression strategy. Specifically, we adopt the Markov compression sequence recently proposed by Richtárik et al. (2021), which ideally could lead to contractive compression error for gradient descent if the gradient sequence (to be compressed) is convergent. Note that this is not simply applying the conclusion of Richtárik et al. (2021) to the distributed AMSGrad setting, as Richtárik et al. (2021) only deals with worker-to-server compression. While extending Richtárik et al. (2021) to both worker-to-server and server-to-worker compression may seem straightforward for standard gradient descent, it is particularly tricky for the adaptive gradient method (see Section 5 for details). Therefore, it requires carefully designed algorithms and analyses to build a provably efficient distributed adaptive gradient method.

We summarize our contribution as follows:

- We propose a new communication-compressed distributed AMSGrad approach which solves the bottleneck of applying communication compression strategies for fully-functional adaptive gradient methods in the distributed setting. The proposed method largely reduces the communication cost by enforcing both worker-to-server

and server-to-worker communication compression without any warm-up stages.

- We theoretically prove the convergence of our proposed algorithm in the nonconvex stochastic optimization setting. Specifically, we show that our proposed fully compressed CD-Adam can reach its first-order $\epsilon$-stationary point within $O(1/\epsilon^2)$ iterations, which is the same iteration complexity as the uncompressed vanilla AMSGrad. This suggests that without applying any variance-freezing tricks, the fully compressed distributed adaptive gradient method can still provably converge to its $\epsilon$-stationary point.

- Thorough experiments on various real-world benchmarks show that our proposed CD-Adam reduces the communication cost by around 32x over the original AMSGrad and around 5x over 1-bit Adam.

2 Related Work

Stochastic gradient descent and adaptive gradient methods: Stochastic gradient descent (SGD) (Robbins and Monro, 1951) is broadly used in training large-scale machine learning problems. Despite being simple to implement, SGD can be sensitive to parameters such as learning rate and slow to converge. Adaptive gradient methods were proposed to further improve over SGD. Adam (Kingma and Ba, 2014), one of the most popular adaptive gradient methods, has shown to be fast convergent and also robust to hyperparameters like learning rate. It designs an adaptive learning rate for each different coordinate using the past gradient. Aside from Adam, AdaGrad (Duchi et al., 2011) applied the second moment of the gradient to adaptive the learning rate, RMSProp (Tieleman et al., 2012) further used the second moment of the gradient with a decay rate, AdaDelta (Zeiler, 2012) is an extension of AdaGrad with a non-increasing learning rate. Later on, Reddi et al. (2018) pointed out a non-convergence issue of Adam, and proposed a new AMSGrad algorithm for ensuring the convergence. Zaheer et al. (2018) studied the effect of adaptive denominator constant $\nu$ and mini-batch size in the convergence of adaptive gradient methods. AdaBound (Luo et al., 2019) proposed with both upper and lower bound for the variance term of Adam. AdamW (Loshchilov and Hutter, 2019) proposed to fix the weight decay regularization in Adam by decoupling the weight decay from the gradient update. Chen et al. (2020) proposed a partially adaptive gradient method and proved its convergence in nonconvex settings. Chen et al. (2019); Zhou et al. (2018) showed the convergence rate of a class of adaptive gradient

---

1Here we use “fully-functional” to differentiate with the variance-freezed Adam used in Tang et al. (2021).
methods under the nonconvex stochastic optimization setting. [Alacaoglu et al. (2020)] proposed a new framework to derive data-dependent regret bounds with a constant momentum parameter in various settings.

**Distributed SGD and error feedback:** For communication-efficient distributed SGD, one of the most common strategies is to compress the gradients before uploading. [Alistarh et al. (2017)] provided a theoretical analysis of the centralized compressed distributed SGD. [Seide et al. (2014)] compressed the coordinates of the gradient into ±1 by its sign. [Bernstein et al. (2018)] proposed signSGD and proved its convergence in the nonconvex setting. Variant works has applied kinds of compression methods such as sparsification [Stich et al. (2018); Basu et al. (2019)], quantization [Karimireddy et al. (2019)], and sketching [Ivkin et al. (2019); Mishchenko et al. (2019)] proposed DIANA, which adopts the compression of gradient difference in distributed settings. Horváth et al. (2019) introduced several variants of DIANA including the variance reduction version. Philippenko and Dieuleveut (2020) introduced a bi-directional compression schemes with memory in distributed or federated setting. Error feedback (or error compensation) largely improves the compression error bound and has been shown to be critical for ensuring fast convergence of the compression mentioned above. Seide et al. (2014) showed that with error feedback, even 1-bit gradient communication under SGD still obtains the convergence rate of vanilla SGD. Karimireddy et al. (2019) applied error feedback to signSGD [Bernstein et al. (2018)] under nonconvex setting, Stich et al. (2018) performed error feedback in sparsity strong convex settings, and Stich and Karimireddy (2019) proposed a error-feedback framework also works on nonconvex but single node setting. Tang et al. (2019) proposed bi-directional SGD method combining with error compensation. Gorbunov et al. (2020) proposed a framework with variants of quantized SGD including EC-SGD and EC-SGD-DIANA. Liu et al. (2020); Zheng et al. (2019); Philippenko and Dieuleveut (2021) also studied combining bi-directional compression and error feedback strategy. There are also more works on the distributed SGD under nonconvex optimization (Koloskova et al. (2019); Basu et al. (2019)).

**Communication-efficient distributed adaptive gradient method:** There are only few attempts in developing communication-efficient distributed adaptive gradient methods. 1-bit Adam [Tang et al. (2021)] adopted a variance-freezed Adam by pointing out that the variance term of Adam becomes stable in later training stages. Combined with error feedback, 1-bit Adam achieves the same convergence rate as distributed SGD. Zhong et al. (2021) proposed a gradient compressed version of LANS [Zheng et al. (2020). Chen et al. (2021a)] developed a distributed quantized Adam with error feedback. However, the proposed algorithm in Chen et al. (2021a) can only converge to its ε-stationary point with worker-to-server compression alone but not fully compressed Adam with both worker-to-server and server-to-worker compression. Another related work is CADA [Chen et al. (2021b)], which reduced the communication rounds instead of performing communication compression. CADA adaptively reused the stale update parameters, and it achieved the same convergence rate as vanilla AMSGrad for nonconvex optimization. Note that CADA’s strategy of reducing communication rounds is orthogonal to ours can possibly be combined with ours for further improvements.

### 3 Problem Formulation

In this paper, we aim to solve the following distributed optimization problem under nonconvex stochastic optimization setting:

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\xi(i)} f_i(x; \xi(i)), \tag{3.1}
\]

where \(d\) denotes the dimension of the model, \(n\) denotes the total amount of workers, \(\xi(i)\) denotes the stochastic noise variable, and \(f_i(x)\) denotes the loss function on the \(i\)-th worker. In the stochastic setting, we cannot obtain the full gradient of loss function \(f_i(x)\). Instead, we can only obtain the unbiased estimators of \(\nabla f_i(x)\), i.e., \(\nabla f_i(x; \xi(i))\).

First, let us revisit the vanilla distributed setting of AMSGrad [Reddi et al. (2018)]. Consider a distributed learning system that contains a parameter server and \(n\) workers, with each worker \(i\) owns its local data from distribution \(D_i\). Let \(x_t\) denotes the current model at \(t\)-th iteration. The server will first broadcast \(x_t\) to all the workers in each iteration. Each worker \(i\) computes the stochastic gradient \(g_i^{(t)} = \nabla f_i(x_t; \xi(i))\) using the local samples, and then uploads \(g_i^{(t)}\) to the server. The server then aggregates the stochastic gradients and obtains \(g_t = \frac{1}{n} \sum_{i=1}^{n} g_i^{(t)}\). The model uses \(g_t\) to update parameter via vanilla single node AMSGrad:

\[
m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \quad \tilde{v}_t = \max(\tilde{v}_{t-1}, v_t), \quad x_{t+1} = x_t - \alpha_t \tilde{v}_t^{-1/2} m_t,
\]

where \(\tilde{v}_t = \text{diag}(\tilde{v}_t + \nu)\), \(\alpha_t > 0\) is the step size, \(\nu > 0\) is a small number for numerical stability, and \(\beta_1, \beta_2 \in [0, 1]\) are adaptive hyperparameters. This vanilla distributed AMSGrad achieves the same convergence rate as its centralized version [Reddi et al.
Upon compression, the server collects the compressed gradients $\hat{\mathbf{g}}_t^{(i)}$ from all workers and update model parameters via vanilla SGD. Karimireddy et al. (2019) showed that the compression error of error feedback is bounded by constant if the following assumption hold for the biased compressor.

**Assumption 4.1 (Biased compressor).** Consider a biased operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$, there exists a constant $0 < \pi \leq 1$ such that

$$\mathbb{E}_{C}[\|C(x) - x\|^2] \leq \pi\|x\|^2, \quad \forall x \in \mathbb{R}^d. \quad (4.1)$$

Note that $\pi = 0$ leads to $\mathcal{C}(x) = x$. Assumption 4.1 is a common assumption for biased compressor (Richtárik et al., 2021) [Karimireddy et al., 2019; Stich et al., 2018]. Canonical examples of the compressor satisfying Assumption 4.1 include top-$k$ or random-$k$ as well as scaled sign compressor. With Assumption 4.1 the distributed SGD with error feedback (Karimireddy et al., 2019) in nonconvex setting achieves the same convergence rate as vanilla SGD.

### Drawbacks of error feedback on adaptive gradient methods:
While error feedback guarantees bounded accumulated gradient compression error, in adaptive gradient methods (Kingma and Ba, 2014; Reddi et al., 2018), the variance term $v_t$, which is the moving average of the quadratic of gradient, can be unstable due to accumulating gradient compression error (Tang et al., 2021). Specifically, let’s denote $\mathbf{g}_t$ as the averaged fresh gradient without compression (average of all $\mathbf{g}_t^{(i)}$), denote $\tilde{\mathbf{g}}_t$ as the averaged compressed gradient (average of all $\hat{\mathbf{g}}_t^{(i)}$). The updating rule for $v_{t+1}$ follows:

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2) \|\mathbf{g}_t\|^2$$

$$= \beta_2 v_t + (1 - \beta_2) [\|\mathbf{g}_t - \tilde{\mathbf{g}}_t + \hat{\mathbf{g}}_t\|^2]$$

$$= \beta_2 v_t + (1 - \beta_2) [\|\tilde{\mathbf{g}}_t - \mathbf{g}_t\|^2 + (1 - \beta_2) (\hat{\mathbf{g}}_t - \mathbf{g}_t)^2]$$

$$+ 2(1 - \beta_2) (\hat{\mathbf{g}}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t). \quad (4.2)$$

Tang et al. (2021) claims that the inner product term in (4.2) can possibly be canceled out during training, while the quadratic term will certainly accumulate. Since the traditional error feedback can only guarantee constant compression error, the accumulation of the quadratic error will make the variance diverge. Therefore, the communication-efficient adaptive gradient method actually requires a stronger gradient compression error bound to obtain a stable variance term.
5 Proposed Method

In this section, we formally develop our proposed Communication-compressed Distributed Adaptive gradient Method (CD-Adam). Our proposed method consists of two key components: Markov compression sequence and worker-side model update design, which jointly provide a better compression error bound. We first investigate the definition and property of Markov compression sequence \( \text{[Richtárik et al. 2021]} \).

Markov compression sequence: Markov compression sequence is introduced\(^4\) in Richtárik et al. (2021). Given a biased compressor \( C(\cdot) \) and a sequence of vectors \( \{w_t\} \), Markov compression sequence \( \{\tilde{w}_t\} \) can be recursively defined as: \( \tilde{w}_0 = C(w_0) \), \( \tilde{w}_{t+1} = \tilde{w}_t + C(w_{t+1} - \tilde{w}_t) \). The main advantage of Markov compression sequence lies in that the compression error can be largely reduced if the underlying sequence \( \{w_t\} \) is convergent:

\[
\|\tilde{w}_{t+1} - w_{t+1}\|^2 = \||\tilde{w}_t + C(w_{t+1} - \tilde{w}_t) - w_{t+1}\|^2 \\
\leq \pi\|w_{t+1} - \tilde{w}_t\|^2 \\
\leq \pi(1 + \gamma\|\tilde{w}_t - w_t\|^2) + \pi(1 + \gamma^{-1})\|w_{t+1} - w_t\|^2,
\]

where the last inequality holds due to Young’s inequality. As can be seen from (5.1), the compression error \( \|\tilde{w}_{t+1} - w_{t+1}\| \) is directly controlled by \( \|w_{t+1} - \tilde{w}_t\| \). If the sequence of \( \{w_t\} \) is convergent, Markov compression error at \( t \)-step will be much smaller than the constant error bound obtained by plain error feedback. In particular, Richtárik et al. (2021) showed that if the sequence of \( \{w_t\} \) achieves a linear rate of convergence, the compression error of the Markov compression can converge to 0 as \( t \to \infty \).

Worker-side model update: Richtárik et al. (2021) only considered a one-way compression, i.e., workers compress the gradient difference \( g_t^{(i)} - \tilde{g}_{t-1}^{(i)} \) and update \( C(g_t^{(i)} - \tilde{g}_{t-1}^{(i)}) \) to the server, while the server only updates model \( x_{t+1} \) and broadcasts it to workers. In this case, the server broadcast still results in high communication overhead. Therefore, to further reduce the communication cost, we change the model update style from server-side update to worker-side update, i.e., the model \( x_t \) is stored and updated on each worker instead of on the server. The reason is that, in server-side update style, the server will need to compress the model update parameter \( V_t^{-1/2} \) for reducing the communication cost, however, it is difficult to obtain the convergence of the underlying sequence \( V_t^{-1/2} \), which is essential for the proof of the Markov compression sequence. Thus, it is difficult to get a better compression error bound for the model update. In our design, the server simply aggregates the compressed gradients and compressed it again before sending back to the workers. The compression errors via Markov compressor are directly related to fresh stochastic gradients or aggregated gradients, whose convergences are much easier to obtain. This two-way compression together with worker-side model update take the full advantage of the Markov compressor for establishing the theoretical guarantees of our proposed method.

Algorithm overview: Combining Markov compression sequence and worker-side model update design, we propose a new communication-compressed distributive gradient method, which is summarized in Algorithm \( \square \). Specifically, at \( t \)-th iteration, each worker first computes the stochastic gradient \( g_t^{(i)} \) with a mini-batch size \( \tau \), then builds the Markov compression sequence \( \tilde{g}_t^{(i)} \) with \( c_t^{(i)} = C(g_t^{(i)} - \tilde{g}_{t-1}^{(i)}) \) and \( \tilde{g}_{t-1}^{(i)} \). Note that the Markov compression sequence \( \tilde{g}_t^{(i)} \) is no longer bits-compressed, therefore, only \( c_t^{(i)} \) is being uploaded to the server. On the other hand, the server updates the aggregated compressed gradient by \( \tilde{g}_t = \tilde{g}_{t-1} + \frac{1}{n} \sum_{i=1}^{n} c_t^{(i)} \) as the last iterate \( \tilde{g}_{t-1} \) has been stored locally in the server. The server then builds another Markov compression sequence \( \tilde{g}_t \) based on \( \tilde{g}_{t-1} \) and \( c_t = C(\tilde{g}_t - \tilde{g}_{t-1}) \) and sends \( c_t \) to all workers. Similarly, each worker can recover the Markov compression sequence \( \tilde{g}_t \) upon receiving \( c_t \), since the last iterate \( \tilde{g}_{t-1} \) has been locally stored in each worker. Then each worker uses the double-compressed gradient \( \tilde{g}_t \) for updating model \( x_{t+1} \) as in standard AMSGrad.

Algorithm \( \square \) is indeed communication-efficient for distributed learning. It improves the communication efficiency by reducing the communication bits during transmission. Only the compressed vectors instead of the full precision vectors are transferred for both server-to-worker and worker-to-server communications. Specifically, if we adopt the scaled sign compressor, CD-Adam only takes 1-bit\(^5\) instead of 32-bits per dimension in each communication round. This greatly reduces the communication costs for distributed implementation of adaptive methods. Note that compared with 1-bit Adam (Fang et al. 2021), which performs a few epochs of uncompressed Adam at the beginning of training, our proposed method is much more communication-efficient (see Figure \( \square \) as we start the compression from the very first iteration.

\( ^4 \)Note that Künstner (2017) also proposed QDGD-F algorithm which is similar to the Markov compression sequence.

\( ^5 \)Rigorously, the scaling number will also take 32-bits for communication, so the overall cost for compressing a \( d \)-dimensional vector should be \( 32 + d \) bits.
Algorithm 1 Communication-Compressed Distributed Adaptive Gradient Method (CD-Adam)

Input: initial point \( x_1 \), non-increasing step size \( \{\alpha_t\}_t \), \( \beta_1, \beta_2, \nu \), batch size \( \tau \), compressor \( C(\cdot) \).

1. \( \mathbf{g}_0^{(i)} \leftarrow 0, \mathbf{m}_0 \leftarrow 0, \mathbf{v}_0 \leftarrow 0, \mathbf{g}_0 = C(\mathbf{g}_0^{(i)}), \mathbf{g}_0 = C(\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_0^{(i)}) \)
2. for \( t = 1 \) to \( T \) do
3. (On \( i \)-th worker)
4. Compute local stochastic gradient \( \mathbf{g}_t^{(i)} \) with batch size \( \tau \)
5. Compress \( \mathbf{c}_t^{(i)} = C(\mathbf{g}_t^{(i)} - \mathbf{g}_{t-1}^{(i)}) \)
6. Send \( \mathbf{c}_t^{(i)} \) to the server and update local state \( \mathbf{g}_t^{(i)} = \mathbf{g}_{t-1}^{(i)} + \mathbf{c}_t \)
7. (On Server)
8. Update \( \mathbf{g}_t = \mathbf{g}_{t-1} + \frac{1}{n} \sum_{i=1}^{n} \mathbf{c}_t^{(i)} \)
9. Compress \( \mathbf{c}_t = C(\mathbf{g}_t - \mathbf{g}_{t-1}) \)
10. Send \( \mathbf{c}_t \) to all the workers and update local state \( \mathbf{g}_t : \mathbf{g}_t = \mathbf{g}_{t-1} + \mathbf{c}_t \)
11. (On \( i \)-th worker)
12. Update \( \mathbf{g}_t = \mathbf{g}_{t-1} + \mathbf{c}_t \)
13. \( \mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \)
14. \( \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 \)
15. \( \widetilde{\mathbf{v}}_t = \max(\mathbf{v}_t - 1, \mathbf{v}_t) \)
16. Update \( x_{t+1} = x_t - \alpha_t \widetilde{\mathbf{v}}_t^{1/2} \mathbf{m}_t \) with \( \widetilde{\mathbf{v}}_t = \text{diag}(\mathbf{v}_t + \nu) \)
17. end for

6 Convergence Analysis

In this section, we present the convergence results of our proposed Communication-Compressed Distributed Adaptive Gradient Method (CD-Adam). Before we jump into the main theorem, let us first introduce the additional assumptions needed.

Assumption 6.1 (Smoothness). Each component loss function on the \( i \)-th worker \( f_i(x) = E[f(x; \xi^{(i)})] \) is \( L \)-smooth, i.e., \( \forall x, y \in \mathbb{R}^d, \)
\[
|f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle| \leq \frac{L}{2} \|x - y\|_2^2.
\]

Assumption 6.1 is a standard assumption for non-convex stochastic optimization [Koloskova et al., 2019; Basu et al., 2019; Richtárik et al., 2021]. Note that the \( L \)-smoothness assumption on each worker’s loss \( f_i(\cdot) \) implies the \( L \)-smoothness condition on \( f(\cdot) \). Assumption 6.1 also implies the \( L \)-gradient Lipschitz condition, i.e., \( \|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq L \|x - y\|_2 \).

Assumption 6.2 (Bounded gradient). Each component loss function on the \( i \)-th worker \( f_i(x) \) has \( G \)-bounded stochastic gradient on \( \ell_2 \) and has \( G_{\infty} \)-bounded stochastic gradient on \( \ell_{\infty} \), i.e., for all \( \xi \),
\[
\|\nabla f_i(x; \xi)\|_2 \leq G, \quad \|\nabla f_i(x; \xi)\|_{\infty} \leq G_{\infty}.
\]

Assumption 6.2 is a common assumption in adaptive gradient literature [Kingma and Ba, 2014; Chen et al., 2019; Zhou et al., 2018; Chen et al., 2018]. It also implies the bounded gradient for \( f(x) \), i.e., \( \|\nabla f(x; \xi)\|_2 \leq G, \|\nabla f(x; \xi)\|_{\infty} \leq G_{\infty} \).

Figure 1: Comparison of training loss and testing accuracy against communication bits of training ResNet-18 on CIFAR-10. Our proposed method achieves around 32× communication cost improvement over the original AMSGrad and around 5× over 1-bit Adam.

Assumption 6.3 implies that the variance of stochastic gradients is bounded on each worker. The assumption of bounded variance for stochastic gradient has also been used in [Basu et al., 2019; Karimireddy et al., 2019; Horváth et al., 2019].

Now we are ready to present our main theorem.

Theorem 6.4. Under Assumptions 4.1, 6.1, 6.2 and 6.3, suppose \( \alpha_t = \alpha \), denote \( N \) as the number of data samples on each worker, the iterates of Algorithm 1 achieves an \( \epsilon \)-stationary point of (3.1), i.e., \( \min_{t \in \mathbb{T}} E[\|\nabla f(x_t)\|_2^2] \leq \epsilon \), if \( \alpha \leq \frac{n \epsilon}{6n M_3 + 6M_4 \sigma^2 + \frac{N(3M_3 \sigma)^2}{(N-1)^2} + \frac{3M_2}{\epsilon^2}}, \)
\[
T \geq \frac{36M_1 M_3}{\epsilon^2} + \frac{36M_1 M_4 \sigma^2}{n \epsilon^2} + \frac{3M_2}{\epsilon}, \tag{6.1}
\]
where
\[
M_1 = C \cdot \Delta f, \quad M_2 = \frac{CGG}{(1 - \beta_1) \sqrt{\nu}}, \quad M_3 = \frac{32CC_1 \sqrt{\nu}}{\nu(1 - \sqrt{\nu})^2}, \quad M_4 = \frac{4CC_1 \nu}{\nu(1 - \sqrt{\nu})^2}.
\]
Figure 2: Gradient norm comparison of different compressing strategies on nonconvex logistic regression trained by AMSGrad with scaled sign compressor. The upper row shows the norm convergence with respect to the communication cost, and the lower row is with respect to the training iteration.

and

\[ \Delta f = f(x_1) - \inf_x f(x), \tilde{G} = \frac{(1 + \sqrt{\pi})^2}{(1 - \sqrt{\pi})^2} G, \]

\[ C = 2(\tilde{G}^2 + \nu)^{1/2}, C_1 = 2L + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2, \]

\[ \tilde{G}_\infty = \frac{(1 + \sqrt{\pi})^2}{(1 - \sqrt{\pi})^2} G_\infty. \]

Remark 6.5. Theorem 6.4 suggests that Algorithm 1 reaches the \( \epsilon \)-stationary point of (3.1) with \( O(1/\epsilon^2) \) iterations, which matches the iteration complexity of uncompressed vanilla AMSGrad. This suggests that our proposed method is indeed provably efficient, and the fully compressed distributed adaptive gradient method can reach \( \epsilon \)-stationary point with the same iteration complexity as the uncompressed counterpart without applying any variance-freezing tricks.

Remark 6.6. Theorem 6.4 also suggests that a larger number of workers \( n \) can lead to a better iteration complexity as shown in (6.1).

In addition, the condition on the mini-batch size \( \tau \) is needed for controlling the stochastic variance in a mini-batch. Note that \( \tau \) is always smaller than the number of data samples on each worker \( N \) as the full batch gradient does not incur stochastic variance.

7 Experiments

In this section, we present empirical results of our proposed communication-compressed distributed adaptive gradient method and compare it with other communication-compressed distributed baselines. Specifically, we first present the experimental results on a nonconvex logistic regression problem as an illustrative case study. Then we present the deep learning experiments of image classification on standard benchmarks.

7.1 Illustrative case study: nonconvex logistic regression

We consider minimizing the following logistic regression problem with a nonconvex regularizer similar to Richtárik et al. (2021),

\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} \log \left( 1 + \exp \left( -y_i a_i^\top x \right) \right) + \lambda \sum_{j=1}^{d} \frac{a_j^2}{1 + x_j^2}, \]

where \( a_i \in \mathbb{R}^d \), \( y_i \) are training data with value \( \pm 1 \), and \( \lambda \) is the regularizer parameter. We set \( \lambda = 0.1 \) in the following nonconvex logistic regression experiments.

We use datasets phishing, mushrooms, a9a and w8a from LibSVM (Chang and Lin, 2011). We equally separate each dataset to \( n = 20 \) parts, and we use full batch gradients in this experiment. For each method, we choose the best step size starting from 0.001 and increase it by adding 0.002 till achieving 0.01.

Let’s first study the effect of using different compression strategies on AMSGrad, though the other strategies do not enjoy any theoretical guarantees. Figure 2 compares the gradient norm convergence of our
Figure 3: Comparison on gradient norm, training loss and testing accuracy among our proposed method and the baselines when training ResNet-18 model on CIFAR-10.

Our proposed CD-Adam with AMSGrad using error feedback, naive compression and without compression at all. For all compressed methods, we use the scaled sign as a canonical example of biased compressor \( C(\cdot) \). We observe that our proposed CD-Adam achieves the best performances on all four datasets against the other three compression strategies. Specifically, if we compare the communication cost, CD-Adam achieves much better communication efficiency against uncompressed AMSGrad. Compared with the error feedback and the naive compression, CD-Adam still achieves a much smaller gradient norm under the same communication budgets. If we compare the performance of different compression strategies under the same iteration complexity, our proposed algorithm achieves roughly the same final gradient norm as the uncompressed AMSGrad across all four datasets. In fact, it achieves a much better convergence result comparing with the error feedback and the naive compression strategies, whose gradient norm stops decreasing at the early stages.

### 7.2 Deep learning experiments on image classification

We compare CD-Adam with several state-of-the-art communication-compressed distributed learning algorithms that are provably efficient, including: (1) EF21 \cite{richtarik2021communication} (2) 1-bit Adam \cite{tang20211}. We adopt CD-Adam with scaled sign compressor in this experiment. Note that the original EF21 paper \cite{richtarik2021communication} adopts the top-K compressor as the base compressor \( C(\cdot) \) with only worker-to-server compression applied. For a fair comparison, we further extend EF21 to allow both worker-to-server and server-to-worker compression and choose \( K = 0.016d \) such that the communication compression ratio keeps roughly the same as the scaled-sign based compressor.

We train CIFAR10 \cite{krizhevsky2009learning} using three popular models: ResNet-18 \cite{he2016deep}, VGG-16 \cite{simonyan2014very} and WideResNet-16-4 \cite{zagoruyko2016wide}. The image classification dataset CIFAR10 includes a training set of 50000 images and a test set of 10000 images, where each image is assigned one of the 10 labels. The dataset is split into \( n = 8 \) equal parts, which are distributed to 8 workers. The mini-batch size for each worker is set to be 128. We set the learning rate as \( 1 \times 10^{-1} \) for SGD, which is used in EF21, and \( 1 \times 10^{-4} \) for 1-bit Adam and our proposed CD-Adam. We also set \( \beta_1 = 0.9 \) and \( \beta_2 = 0.99 \) for 1-bit Adam and our CD-Adam. For 1-bit Adam, its warm-up epochs are set as 13 according to its original paper \cite{tang20211}. The normalized weight decay is set to \( 5 \times 10^{-4} \) for all methods. We test all methods for 100 epochs and decay the learning rate by 0.1 at the 50th and 75th epoch. Due to the space limit, we leave the experimental results on VGG-16 and WideResNet-16-4 models in the supplemental materials.

We show that our proposed algorithm enjoys a fast convergence speed with high accuracy and less communication cost comparing with EF21 and 1-bit Adam. The first and second plots in Figure 3 demonstrate the training gradient norm in terms of epochs and communication bits respectively. We can observe that our proposed CD-Adam obtains a smaller gradient norm than EF21 and 1-bit Adam under the same epoch or communication budget. Note that since 1-bit Adam will need to run uncompressed Adam for a few epochs as a warm-up, its per communication bits performance is actually much worse (has been shown in Figure 1). The third plot in Figure 3 shows the training loss against communication bits on CIFAR-10 dataset. At the early stage of the training process, CD-Adam and EF21 obtain a similar speed of reducing the training loss, while 1-bit Adam is much slower due to the warm-up process. At later stages, CD-Adam shows a clear advantage compared with EF21. The last plot in Figure 3 shows the test accuracy against communication bits. Similar phenomenon can be observed that CD-Adam achieves the overall best test accuracy compared with EF21 and 1-bit Adam, under the same communication budget. These results suggest that our pro-
posed CD-Adam is indeed much more communication-efficient while maintaining a high accuracy compared with other communication compression baselines.

8 Conclusion

In this paper, we propose a communication-compressed distributed adaptive gradient method which solves the bottleneck of applying communication compression strategies for adaptive gradient methods in distributed settings. We provide theoretical convergence analysis in the nonconvex stochastic optimization setting and show that our proposed algorithm converges to an ϵ-stationary point with the same iteration complexity as the uncompressed vanilla AMSGrad. Furthermore, compared with prior work which adopts variance-freezed Adam, our proposed algorithm is fully adaptive during the entire training process. It does not require any warm-up procedure with uncompressed communication, leading to better empirical results in terms of both performance and communication cost.

References


international Conference on Machine Learning, pages 4120–4129. PMLR.


A  Compression Schemes

Here we provide several compressors satisfy Assumption 4.1.

**Rand-k** (Stich et al., 2018): For $1 \leq k \leq d$, denotes $S$ as chosen $k$ elements randomly in $[d] = \{1, 2, \ldots, d\}$, i.e. $S \subseteq [d] = \{1, 2, \ldots, d\}$, $|S| = k$. Denotes $e_1, e_2, \ldots, e_d$ as standard unit basis vectors in $\mathbb{R}^d$, for $x \in \mathbb{R}^d$, the compressor $C_{\text{rand}} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as

$$C_{\text{rand}}(x) = \sum_{i \in S} x_i e_i.$$ 

For Rand-k compressor, we have

$$\mathbb{E}[\|C_{\text{rand}}(x) - x\|_2^2] = \mathbb{E}[\left\| \sum_{i \in S} x_i e_i \right\|_2^2] = \left(1 - \frac{k}{d}\right)\|x\|^2_2. \quad (A.1)$$

**Top-k** (Stich et al., 2018): For $1 \leq k \leq d$ and $x \in \mathbb{R}^d$, the order of $x$ is ordered by the magnitude of coordinates with $|x_{(1)}| \leq |x_{(2)}| \cdots \leq |x_{(d)}|$. The compressor $C_{\text{Top}} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as

$$C_{\text{Top}}(x) = \sum_{i = d-k+1}^{d} x(i)e(i).$$

For Top-k compressor, considering the definition of rand-k compressor, we have

$$\|C_{\text{Top}}(x) - x\|_2 \leq \|C_{\text{rand}}(x) - x\|_2;$$

thus we conclude

$$\|C_{\text{Top}}(x) - x\|_2 \leq \left(1 - \frac{k}{d}\right)\|x\|^2_2.$$

**Scaled sign** (Karimireddy et al., 2019): For $1 \leq k \leq d$ and $x \in \mathbb{R}^d$, the scaled sign compressor $C_{\text{sign}} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as

$$C_{\text{sign}}(x) = \|x\|_1 \cdot \text{sign}(x)/d,$$

where the sign($x$) takes sign of each coordinates. For scaled sign compressor, we have

$$\|C_{\text{sign}}(x) - x\|_2^2 = \sum_{i = 1}^{d} \left(\frac{1}{d} \sum_{j = 1}^{d} |x_j| - |x_i|\right)^2$$

$$= \sum_{i = 1}^{d} \left(\frac{1}{d} \sum_{j = 1}^{d} |x_j|\right)^2 - 2 \left(\sum_{i = 1}^{d} |x_i|\right) \left(\frac{1}{d} \sum_{j = 1}^{d} |x_j|\right) + \sum_{i = 1}^{d} |x_i|^2$$

$$= d \cdot \frac{1}{d} \|x\|^2_2 - \frac{2}{d} \|x\|^2_1 + \|x\|^2_2$$

$$= \|x\|^2_2 - \frac{1}{d} \|x\|^2_1$$

$$= \left(1 - \frac{\|x\|^2_1}{d\|x\|^2_2}\right)\|x\|^2_2. \quad (A.2)$$

Thus we conclude that

$$\|C_{\text{sign}}(x) - x\|_2^2 = \left(1 - \frac{\|x\|^2_1}{d\|x\|^2_2}\right)\|x\|^2_2.$$

B  Main Theorem for Algorithm 1

We provide more following lemmas to support the proof of the main theorem.

**Lemma B.1.** For compressors $C(\cdot)$ satisfying Assumption 4.1 we have $\mathbb{E}[\|C(x)\|_2] \leq (1 + \sqrt{\pi})\|x\|_2$. 
Lemma [B.1] can be easily verified by applying triangle inequality to Assumption 4.1.

**Lemma B.2.** Under Assumptions 4.1 and 6.2 we have \( \|\nabla f(x)\|_2 \leq G, \|\tilde{g}_t\|_2 \leq \frac{1+\sqrt{\pi}}{1-\sqrt{\pi}}G = \hat{G}, \|\tilde{v}_t\|_2 \leq \hat{G}^2 \) and \( |\mu\|_2 \leq \hat{G} \). Additionally, we have \( \|\nabla f(x)\|_\infty \leq G_\infty, \|\tilde{g}_t\|_\infty \leq \frac{1+\sqrt{\pi}}{1-\sqrt{\pi}}G_\infty = \hat{G}_\infty, \|\tilde{v}_t\|_\infty \leq \frac{1+\sqrt{\pi}}{1-\sqrt{\pi}}G_\infty. \)

**Lemma B.2** shows that the compressed gradients and the update parameters can be bounded by constants, it is a quite standard result for the adam-type convergence analysis.

**Lemma B.3.** For the mini-batch stochastic gradient on each worker, its variance satisfies

\[
\mathbb{E}[\|g^{(i)}_t - \nabla f_t(x_t)\|_2^2] \leq \sigma^2 \frac{N - \tau}{\tau(N - 1)},
\]

where \( \tau \) denotes the batch size and \( N \) denotes the total amount of data on each worker.

**Lemma B.4.** Let \( \beta_1, \beta_2 \) be the weight parameters such that

\[
\begin{align*}
\mu_t &= \beta_1 \mu_{t-1} + (1 - \beta_1) \tilde{g}_t, \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) \tilde{g}_t^2,
\end{align*}
\]

set \( \alpha_t = \alpha, t = 1, \ldots, T \) be the step sizes. Under Assumption 6.2 and 6.3 we have the following two results:

\[
\sum_{t=1}^{T} \alpha_t^2 \|\tilde{v}_t^{-1/2} \mu_t\|_2^2 \leq \frac{32T\alpha^2 \hat{G}^2}{\nu} + \frac{4T\alpha^2 \sigma^2}{\nu n} + \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(x_t)\|_2^2],
\]

and

\[
\sum_{t=1}^{T} \alpha_t^2 \|\tilde{v}_t^{-1/2} \tilde{g}_t\|_2^2 \leq \frac{32T\alpha^2 \hat{G}^2}{\nu} + \frac{4T\alpha^2 \sigma^2}{\nu n} + \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(x_t)\|_2^2],
\]

**Lemma B.4** provides upper bounds for the summation of \( \alpha_t^2 \|\tilde{v}_t^{-1/2} \mu_t\|_2^2 \) and \( \alpha_t^2 \|\tilde{v}_t^{-1/2} \tilde{g}_t\|_2^2 \).

**Lemma B.5.** Denote \( \tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} g^{(i)}_t \) as the aggregated fresh gradient and denote \( \tilde{g}_t = \frac{1}{n} \sum_{i=1}^{n} g^{(i)}_t \) as the aggregated worker-side compressed gradient. The compression error of the worker-side compression \( \|\tilde{g}_t - g_t\|_2 \) satisfies

\[
\mathbb{E}[\|\tilde{g}_t - g_t\|_2] \leq \frac{L\hat{G}\sqrt{\pi}}{1 - \sqrt{\pi}} \alpha_1 \sqrt{d\nu^{-1/2}} + \frac{2\sqrt{\pi} \sigma}{1 - \sqrt{\pi}} \sqrt{\frac{N - \tau}{\tau(N - 1)}}, \tag{B.1}
\]

where \( N \) is the total number of data samples on each worker and \( \tau \) is the batch-size.

**Lemma B.6.** The compression error of the server-side compression \( \|\tilde{g}_t - \tilde{g}_t\|_2 \) satisfies

\[
\mathbb{E}[\|\tilde{g}_t - \tilde{g}_t\|_2] \leq \frac{L\hat{G}\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} \cdot \alpha_1 \sqrt{d\nu^{-1/2}} + \frac{2\sqrt{\pi}(1 + \sqrt{\pi})^\sigma}{(1 - \sqrt{\pi})^2} \sqrt{\frac{N - \tau}{\tau(N - 1)}}, \tag{B.2}
\]

where \( N \) is the total number of data samples on each worker and \( \tau \) is the batch-size.

**Lemma B.3** and **B.6** provide upper bounds for the worker-side and the server-side compression errors respectively. Following Zhou et al. (2018), we define an auxiliary sequence \( z_t \) as follows: let \( x_0 = x_1 \), and for each \( t \geq 1 \),

\[
z_t = x_t + \frac{\beta_1}{1 - \beta_1} (x_t - x_{t-1}). \tag{B.3}
\]

The following three lemmas directly follows Lemma 7.1, 7.2, and 7.3 from Zhou et al. (2018).
Lemma B.7. Let \( z_t \) be defined in \([B.3]\). For \( t \geq 2 \), we have
\[
\begin{aligned}
z_{t+1} - z_t &= \frac{\beta_1}{1 - \beta_1} \left[ I - (\alpha_t \hat{V}_t^{-1/2}) (\alpha_{t-1} \hat{V}_{t-1}^{-1/2})^{-1} \right] (x_{t-1} - x_t) - \alpha_t \hat{V}_t^{-1/2} g_t, \\
&= \frac{\beta_1}{1 - \beta_1} \left[ (\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}) m_{t-1} - \alpha_t \hat{V}_t^{-1/2} g_t, \right. \tag{B.4}
\end{aligned}
\]
and
\[
\begin{aligned}
z_{t+1} - z_t &= \frac{\beta_1}{1 - \beta_1} \left( \alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2} \right) m_{t-1} - \alpha_t \hat{V}_t^{-1/2} g_t, \\
&= \frac{\beta_1}{1 - \beta_1} \left( \alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2} \right) m_{t-1} - \alpha_t \hat{V}_t^{-1/2} g_t, \tag{B.5}
\end{aligned}
\]
Specifically, for \( t = 1 \), we have
\[
z_2 - z_1 = -\alpha_1 \hat{V}_1^{-1/2} g_1. \tag{B.6}
\]

Lemma B.8. Let \( z_t \) be defined in \([B.3]\). For \( t \geq 2 \), we have
\[
\begin{aligned}
\|z_{t+1} - z_t\|_2 &\leq \|\alpha_t \hat{V}_t^{-1/2} g_t\|_2 + \frac{\beta_1}{1 - \beta_1} \|x_{t-1} - x_t\|_2.
\end{aligned}
\]

Lemma B.9. Let \( z_t \) be defined in \([B.3]\). For \( t \geq 2 \), we have
\[
\begin{aligned}
\|\nabla f(z_t) - \nabla f(x_t)\|_2 &\leq L \left( \frac{\beta_1}{1 - \beta_1} \right) \cdot \|x_{t-1} - x_t\|_2.
\end{aligned}
\]
Together with previous Lemmas, now we are ready to prove Theorem 6.4.

Proof of Theorem 6.4. Since \( f \) is \( L \)-smooth, we have:
\[
\begin{aligned}
f(z_{t+1}) &\leq f(z_t) + \nabla f(z_t)^T (z_{t+1} - z_t) + \frac{L}{2} \|z_{t+1} - z_t\|_2^2 \\
&= f(z_t) + \nabla f(z_t)^T (z_{t+1} - z_t) + (\nabla f(z_t) - \nabla f(x_t))^T (z_{t+1} - z_t) + \frac{L}{2} \|z_{t+1} - z_t\|_2^2. \tag{B.7}
\end{aligned}
\]
When \( t = 1 \), for \([B.7]\) we have
\[
\begin{aligned}
f(z_2) &\leq f(z_1) + \nabla f(x_1)^T (z_2 - z_1) + (\nabla f(z_1) - \nabla f(x_1))^T (z_2 - z_1) + \frac{L}{2} \|z_2 - z_1\|_2^2 \\
&= f(z_1) - \nabla f(x_1)^T \alpha_1 \hat{V}_1^{-1/2} g_1 + 2L \|\alpha_1 \hat{V}_1^{-1/2} g_1\|_2^2. \tag{B.8}
\end{aligned}
\]
When \( t \geq 2 \), we have
\[
\begin{aligned}
\nabla f(x_t)^T (z_{t+1} - z_t) &= \nabla f(x_t)^T \left[ \frac{\beta_1}{1 - \beta_1} (\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}) m_{t-1} - \alpha_t \hat{V}_t^{-1/2} g_t \right] \\
&= \frac{\beta_1}{1 - \beta_1} \nabla f(x_t)^T (\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}) m_{t-1} - \alpha_t \hat{V}_t^{-1/2} g_t \\
&= \frac{\beta_1}{1 - \beta_1} \nabla f(x_t)^T (\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}) m_{t-1} + \nabla f(x_t)^T (\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}) g_t \\
&- \nabla f(x_t)^T \alpha_t \hat{V}_t^{-1/2} g_t \\
&\leq \frac{1}{1 - \beta_1} \|\nabla f(x_t)\|_2 \cdot \|\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}\|_{\infty, \infty} \cdot \|m_{t-1}\|_2 - \nabla f(x_t)^T \alpha_{t-1} \hat{V}_{t-1}^{-1/2} g_t \\
&\leq \frac{1}{1 - \beta_1} G\tilde{G} \left[ \|\alpha_{t-1} \hat{V}_{t-1}^{-1/2}\|_{\infty, \infty} - \|\alpha_t \hat{V}_t^{-1/2}\|_{\infty, \infty} \right] - \nabla f(x_t)^T \alpha_{t-1} \hat{V}_{t-1}^{-1/2} g_t, \tag{B.9}
\end{aligned}
\]
where the first equality holds due to \([B.5]\) in Lemma B.7. The first inequality holds because for vector \( x, y \) and positive diagonal matrix \( A \), we have \( x^T Ay \leq \|x\|_2 \cdot \|A\|_{\infty, \infty} \cdot \|y\|_2 \). The second inequality holds due to Lemma B.7 and \( \alpha_{t-1} \hat{V}_{t-1}^{-1/2} \geq \alpha_t \hat{V}_t^{-1/2} \geq 0 \). Thus for diagonal matrices \( \alpha_{t-1} \hat{V}_{t-1}^{-1/2} \) and \( \alpha_t \hat{V}_t^{-1/2} \), we directly conclude \( \|\alpha_{t-1} \hat{V}_{t-1}^{-1/2} - \alpha_t \hat{V}_t^{-1/2}\|_{\infty, \infty} = \|\alpha_{t-1} \hat{V}_{t-1}^{-1/2}\|_{\infty, \infty} - \|\alpha_t \hat{V}_t^{-1/2}\|_{\infty, \infty} \).
Next we bound the last two terms in (B.7):

\[
(\nabla f(z_t) - \nabla f(x_t))^T (z_{t+1} - z_t) + \frac{L}{2} \|z_{t+1} - z_t\|^2
\]

\[
\leq \|\nabla f(z_t) - \nabla f(x_t)\|_2 \cdot \|z_{t+1} - z_t\| + \frac{L}{2} \|z_{t+1} - z_t\|^2
\]

\[
\leq L \left( \frac{\beta_1}{1 - \beta_1} \right) \|x_t - x_{t-1}\| \cdot \left[ \left\| \alpha_t \hat{V}^{-1/2}_t \tilde{g}_t \right\|_2 + \frac{\beta_1}{1 - \beta_1} \|x_t - x_{t-1}\| \right] + \frac{L}{2} \left[ \|\alpha_t \hat{V}^{-1/2}_t \tilde{g}_t\|_2 + \frac{\beta_1}{1 - \beta_1} \|x_{t-1} - x_t\|^2 \right]
\]

\[
\leq \frac{L}{2} \|\alpha_t \hat{V}^{-1/2}_t \tilde{g}_t\|_2^2 + \frac{2L}{2} \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|x_t - x_{t-1}\|^2 + L \|\alpha_t \hat{V}^{-1/2}_t \tilde{g}_t\|_2^2 + L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|x_{t-1} - x_t\|^2
\]

\[
\leq 2L \|\alpha_t \hat{V}^{-1/2}_t \tilde{g}_t\|_2^2 + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|x_t - x_{t-1}\|^2,
\]

where the second inequality holds because of Lemma B.8 and B.9, the third inequality holds by applying Young’s inequality to the first term and applying Cauchy inequality to the second term.

For \( t = 1 \), recalling (B.8), taking expectation and rearranging terms, we have

\[
\mathbb{E}[f(x_2) - f(z_1)]
\]

\[
\leq \mathbb{E} \left[ -\nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 \tilde{g}_1 + 2L \|\alpha_1 \hat{V}^{-1/2}_1 \tilde{g}_1\|^2 \right]
\]

\[
= \mathbb{E} \left[ -\nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 \tilde{g}_1 - \nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 (\tilde{g}_1 - \tilde{g}_1) - \nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 (\tilde{g}_1 - \tilde{g}_1) + 2L \|\alpha_1 \hat{V}^{-1/2}_1 \tilde{g}_1\|^2 \right]
\]

\[
\leq \mathbb{E} \left[ -\alpha_1 \|\nabla f(x_1)\|^2\left( \tilde{G}_2^2 + \nu \right)^{-1/2} - \nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 (\tilde{g}_1 - \tilde{g}_1) - \nabla f(x_1)^T \alpha_1 \hat{V}^{-1/2}_1 (\tilde{g}_1 - \tilde{g}_1)
\]

\[
+ 2L \|\alpha_1 \hat{V}^{-1/2}_1 \tilde{g}_1\|^2 \right],
\]

where the first term of the last inequality holds because

\[
\mathbb{E}[ -\alpha_1 \nabla f(x_1)^T \hat{V}^{-1/2}_1 \nabla f(x_1)] = \mathbb{E}[ -\alpha_1 \nabla f(x_1)^T \hat{V}^{-1/2}_1 \nabla f(x_1)]
\]

\[
\leq -\alpha_1 (\tilde{G}_2^2 + \nu)^{-1/2} \mathbb{E}[\|\nabla f(x_1)\|^2].
\]

For \( t \geq 2 \), substituting (B.9) and (B.10) into (B.7), taking expectation and rearranging terms, we have

\[
\mathbb{E} \left[ f(z_{t+1}) + \frac{\tilde{G}_2 \|\alpha_t \hat{V}^{-1/2}_t \tilde{g}_t\|_{\infty, \infty}}{1 - \beta_t} - \left( f(z_t) + \frac{\tilde{G}_2 \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{m}_{t-1}\|_{\infty, \infty}}{1 - \beta_t} \right) \right]
\]

\[
\leq \mathbb{E} \left[ -\nabla f(x_t)^T \alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{g}_t + 2L \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{g}_t\|^2 + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|x_t - x_{t-1}\|^2 \right]
\]

\[
= \mathbb{E} \left[ -\nabla f(x_t)^T \alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{g}_t - \nabla f(x_t)^T \alpha_{t-1} \hat{V}^{-1/2}_{t-1} (\tilde{g}_t - \tilde{g}_t) - \nabla f(x_t)^T \alpha_{t-1} \hat{V}^{-1/2}_{t-1} (\tilde{g}_t - \tilde{g}_t)
\]

\[
+ 2L \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{g}_t\|^2 + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{m}_{t-1}\|^2 \right]
\]

\[
\leq \mathbb{E} \left[ -\alpha_{t-1} \|\nabla f(x_t)\|^2\left( \tilde{G}_2^2 + \nu \right)^{-1/2} + 2L \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{g}_t\|^2 + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|\alpha_{t-1} \hat{V}^{-1/2}_{t-1} \tilde{m}_{t-1}\|^2
\]

\[
+ \alpha_{t-1} \nu^{-1/2} \tilde{G}_2 \|\tilde{g}_t - \tilde{g}_t\|_2 + \alpha_{t-1} \nu^{-1/2} \tilde{G}_2 \|\tilde{g}_t - \tilde{g}_t\|_2 \right].
\]
where the equality holds because \( \bar{g}_t = g_t + (\bar{g}_t - g_t) + (\bar{g}_t - \bar{g}_t) \). The second inequality holds because of

\[
\mathbb{E}[-\nabla f(x_t) \nabla_{t-1} \tilde{\nabla}_{t-1}^{-1/2} g_t] = \mathbb{E}[-\nabla f(x_t) \nabla_{t-1} \tilde{\nabla}_{t-1}^{-1/2} \mathbb{E}[g_t]] \\
= \mathbb{E}[-\nabla f(x_t) \nabla_{t-1} \tilde{\nabla}_{t-1}^{-1/2} \nabla f(x_t)] \\
\leq -\alpha_{t-1} (G^2 + \nu)^{-1/2} \mathbb{E}[\|\nabla f(x_t)\|^2], \quad (B.13)
\]

\[
\mathbb{E}[-\nabla f(x_t) \nabla_{t-1} \tilde{\nabla}_{t-1}^{-1/2} (\bar{g}_t - g_t)] \leq \alpha_{t-1} \nu^{-1/2} \|\nabla f(x_t)\|_2 \cdot \mathbb{E}[\|\bar{g}_t - g_t\|_2] \\
\leq \alpha_{t-1} \nu^{-1/2} G \cdot \mathbb{E}[\|\bar{g}_t - g_t\|_2], \quad (B.14)
\]

\[
\mathbb{E}[-\nabla f(x_t) \nabla_{t-1} \tilde{\nabla}_{t-1}^{-1/2} (\bar{g}_t - \bar{g}_t)] \leq \alpha_{t-1} \nu^{-1/2} \|\nabla f(x_t)\|_2 \cdot \mathbb{E}[\|\bar{g}_t - \bar{g}_t\|_2] \\
\leq \alpha_{t-1} \nu^{-1/2} G \cdot \mathbb{E}[\|\bar{g}_t - \bar{g}_t\|_2]. \quad (B.15)
\]

Summing up (B.12) for \( t = 2 \) to \( T \), rearranging and adding with (B.11), we have

\[
(\tilde{G}_T^2 + \nu)^{-1/2} \sum_{t=1}^T \alpha_t \mathbb{E}[\|\nabla f(x_t)\|^2_2] \\
\leq \mathbb{E}\left[ f(z_1) + \frac{GG\|\alpha_t \hat{V}_{t-1}^{-1/2}\|_{\infty, \infty}}{1 - \beta_1} \right] + \frac{GG\|\alpha_t \hat{V}_{T}^{-1/2}\|_{\infty, \infty}}{1 - \beta_1} \\
+ 2L \sum_{t=1}^T \mathbb{E}[\|\alpha_t \hat{V}_{t-1}^{-1/2} g_t\|_2] + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \sum_{t=1}^T \mathbb{E}[\|\alpha_t \hat{V}_{t-1}^{-1/2} m_{t-1}\|_2] \\
+ \alpha_{t-1} \nu^{-1/2} G \cdot \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - g_t\|_2] + \alpha_{t-1} \nu^{-1/2} G \cdot \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - \bar{g}_t\|_2] \\
\leq \mathbb{E}\left[ \Delta f + \frac{\alpha_1 \tilde{G} \nu^{-1/2}}{1 - \beta_1} \right] + 2L \sum_{t=1}^T \mathbb{E}[\|\alpha_t \hat{V}_{t-1}^{-1/2} g_t\|_2] + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \sum_{t=1}^T \mathbb{E}[\|\alpha_t \hat{V}_{t-1}^{-1/2} m_{t-1}\|_2] \\
+ \alpha_{t-1} \nu^{-1/2} G \cdot \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - g_t\|_2] + \alpha_{t-1} \nu^{-1/2} G \cdot \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - \bar{g}_t\|_2], \quad (B.16)
\]

where \( \Delta f = f(x_1) - \inf_x f(x) \).

By Lemma B.4 with the constant stepsize \( \alpha_t = \alpha \), we have

\[
\sum_{t=1}^T \alpha_t^2 \mathbb{E}[\|\hat{V}_{t-1}^{-1/2} m_t\|^2_2] \leq \frac{32T \alpha^2 \tilde{G}^2}{\nu} + \frac{4T \alpha^2 \sigma^2}{\nu m} + \frac{4 \alpha^2}{\nu} \sum_{t=1}^T \mathbb{E}[\|\nabla f(x_t)\|^2_2], \quad (B.17)
\]

and

\[
\sum_{t=1}^T \alpha_t^2 \mathbb{E}[\|\hat{V}_{t-1}^{-1/2} g_t\|^2_2] \leq \frac{32T \alpha \tilde{G}^2}{\nu} + \frac{4T \alpha \sigma^2}{\nu m} + \frac{4 \alpha^2}{\nu} \sum_{t=1}^T \mathbb{E}[\|\nabla f(x_t)\|^2_2]. \quad (B.18)
\]

By Lemma B.5 and B.6, we conclude the last two terms of (B.16) as

\[
\sum_{t=1}^T \alpha_t \nu^{-1/2} G \cdot \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - g_t\|_2] + \sum_{t=1}^T \mathbb{E}[\|\bar{g}_t - \bar{g}_t\|_2] \\
\leq \frac{\sqrt{\pi}}{1 - \sqrt{\pi}} LTG \sqrt{\alpha_t} \nu^{-1/2} + \frac{\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} LTG \sqrt{\alpha_t} \nu^{-1/2} \\
+ \frac{2 \sqrt{\pi}}{1 - \sqrt{\pi}} T G \sqrt{\alpha_t} \nu^{-1/2} \sqrt{\frac{N - \tau}{\tau(N - 1)}} + \frac{2 \sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} T G \sqrt{\alpha_t} \nu^{-1/2} \sqrt{\frac{N - \tau}{\tau(N - 1)}} \\
\leq \frac{2 \sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} \sqrt{\frac{N - \tau}{\tau(N - 1)}} T G \sqrt{\alpha_t} \nu^{-1} \sqrt{d} + \frac{4 \sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} G T \sqrt{\alpha_t} \nu^{-1} \sqrt{d}, \quad (B.19)
\]
Substituting (B.17), (B.18) and (B.19) into (B.16), set the constant stepsize \( \alpha_t = \alpha \) and rearranging, we have

\[
(G_\infty^2 + \nu)^{-1/2} \sum_{t=1}^{T} \alpha E[\|\nabla f(x_t)\|_2^2] \\
\leq E\left[\Delta f + \frac{\alpha G\nu^{-1/2} 2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2}{\nu} \right] 2T^2 \frac{\alpha^2 \tilde{G}^2}{\nu} + \left[2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2\right] \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|_2^2] \\
+ \frac{2\sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} LCGT \alpha^2 \sqrt{d} + \frac{4\sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} GT \alpha \cdot \sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}}.
\]

Hence we obtain

\[
\left(\alpha(G_\infty^2 + \nu)^{-1/2} - \left[2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2\right] \frac{4\alpha^2}{\nu}\right] \sum_{t=1}^{T} \alpha E[\|\nabla f(x_t)\|_2^2] \\
\leq E\left[\Delta f + \frac{\alpha G\nu^{-1/2} 2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2}{\nu} \right] 2T^2 \frac{\alpha^2 \tilde{G}^2}{\nu} + \left[2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2\right] \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|_2^2] \\
+ \frac{2\sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} LCGT \alpha^2 \sqrt{d} + \frac{4\sqrt{\pi}}{\nu(1 - \sqrt{\pi})^2} GT \alpha \cdot \sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}}.
\]

Let \( 2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2 \alpha^2 \nu \leq \frac{\nu}{8(G_\infty^2 + \nu)^{1/2}} \), we obtain the condition on learning rate \( \alpha \leq \frac{\nu}{8G_\infty^2 (\frac{1}{\nu} + \sqrt{\pi})^2} = \frac{\nu}{4CC_1} \). Then we get

\[
\min_{t \in [T]} E[\|\nabla f(x_t)\|_2^2] \leq \frac{1}{T} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|_2^2] \leq \frac{M_1}{T\alpha} + \frac{M_2}{T} + \alpha M_3 + \frac{\alpha M_4 \sigma^2}{n} + M_5 \sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}},
\]

where

\[
M_1 = C \cdot \Delta f, \quad M_2 = CCG \sqrt{\frac{1}{\nu}}, \quad M_3 = \frac{32CC_1 \tilde{G}^2}{\nu} \sqrt{\frac{\pi CLG}{(1 - \sqrt{\pi})^2}},
\]

\[
M_4 = \frac{4CC_1}{\nu}, \quad M_5 = \frac{4\sqrt{\pi}CG}{\sqrt{\nu(1 - \sqrt{\pi})^2}},
\]

and

\[
\Delta f = f(x_1) - \inf_{x} f(x), \tilde{G} = C_2 G, \tilde{G}_\infty = C_2 G_\infty,
\]

\[
C = 2(G_\infty^2 + \nu)^{1/2}, C_1 = 2L + 3L\left(\frac{\beta_1}{1 - \beta_1}\right)^2, C_2 = \frac{(1 + \sqrt{\pi})^2}{(1 - \sqrt{\pi})^2}.
\]

After \( T \) iterations, it requires

\[
\frac{M_1}{T\alpha} + \alpha \left(M_3 + \frac{M_4 \sigma^2}{n}\right) \leq \frac{\epsilon}{3}, \quad \frac{M_2}{T} \leq \frac{\epsilon}{3}, \quad M_5 \sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}} \leq \frac{\epsilon}{3}.
\]

Note that \( \frac{\nu}{6M_5 T + 6M_4 \sigma^2} \leq \frac{\nu}{32CC_1 G^2} \leq \frac{\nu}{4CC_1} \), therefore, we get

\[
\alpha \leq \frac{ne}{6n M_3 + 6M_4 \sigma^2}, \quad \tau \geq \left[\frac{N(3M_2 \sigma^2)^2}{(N - 1) \epsilon^2 + (3M_2 \sigma)^2}\right] T \geq \left[\frac{36M_1 M_3}{\epsilon^2} + \frac{36M_1 M_4 \sigma^2}{n \epsilon^2} + 3M_2 \right].
\]

This concludes the proof.
C Supporting Lemmas for Algorithm 1

Proof of Lemma B.2. Since by Assumption A.1, $f$ has $G$-bounded stochastic gradient, for any $x$ and $\xi$, there is $\|\nabla f(x; \xi)\|_2 \leq G$. We have

$$\|\nabla f(x)\|_2 = \|E_\xi \nabla f(x; \xi)\|_2 \leq E_\xi \|\nabla f(x; \xi)\|_2 \leq G.$$  

First we prove the upper bound for compressed gradient on $i$-th worker by induction. For $t = 1$, it is obvious that $\|\tilde{g}_i^{(1)}\|_2 \leq \|C(g_i^{(1)})\|_2 \leq \|g_i^{(1)}\|_2 \leq G$. For $t > 1$, suppose we have $\|\tilde{g}_i^{(t)}\|_2 \leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} G = \tilde{G}$, our goal is to show that $\|\tilde{g}_i^{(t+1)}\|_2 \leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} G$:

$$\|\tilde{g}_i^{(t+1)}\|_2 = \|\tilde{g}_i^{(t)} + C(g_i^{(t+1)} - \hat{g}_t^{(t)})\|_2$$

$$\leq \|\tilde{g}_i^{(t)} + C(g_i^{(t+1)} - \hat{g}_t^{(t)}) - g_i^{(t)} + g_i^{(t+1)}\|_2$$

$$\leq \sqrt{\pi} \|g_i^{(t+1)} - g_i^{(t)}\|_2 + \|g_i^{(t+1)}\|_2$$

$$\leq \sqrt{\pi} (G + \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} G) + G$$

$$\leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} G := \tilde{G}, \quad (C.1)$$

where the first inequality holds due to triangle inequality and the second inequality holds by Assumption 4.1.

Then we bound the aggregated compressed gradient $\tilde{g}_t$:

$$\|\tilde{g}_t\|_2 = \left\|\frac{1}{n} \sum_{i=1}^{n} \tilde{g}_i^{(t)}\right\|_2 \leq \frac{1}{n} \sum_{i=1}^{n} \left\|\tilde{g}_i^{(t)}\right\|_2 \leq \tilde{G}. \quad (C.2)$$

Then we prove the upper bound for compressed aggregated gradient $\tilde{g}$ on the server by induction. Similarly, for $t = 1$, we have $\|\tilde{g}_1\|_2 \leq \|C(\tilde{g}_1)\|_2 \leq \|\tilde{g}_1\|_2 \leq \tilde{G}$. For $t > 1$, suppose we have $\|\tilde{g}_t\|_2 \leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} G$, then we have

$$\|\tilde{g}_{t+1}\|_2 = \|\tilde{g}_t + C(\tilde{g}_{t+1} - \hat{g}_t)\|_2$$

$$\leq \|\tilde{g}_t + C(\tilde{g}_{t+1} - \hat{g}_t) - \tilde{g}_t + \tilde{g}_{t+1}\|_2 + \|\tilde{g}_{t+1}\|_2$$

$$\leq \sqrt{\pi} \|\tilde{g}_{t+1} - \tilde{g}_t\|_2 + \|\tilde{g}_{t+1}\|_2$$

$$\leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} \tilde{G} := \tilde{G}. \quad (C.3)$$

Next we bound $\|m_t\|_2$ by induction. We prove this by induction. For $t = 0$, we have $\|m_0\|_2 = 0 \leq \tilde{G}$. For $t \geq 1$, suppose that $\|m_t\|_2 \leq \tilde{G}$, then for $m_{t+1}$, we have

$$\|m_{t+1}\|_2 = \|\beta_1 m_t + (1 - \beta_1) \tilde{g}_{t+1}\|_2$$

$$\leq \beta_1 \|m_t\|_2 + (1 - \beta_1) \|\tilde{g}_{t+1}\|_2$$

$$\leq \beta_1 \tilde{G} + (1 - \beta_1) \tilde{G}$$

$$= \tilde{G}.$$  

Thus, for any $t \geq 0$, we have $\|m_t\|_2 \leq \tilde{G}$.

Finally we bound $\|\tilde{v}_t\|_2$ again by induction. For $t = 1$, we have $\|v_0\|_2 = \|\tilde{v}_0\|_2 = 0 \leq \tilde{G}^2$. For $t > 1$, suppose we
For the squared norm of compressed gradient, we have
\[ \|v_{t+1}\|_2 = \|\beta_2 v_t + (1 - \beta_2) \hat{g}_{t+1}\|_2 \]
\[ \leq \beta_2 \|v_t\|_2 + (1 - \beta_2) \|\hat{g}_{t+1}\|_2 \]
\[ \leq \beta_2 \hat{G}^2 + (1 - \beta_2) \hat{G}^2 \]
\[ = \hat{G}^2, \]
and by definition, we have \( \|\hat{v}_{t+1}\|_2 = \max\{\|\hat{v}_t\|_2, \|v_{t+1}\|_2\} \leq \hat{G}^2. \) Thus, for any \( t \geq 0, \) we have \( \|\hat{v}_t\|_2 \leq \hat{G}^2. \) The proof of \( \ell_\infty \)-boundedness is similar to the proof of \( \ell_2 \)-boundedness. This concludes the proof.

Proof of Lemma B.3. By definition, we have
\[ \mathbb{E}[\|g_t^{(i)} - \nabla f_i(x_t)\|_2^2] = \mathbb{E}[\left\| \frac{1}{\tau} \sum_{j \in S_t^i} g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2], \]
where \( S_t^i \) is a set of batch index with \( \tau \) samples on the \( i \)-th worker, then for \( \forall j \in S_t^i, \) we have
\[
\begin{align*}
\mathbb{E}
&\left[ \left\| \frac{1}{\tau} \sum_{j \in S_t^i} g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2 \right] \\
&= \frac{1}{\tau^2} \mathbb{E}
\left[ \sum_{j,j' \in S_t^i} [g_t^{(i,j)} - \nabla f_i(x_t)]^\top [g_t^{(i,j')} - \nabla f_i(x_t)] \right] \\
&= \frac{1}{\tau^2} \mathbb{E}
\left[ \sum_{j \neq j' \in S_t^i} [g_t^{(i,j)} - \nabla f_i(x_t)]^\top [g_t^{(i,j')} - \nabla f_i(x_t)] \right] + \frac{1}{\tau^2} \mathbb{E}
\left[ \left\| g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2 \right] \\
&= \frac{\tau - 1}{\tau(N - 1)} \mathbb{E}
\left[ \sum_{j \neq j' \in [N]} [g_t^{(i,j)} - \nabla f_i(x_t)]^\top [g_t^{(i,j')} - \nabla f_i(x_t)] \right] + \frac{1}{\tau} \mathbb{E}
\left[ \left\| g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2 \right] \\
&\leq \frac{\tau - 1}{\tau (N - 1)} \mathbb{E}
\left[ \left\| g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2 \right] \\
&\leq \frac{\tau - 1}{\tau (N - 1)} \sigma^2, \quad (C.4)
\end{align*}
\]
where the second equality holds by the sampling strategy, the third one follows by \( \mathbb{P}\{j, j' \in S_t^i\} = \frac{\tau(\tau - 1)}{N(N - 1)}. \) The inequality holds due to Assumption 6.3. Therefore, we have
\[ \mathbb{E}
\left[ \left\| \frac{1}{\tau} \sum_{j \in S_t^i} g_t^{(i,j)} - \nabla f_i(x_t) \right\|_2^2 \right] \leq \frac{N - \tau}{\tau (N - 1)} \sigma^2. \]
This concludes the proof.

Proof of Lemma B.4. For the squared norm of compressed gradient, we have
\[
\begin{align*}
\mathbb{E}[\|\hat{g}_t\|_2^2] &= \mathbb{E}[\|\hat{g}_t - \tilde{g}_t + \tilde{g}_t - g_t + g_t - \nabla f(x_t) + \nabla f(x_t)\|_2^2] \\
&\leq 4\mathbb{E}[\|\hat{g}_t - \nabla f(x_t)\|_2^2] + 4\mathbb{E}[\|\tilde{g}_t - \hat{g}_t\|_2^2] + 4\mathbb{E}[\|\tilde{g}_t - g_t\|_2^2] + 4\mathbb{E}[\|\nabla f(x_t)\|_2^2]. \quad (C.5)
\end{align*}
\]
For the stochastic variance, we have
\[ \mathbb{E} \| \hat{g}_t - \nabla f(x_t) \|^2 = \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} [g_t^{(i)} - \nabla f_i(x_t)] \right\|^2 \right] \]
\[ = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \| g_t^{(i)} - \nabla f_i(x_t) \|^2 \]
\[ = \frac{1}{n} \mathbb{E} \| g_t^{(i)} - \nabla f_i(x_t) \|^2 \]
\[ \leq \frac{\sigma^2 (N - \tau)}{n \tau (N - 1)} \]
\[ \leq \frac{\sigma^2}{n}, \tag{C.6} \]
where the first inequality holds by Lemma B.3 and the second one holds by $1 \leq \tau \leq N$. For the squared norm of gradient compression error, we have
\[ \mathbb{E} \| \hat{g}_t - g_t \|^2 \leq \mathbb{E} \| \hat{g}_t \|^2 + \mathbb{E} \| g_t \|^2 \]
\[ \leq 2 \bar{G}^2 + 2 G^2, \tag{C.7} \]
where the second inequality follows by Lemma B.2. Similarly, for the squared norm of gradient compression error on the server side, we have
\[ \mathbb{E} \| \hat{g}_t - \hat{g}_t \|^2 \leq \mathbb{E} \| \hat{g}_t \|^2 + \mathbb{E} \| \hat{g}_t \|^2 \]
\[ \leq 2 \bar{G}^2 + 2 G^2, \tag{C.8} \]
where the second inequality follows by Lemma B.2. Summing over $t = 1, \ldots, T$, we have
\[ \sum_{t=1}^{T} \mathbb{E} \| \hat{g}_t \|^2 \leq 4 T (2 \bar{G}^2 + 2 \bar{G}^2) + 4 T (2 \bar{G}^2 + 2 G^2) + \frac{4 T \sigma^2}{n} + 4 \sum_{t=1}^{T} \mathbb{E} \| \nabla f(x_t) \|^2 \]
\[ \leq 32 T \bar{G}^2 + \frac{4 T \sigma^2}{n} + 4 \sum_{t=1}^{T} \mathbb{E} \| \nabla f(x_t) \|^2, \tag{C.9} \]
where the second inequality follows by $G \leq \bar{G} \leq \bar{G}$. By the updating rule, we have
\[ \mathbb{E} \| m_t \|^2 = \mathbb{E} \left[ \left\| (1 - \beta_1) \sum_{\tau=1}^{t} \beta_1^{t-\tau} \tilde{g}_\tau \right\|^2 \right] \]
\[ \leq (1 - \beta_1)^2 \sum_{i=1}^{d} \mathbb{E} \left[ \left( \sum_{\tau=1}^{t} \beta_1^{t-\tau} \tilde{g}_{\tau,i} \right)^2 \right] \]
\[ \leq (1 - \beta_1)^2 \sum_{i=1}^{d} \mathbb{E} \left[ \left( \sum_{\tau=1}^{t} \beta_1^{t-\tau} \right) \left( \sum_{\tau=1}^{t} \beta_1^{t-\tau} \tilde{g}_{\tau,i} \right) \right] \]
\[ \leq (1 - \beta_1) \sum_{\tau=1}^{t} \beta_1^{t-\tau} \mathbb{E} \| \tilde{g}_\tau \|^2 \]
\[ \leq 32 \bar{G}^2 + \frac{4 \sigma^2}{n} + 4 (1 - \beta_1) \sum_{\tau=1}^{t} \beta_1^{t-\tau} \mathbb{E} \| \nabla f(x_t) \|^2, \tag{C.10} \]
where the second inequality holds by applying Cauchy-Schwarz inequality, and the third inequality holds by summation of series. Hence summing over $t = 1, \cdots, T$, we have
\[ \sum_{t=1}^{T} \mathbb{E} \| m_t \|^2 \leq 32 T \bar{G}^2 + \frac{4 T \sigma^2}{n} + 4 \sum_{t=1}^{T} \mathbb{E} \| \nabla f(x_t) \|^2, \tag{C.11} \]
Therefore, with the condition \( \alpha_t = \alpha \) in Lemma B.4, we have

\[
\sum_{t=1}^{T} \alpha_t^2 \mathbb{E} \left[ \left\| \nabla f_t \right\|^2 \right] \leq \frac{32T \alpha^2 \tilde{G}^2}{\nu} + \frac{4T \alpha^2 \sigma^2}{\nu n} + \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(x_t) \right\|^2 \right].
\]

(C.12)

We also have

\[
\sum_{t=1}^{T} \alpha_t^2 \mathbb{E} \left[ \left\| \nabla f_t \right\|^2 \right] \leq \frac{32T \alpha^2 \tilde{G}^2}{\nu} + \frac{4T \alpha^2 \sigma^2}{\nu n} + \frac{4\alpha^2}{\nu} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(x_t) \right\|^2 \right].
\]

(C.13)

\[\square\]

**Proof of Lemma B.5** Considering on the \( i \)-th worker, we have

\[
\mathbb{E} \left[ \left\| \tilde{g}_t^{(i)} - g_t^{(i)} \right\| \right] = \mathbb{E} \left[ \left\| \tilde{g}_{t-1}^{(i)} + C(g_t^{(i)} - \tilde{g}_{t-1}^{(i)}) - g_t^{(i)} \right\| \right]
\]

\[
\leq \sqrt{\pi} \mathbb{E} \left[ \left\| g_t^{(i)} - \tilde{g}_{t-1}^{(i)} \right\| \right]
\]

\[
\leq \sqrt{\pi} \mathbb{E} \left[ \left\| g_t^{(i)} - g_t^{(i)} \right\| \right] + \sqrt{\pi} \mathbb{E} \left[ \left\| g_t^{(i)} - \tilde{g}_{t-1}^{(i)} \right\| \right],
\]

(C.14)

where the first inequality holds because of Assumption 4.1. For the compression error between the local stochastic gradient and the gradient of loss function, by Lemma B.3 we have

\[
\mathbb{E} \left[ \left\| g_t^{(i)} - \nabla f_t(x_t) \right\| \right] = \frac{N - \tau}{\tau (N - 1)} \sigma^2.
\]

Since for all variables \( X \), there is \( [\mathbb{E}X]^2 \leq \mathbb{E}X^2 \), we have

\[
\mathbb{E} \left[ \left\| g_t^{(i)} - \nabla f_i(x_t) \right\| \right] \leq \mathbb{E} \left[ \left\| g_t^{(i)} - \nabla f_i(x_t) \right\| \right] \leq \sigma \sqrt{\frac{N - \tau}{\tau (N - 1)}}.
\]

Then the expectation of gradient compression error on all workers is

\[
\mathbb{E} \left[ \left\| g_t - g_t \right\| \right] = \mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} g_t^{(i)} - \frac{1}{n} \sum_{i=1}^{n} g_t^{(i)} \right\| \right]
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_t^{(i)} - g_t^{(i)} \right\| \right]
\]

\[
\leq \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_{t-1}^{(i)} - g_{t-1}^{(i)} \right\| \right] + \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_t^{(i)} - g_t^{(i)} \right\| \right]
\]

\[
\leq \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_{t-1}^{(i)} - g_{t-1}^{(i)} \right\| \right] + \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| \nabla f_i(x_t) - \nabla f_i(x_{t-1}) \right\| \right]
\]

\[
+ \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_t^{(i)} - \nabla f_i(x_t) \right\| \right] + \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_t^{(i)} - \nabla f_i(x_{t-1}) \right\| \right]
\]

\[
\leq \sqrt{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| g_{t-1}^{(i)} - g_{t-1}^{(i)} \right\| \right] + \sqrt{\pi} L \left\| x_t - x_{t-1} \right\| + 2\sqrt{\pi} \sigma \sqrt{\frac{N - \tau}{\tau (N - 1)}},
\]

(C.15)

where the first inequality holds by extended triangle inequality, the second inequality holds by applying (C.14).
and the last inequality holds by $L$-smoothness. Telescoping the right hand side of (C.15), we have

$$
\mathbb{E}[\|\mathbf{g}_t - \mathbf{g}_t\|_2] \leq \sqrt{\frac{\pi}{n}} \sum_{i=1}^{n} \mathbb{E}[\|\mathbf{g}^{(i)}_{t-1} - \mathbf{g}^{(i)}_{t-1}\|_2] + \sqrt{\pi} L \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 + 2\sqrt{\pi} \sigma \sqrt{\frac{N - \tau}{\pi(N-1)}}
$$

$$
\leq L \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j} \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2 + 2\sigma \sqrt{\frac{N - \tau}{\pi(N-1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j}
$$

$$
= L \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j} \|\mathbf{m}_{j-1}\|_2 + 2\sigma \sqrt{\frac{N - \tau}{\pi(N-1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j}
$$

$$
\leq L \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j} \|\mathbf{V}_{j-1}\|_2 + 2\sigma \sqrt{\frac{N - \tau}{\pi(N-1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j}
$$

$$
\leq G \sum_{j=2}^{t} (\sqrt{\pi})^{t+1-j} \|\mathbf{V}_{j-1}\|_2 + 2\sigma \sqrt{\frac{N - \tau}{\pi(N-1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j},
$$

(C.16)

where the second inequality holds by $L$-smooth and the recursion (C.15). The third inequality holds by taking iteration of $\|\mathbf{g}_t - \mathbf{g}_t\|$. The forth inequality holds by the update rule of $\mathbf{x}$ and for matrix $\mathbf{A}$ and vector $\mathbf{x}$, we have $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2$. The last inequality holds because we set the initial momentum $\mathbf{m}_0 = 0$ and by Lemma B.2 there is an upper bound $\|\mathbf{m}_t\|_2 \leq G$.

Since $\alpha_j \mathbf{V}_{j-1}^{1/2} \geq \alpha_j \mathbf{V}_{j-1}^{1/2} \geq 0$, we have

$$
\mathbb{E}[\|\mathbf{g}_t - \mathbf{g}_t\|_2] \leq LG \sum_{j=2}^{t} (\sqrt{\pi})^{t+1-j} \|\mathbf{V}_{j-1}^{1/2}\|_2 + 2\sigma \sqrt{\frac{N - \tau}{\pi(N-1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j}
$$

$$
\leq \frac{\sqrt{\pi}}{1 - \sqrt{\pi}} LG \alpha_1 \sqrt{d
u} - \sqrt{\nu} \sqrt{\frac{N - \tau}{n\tau(N-1)}},
$$

(C.17)

where the last inequality holds because $\|\mathbf{V}_{j-1}^{1/2}\|_2 \leq \sqrt{d
u} - \sqrt{\nu}$. This completes the proof.

**Proof of Lemma 4.6.** By Assumption 4.1 we have

$$
\mathbb{E}[\|\mathbf{g}_t - \mathbf{g}_t\|_2] \leq \mathbb{E}[\|\mathbf{g}_{t-1} + C(\mathbf{g}_t - \mathbf{g}_{t-1}) - \mathbf{g}_t\|_2]
$$

$$
\leq \sqrt{\pi} \mathbb{E}[\|\mathbf{g}_t - \mathbf{g}_{t-1}\|_2]
$$

$$
\leq \sqrt{\pi} \mathbb{E}[\|\mathbf{g}_{t-1} - \mathbf{g}_{t-1}\|_2] + \sqrt{\pi} \mathbb{E}[\|\mathbf{g}_t - \mathbf{g}_{t-1}\|_2]
$$

$$
\leq \sum_{j=1}^{t} (\sqrt{\pi})^{t+1-j} \mathbb{E}[\|\mathbf{g}_j - \mathbf{g}_{j-1}\|_2],
$$

(C.18)

the last inequality holds by telescoping the right hand side of the previous inequality, with the similar approach
in (C.16), we have
\[
E[\|\tilde{g}_j - \tilde{g}_{j-1}\|_2] = E\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_j^{(i)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{g}_{j-1}^{(i)} \right\|_2 \right]
\leq \frac{1}{n} \sum_{i=1}^{n} E[\|C(\tilde{g}_j^{(i)}) - \tilde{g}_{j-1}^{(i)}\|_2]
\leq \frac{1}{n} \sum_{i=1}^{n} E[\|\tilde{g}_j^{(i)} - \tilde{g}_{j-1}^{(i)}\|_2]
\leq \frac{1 + \sqrt{\pi}}{n} \sum_{i=1}^{n} E[\|\tilde{g}_j^{(i)} - \tilde{g}_{j-1}^{(i)}\|_2] + \frac{1 + \sqrt{\pi}}{n} \sum_{i=1}^{n} E[\|\tilde{g}_{j-1}^{(i)} - \tilde{g}_j^{(i)}\|_2],
\]

where we apply Jensen’s inequality to get the first inequality. The second inequality holds due to Lemma B.1, the third inequality holds by triangle inequality, then similar to previous proof, we have
\[
E[\|\tilde{g}_j - \tilde{g}_{j-1}\|_2] \leq \frac{1 + \sqrt{\pi}}{n} \sum_{i=1}^{n} E[\|\tilde{g}_j^{(i)} - \tilde{g}_{j-1}^{(i)}\|_2] + \frac{1 + \sqrt{\pi}}{n} \sum_{i=1}^{n} E[\|\tilde{g}_{j-1}^{(i)} - \tilde{g}_j^{(i)}\|_2]
\leq \frac{(1 + \sqrt{\pi})(1 - \sqrt{\pi})}{1 - \sqrt{\pi}} L\tilde{G}\alpha_1 \|\tilde{V}_{1/2}\|_2 + \frac{1 + \sqrt{\pi}(1 - \sqrt{\pi})}{1 - \sqrt{\pi}} \cdot 2\sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}}
\leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} L\tilde{G}\alpha_1 \|\tilde{V}_{1/2}\|_2 + \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} \cdot 2\sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}},
\]

where inequalities hold by proof of Lemma B.5. Substituting (C.19) into (C.18)
\[
E[\|\tilde{g}_t - \tilde{g}_t\|_2] \leq \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} L\tilde{G}\alpha_1 \cdot \|\tilde{V}_{1/2}\|_2 \sum_{j=1}^{t} (\sqrt{\pi})^{j-1} + \frac{1 + \sqrt{\pi}}{1 - \sqrt{\pi}} \cdot 2\sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{j-1}
\leq \frac{\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} L\tilde{G}\alpha_1 \|\tilde{V}_{1/2}\|_2 + \frac{\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} \cdot 2\sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}} \sum_{j=1}^{t} (\sqrt{\pi})^{j-1}
\leq \frac{\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} L\tilde{G}\alpha_1 \|\tilde{V}_{1/2}\|_2 + \frac{\sqrt{\pi}(1 + \sqrt{\pi})}{(1 - \sqrt{\pi})^2} \cdot 2\sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}},
\]

where the last inequality holds because \|\tilde{V}_{1/2}\|_2 \leq \sqrt{d\nu^{-1/2}}. This concludes the proof. \(\square\)

D Dependency on the Compression Constant

We track the impact of the compressor related parameter \(\pi\). Recall in Assumption 4.1, we assume we have compressors satisfy \(E_C[\|C(x) - x\|_2^2] \leq \pi x^2, \forall x \in \mathbb{R}^4\). Note that \(\pi = 0\) leads to \(C(x) = x\). In fact, a stronger compression means a larger \(\pi\), i.e., \(\pi \rightarrow 1\), may leads to worse convergence. For our convergence bound
\[
\min_{t \in [T]} E[\|\nabla f(x_t)\|_2^2] \leq \frac{1}{T} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|_2^2]
\leq \frac{M_1}{T\alpha} + \frac{M_2}{T} + \alpha M_3 + \frac{\alpha M_4 \sigma^2}{n} + M_5 \sigma \sqrt{\frac{N - \tau}{\tau(N - 1)}},
\]

(D.1)
where

\[ M_1 = C \cdot \Delta f, \quad M_2 = \frac{CG\tilde{G}}{(1 - \beta_1)\sqrt{\nu}}, \quad M_3 = \frac{32CC_1\tilde{G}^2}{\nu} + \frac{2\sqrt{\pi}CLCG\tilde{G}\sqrt{d}}{\nu(1 - \sqrt{\pi})^2}, \]

\[ M_4 = \frac{4CC_1}{\nu}, \quad M_5 = \frac{4\sqrt{C}G\sqrt{d}}{\sqrt{\nu}(1 - \sqrt{\pi})^2}, \]

and

\[ \Delta f = f(x_1) - \inf_x f(x), \tilde{G} = C_2G, \tilde{G}_\infty = C_2G_\infty, \]

\[ C = 2(\tilde{G}_\infty^2 + \nu)^{1/2}, C_1 = 2L + 3L \left( \frac{\beta_1}{1 - \beta_1} \right)^2, C_2 = \frac{(1 + \sqrt{\pi})^2}{(1 - \sqrt{\pi})^2}. \]

Following by Lemma B.2 as \( \pi \to 1 \), we conclude the series of compressed gradient bounds \( \tilde{G}, \tilde{G}_\infty, \tilde{G}_\infty \) as follows

\[ \tilde{G} = 1 + \sqrt{\pi} \frac{G}{1 - \sqrt{\pi}} = O \left( \frac{G}{1 - \pi} \right), \quad \tilde{G}_\infty = 1 + \sqrt{\pi} \frac{G_\infty}{1 - \sqrt{\pi}} = O \left( \frac{G_\infty}{1 - \pi} \right), \]

\[ \tilde{G} = C_2G = (1 + \sqrt{\pi})^2 \frac{G}{(1 - \sqrt{\pi})^2} = O \left( \frac{G}{(1 - \pi)^2} \right), \quad \tilde{G}_\infty = C_2G_\infty = (1 + \sqrt{\pi})^2 \frac{G_\infty}{(1 - \sqrt{\pi})^2} = O \left( \frac{G_\infty}{(1 - \pi)^2} \right), \]

\[ C = 2(\tilde{G}_\infty^2 + \nu)^{1/2} = O(\tilde{G}_\infty) = O \left( \frac{G_\infty}{(1 - \pi)^2} \right). \]

<table>
<thead>
<tr>
<th>Const</th>
<th>Order of ( \pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>( O \left( \Delta f G_\infty(1 - \pi)^{-2} \right) )</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>( O \left( G^2 G_\infty(1 - \pi)^{-4} \right) )</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>( O \left( G^2 G_\infty(1 - \pi)^{-6} + L\sqrt{d}G^2 G_\infty(1 - \pi)^{-6} \right) )</td>
</tr>
<tr>
<td>( M_4 )</td>
<td>( O \left( G_\infty(1 - \pi)^{-2} \right) )</td>
</tr>
<tr>
<td>( M_5 )</td>
<td>( O \left( G\tilde{G}(1 - \pi)^{-4} \right) )</td>
</tr>
<tr>
<td>( T )</td>
<td>( O \left( M_1 M_3 \epsilon^{-2} + M_1 M_4 \sigma^2 \epsilon^{-2} + M_2 \epsilon^{-1} \right) )</td>
</tr>
</tbody>
</table>

Table 1: Theorem constants with dependency on \( \pi \).

Table 1 summarizes the dependency of the theorem constant on \( \pi \). Note that here the total number of iterates \( T \) scales as \( (1 - \pi)^{-8} \), which is higher than that of distributed SGD (Philippenko and Dieuleveut, 2020, 2021). This is mainly due to the fact that the convergence of adaptive gradient methods and the theoretical proof of adaptive gradient methods rely heavily on the bounded gradient assumption. While for communication-compressed adaptive methods, the gradient bound for compressed gradient \( \tilde{g} \) and \( \tilde{g} \) is already dependent on \( \pi \) as shown in the above derivations. In fact, our dependency on gradient bounds matches the existing works studying vanilla adaptive gradient methods (Zhou et al., 2018; Chen et al., 2019), which suggests that our bound is indeed tight.

Note that in practice, the high dependency on \( \pi \) does not make much difference. In fact, the actual \( \pi \) computed using real compressor is in a constant order. For example, in our ResNet-18 experiment with scaled sign compressor, the actual \( \pi \) ranges from \([0.597, 0.713]\) during the entire training procedure, which is indeed a reasonable constant.
E Additional Experiments

In this section, we present the additional experiments of our proposed method. In Section E.1, we provide experiments using the nonconvex logistic regression with Top-k compressor. In Section E.2, we present additional deep learning experiments on other popular state-of-the-art model architectures.

E.1 Nonconvex Logistic Regression Based on Top-k Compressor

The nonconvex logistic regression problem setting follows exactly the same as in Section 7.1. Here we apply the proposed Markov compression sequence based on the Top-k compressor, where $k = 1$ for the nonconvex logistic regression with $d = 300$ parameters.

![Figure 4](image)

Figure 4: Gradient norm comparison of different compression strategies on nonconvex logistic regression trained by AMSGrad with Top-1 based Markov compression sequence. The upper row is the norm convergence with respect to the communication cost, and the lower row is with respect to the training iteration.

Figure 4 shows that our proposed CD-Adam achieved the best performances for all four datasets against the other three compression strategies. Specifically, we can observe that CD-Adam achieves a much better result than AMSGrad with error feedback and naive compression in terms of both communication efficiency and the final performances. Furthermore, compared with uncompressed AMSGrad, our communication efficiency is greatly improved while maintaining similar final performances.

E.2 Deep Learning Experiments on Image Classification

In this part, we provide additional deep learning experiments on other state-of-the-art model architectures. The problem setting follows exactly the same as in Section 7.2. In the following, we show that our proposed algorithm enjoys a fast convergence speed with high accuracy and less communication cost compared with EF21 and 1-bit Adam.

![Figure 5](image)

Figure 5, 7 and 9 compare the performance with respect to the communication cost using different model architectures. The experiments show that our proposed method obtains better training/testing loss and accuracy, especially much better training/testing gradient norm on various CNN model architectures, including ResNet, VGG, and WideResNet. Notice that under this comparison, EF21, which is also equipped with Markov compression sequences on SGD, performs worse than our method, especially in later training stages. While for 1-bit Adam, since it needs an uncompressed warm-up procedure for a few epochs, the communication efficiency is much worse compared with CD-Adam and EF21.
Figure 5 compares the performance with respect to the iteration epochs using different model architectures. Compared with EF21, our proposed method obtains a much better convergence speed in terms of training/testing loss, accuracy, and the gradient norm. Our method also shows a better performance than 1-bit Adam in terms of the final testing accuracy and gradient norm.

Figure 5 is the continuous of Figure 3. It shows that our proposed CD-Adam achieves better performance than EF21 and 1-bit Adam under the same communication cost.

In Figure 6, the left two plots show that at the early stage of the training process, CD-Adam and EF21 obtain a similar speed of reducing the training loss, while 1-bit Adam is quite faster due to the warm-up process. At later stages, CD-Adam shows a clear advantage compared with EF21. Our method also shows better performance on testing loss than 1-bit Adam. From the middle two plots in Figure 6, we can observe that our proposed CD-Adam obtains a smaller gradient norm than EF21 and 1-bit Adam under the same epochs when training.
and testing. The right two plots show that our proposed method achieves high accuracy on both the training and testing sets. At the early stage of the training process, CD-Adam and EF21 obtain a similar speed of increasing the accuracy, while 1-bit Adam is faster due to the warm-up process. At later stages, CD-Adam shows a clear advantage compared with EF21. Our method shows better performance on testing accuracy compared with 1-bit Adam.

Figure 7: Comparison on training/testing loss, gradient norm and accuracy with respect to communication cost among our proposed method and the baselines when training VGG-16 model on CIFAR10.

In Figure 7, the left two plots show that our proposed method obtains better training/testing loss than EF21 and 1-bit Adam, though EF21 intermediately shows a faster decrease in the loss function. From the middle two plots, we can observe that our proposed CD-Adam obtains smaller training/testing gradient norms than EF21 and 1-bit Adam under the same communication budget. The right two plots show that CD-Adam obtains better training/testing accuracy than EF21 and 1-bit Adam, though EF21 intermediately shows a faster increase in accuracy as well.

Figure 8: Comparison on training/testing loss, gradient norm and accuracy with respect to epochs among our proposed method and the baselines when training VGG-16 model on CIFAR10.
In Figure 8, the left two plots show that our proposed method achieves better training/testing loss than EF21, though EF21 intermediate shows a faster decrease in the loss function. Our method shows better performance on testing loss compared with 1-bit Adam. The middle two plots in Figure 8 show that our proposed CD-Adam obtains a smaller gradient norm than EF21 and 1-bit Adam under the same epochs. The right two plots show that our proposed method achieves high accuracy on both the training and testing sets. It shows that our proposed method achieves better training/testing loss than EF21, though EF21 intermediate shows a faster increase in accuracy. Our CD-Adam also shows a better testing accuracy than 1-bit Adam.

Figure 9: Comparison on training/testing loss, gradient norm and accuracy with respect to communication cost among our proposed method and the baselines when training WideResNet-16-4 model on CIFAR10.

In Figure 9, the left two plots and the right two plots show that our proposed method maintains a significant advantage on training/testing loss and accuracy compared with EF21 and 1-bit Adam. The middle two plots show that our proposed CD-Adam achieves smaller gradient norms than EF21 and 1-bit Adam. In addition, 1-bit Adam initially shows a lower gradient norm while its gradient norm diverges later.

Figure 10: Comparison on training/testing loss, gradient norm and accuracy with respect to epochs among our proposed method and the baselines when training WideResNet-16-4 model on CIFAR10.
In Figure 10, the left two plots and right two plots show that our proposed method achieves better training/testing loss and accuracy than EF21. Furthermore, CD-Adam still achieves better performance on testing loss and testing accuracy than 1-bit Adam, while 1-bit Adam keeps a slighter better performance on training loss and training accuracy. The middle two plots in Figure 9 show that CD-Adam achieves better training/testing gradient norms compared with EF21 and 1-bit Adam.

E.3 Additional results

Average runtime and total bits: The overhead on compression as well as the worker-side update is quite small, as shown from Table 2. Note that EF21 takes more time since it adopts top-$k$ which is harder to compute compared to scaled sign compressor.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg Time (s/iter)</th>
<th>Total Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncompressed</td>
<td>1.015</td>
<td>$32d \times 2T$</td>
</tr>
<tr>
<td>EF21</td>
<td>1.402</td>
<td>$\approx (32k \times 2) \times 2T$</td>
</tr>
<tr>
<td>1-bit Adam</td>
<td>1.041</td>
<td>$32d \times 2T + (32 + d) \times 2(T - T_1)$</td>
</tr>
<tr>
<td>CD-Adam</td>
<td>1.134</td>
<td>$(32+d) \times 2T$</td>
</tr>
</tbody>
</table>

Table 2: Average runtime and total bits

Ablation on $n$ and $\tau$: The left plot in Figure 11 shows the effect of the total amount of workers $n$ on the convergence of training loss. It shows that a larger $n$ leads to a faster decrease on training loss but not absolutely leads to a better convergence rate. The right plot in Figure 11 shows the effect of batch size $\tau$ on the convergence of training loss. We observe that a larger number of batch size achieves a faster convergence rate.

Figure 11: Ablation on $n$ and $\tau$