BEER: Fast $O(1/T)$ Rate for Decentralized Nonconvex Optimization with Communication Compression

Haoyu Zhao\(^*\)  Boyue Li\(^†\)  Zhize Li\(^‡,§\)  Peter Richtárik\(^‡\)  Yuejie Chi\(^†\)
Princeton  CMU  KAUST  KAUST  CMU

January 2022

Abstract

Communication efficiency has been widely recognized as the bottleneck for large-scale decentralized machine learning applications in multi-agent or federated environments. To tackle the communication bottleneck, there have been many efforts to design communication-compressed algorithms for decentralized nonconvex optimization, where the clients are only allowed to communicate a small amount of quantized information (aka bits) with their neighbors over a predefined graph topology. Despite significant efforts, the state-of-the-art algorithm in the nonconvex setting still suffers from a slower rate of convergence $O((G/T)^{2/3})$ compared with their uncompressed counterpart, where $G$ measures the data heterogeneity across different clients, and $T$ is the number of communication rounds. This paper proposes BEER, which adopts communication compression with gradient tracking, and shows it converges at a faster rate of $O(1/T)$. This significantly improves over the state-of-the-art rate, by matching the rate without compression even under arbitrary data heterogeneity. Numerical experiments are also provided to corroborate our theory and confirm the practical superiority of BEER in the data heterogeneous regime.

Keywords: decentralized nonconvex optimization, communication compression, fast rate.

1 Introduction

Decentralized machine learning is gaining attention in both academia and industry because of its emerging applications in multi-agent systems such as the internet-of-things (IoT) and networked autonomous systems (Marvasti et al., 2014; Savazzi et al., 2020). One of the key problems in decentralized machine learning is on-device training, which aims to optimize a machine learning model using the datasets stored on (geographically) different clients, and can be formulated as a decentralized optimization problem.

Decentralized optimization aims to solve the following optimization problem without sharing the local datasets with other clients:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} f(x; \mathcal{D}_i) \right\},$$

where $f(x; \mathcal{D}_i) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} f(x; \xi_i)$ for $i \in [n]$, and $n$ is the total number of clients. Here, $x \in \mathbb{R}^d$ is the machine learning model, $f(x; \mathcal{D})$, $f(x; \mathcal{D}_i)$, and $f(x; \xi_i)$ denote the loss functions of the model $x$ on the entire dataset $\mathcal{D}$, the local dataset $\mathcal{D}_i$, and a random data sample $\xi_i$, respectively. Different from the widely studied distributed or federated learning setting where there is a central server to coordinate the parameter sharing across all clients, in the decentralized setting, each client can only communicate with its neighbors over a communication network determined by a predefined network topology.

\(^*\)Department of Computer Science, Princeton University, Princeton, NJ 08540, USA; Email: haoyu@princeton.edu.
\(^†\)Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Emails: {boyuel,yuejiec}@andrew.cmu.edu.
\(^‡\)Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia; Email: {zhize.li,peter.richtarik}@kaust.edu.sa.
\(^§\)Corresponding author.
The main bottleneck of decentralized optimization—when it comes to large-scale machine learning applications—is communication efficiency, due to the large number of clients involved in the network (Savazzi et al., 2020) and the enormous size of machine learning models (Brown et al., 2020), exacerbated by resource constraints such as limited bandwidth availability and stringent delay requirements. One way to reduce the communication cost is communication compression, which only transmits compressed messages (with fewer bits) between the clients using compression operators. The compression operators come with many design choices and offer great flexibility in different trade-offs of communication and computation in practice. Even though communication compression has been extensively applied to distributed or federated optimization with a central server (Stich et al., 2018; Das et al., 2020; Karimireddy et al., 2019; Gorbunov et al., 2021; Richtárík et al., 2021; Li and Richtárík, 2020; Fatkhullin et al., 2021; Li and Richtárík, 2021), its use in the decentralized setting has been relatively sparse. Most of the existing approaches only apply to the strongly convex setting (Reisizadeh et al., 2019; Koloskova et al., 2019b; Liu et al., 2020; Kovalev et al., 2021; Liao et al., 2021; Li et al., 2021b), and only a few consider the general nonconvex setting (Koloskova et al., 2019a; Tang et al., 2019; Singh et al., 2021).

1.1 Our contributions

This paper considers decentralized optimization with communication compression, focusing on the nonconvex setting due to its critical importance in modern machine learning, such as training deep neural networks (LeCun et al., 2015), word embeddings, and other unsupervised learning models (Saunshi et al., 2019).

Unfortunately, existing algorithms (Koloskova et al., 2019a; Tang et al., 2019; Singh et al., 2021) suffer from several important drawbacks in the nonconvex setting: they need strong bounded gradient or bounded dissimilarity assumptions to guarantee convergence, and the convergence rate is order-wise slower than their uncompressed counterpart in terms of the communication rounds (see Table 1).

In this paper, we introduce BEER, which is a decentralized optimization algorithm with communication compression using gradient tracking. BEER not only removes the strong assumptions required in all prior works, but enjoys a faster convergence rate in the nonconvex setting. Concretely, we have the following main contributions (see Tables 1 and 2).

1. We show that BEER converges at a fast rate of $O(1/T)$ in the nonconvex setting, which improves over the state-of-the-art rate $O(1/T^{2/3})$ of CHOCO-SGD (Koloskova et al., 2019a) and Deepsqueeze (Tang et al., 2019), where $T$ is the number of communication rounds. This matches the rate without compression even under arbitrary data heterogeneity across the clients.

2. We also provide the analysis of BEER under the Polyak-Łojasiewicz (PL) condition (Assumption 2.5), and show that BEER converges at a linear rate (see Table 2). Note that strong convexity implies the PL condition, and thus BEER also achieves linear convergence in the strongly convex setting.

3. We run numerical experiments on real-world datasets and show BEER achieves superior or competitive performance when the data are heterogeneous compared with state-of-the-art baselines with and without communication compression.

To the best of our knowledge, BEER is the first algorithm that achieves $O(1/T)$ rate without the bounded gradient or bounded dissimilarity assumptions, supported by a strong empirical performance in the data heterogeneous setting.

1.2 Related works

In this section, we review closely related literature on decentralized optimization, communication-efficient algorithms, and communication compression.

Decentralized optimization Decentralized optimization, which is a special class of linearly constrained (consensus constraint) optimization problems, has been studied for a long time (Bertsekas and Tsitsiklis, 2015; Glowinski and Marrocco, 1975). Many centralized algorithms can be intuitively converted into decentralized counterparts by using gossip averaging (Kempe et al., 2003; Xiao and Boyd, 2004), which mixes parameters from neighboring clients to enforce consensus.
### Table 1: Comparison of convergence rates for existing decentralized methods with communication compression in the nonconvex setting.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Convergence rate</th>
<th>Strong assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuARM-SGD</td>
<td>$O \left( \frac{1}{\sqrt{\tau T}} + \frac{nG^2}{T} \right)$</td>
<td>Bounded Gradient</td>
</tr>
<tr>
<td>DeepSqueeze</td>
<td>$O \left( \left( \frac{G}{T} \right)^{2/3} \right)$</td>
<td>Bounded Dissimilarity</td>
</tr>
<tr>
<td>CHOCO-SGD</td>
<td>$O \left( \left( \frac{G}{T} \right)^{2/3} \right)$</td>
<td>Bounded Gradient</td>
</tr>
<tr>
<td>BEER (Algorithm 1)</td>
<td>$O \left( \frac{1}{T} \right)$</td>
<td>—</td>
</tr>
</tbody>
</table>

### Table 2: Summary of the established convergence rates for the proposed BEER algorithm in the nonconvex setting.

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Convergence rate</th>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$ is $L$-smooth</td>
<td>$\frac{1}{T} \sum_{t=0}^{T-1} E \left| \nabla f(\bar{x}^t) \right|^2 \leq \frac{2(\Phi_0 - \Phi_T)}{\eta T}$</td>
<td>Theorem 4.1</td>
</tr>
<tr>
<td>$f_i$ is $L$-smooth</td>
<td>$\Phi_T \leq (1 - \mu \eta)^T \Phi_0$</td>
<td>Theorem 4.3</td>
</tr>
</tbody>
</table>

### Communication-efficient algorithms

While decentralized optimization is a classical topic, the focus on communication efficiency is relatively new due to the advances in large-scale machine learning. Roughly speaking, there are primarily two kinds of approaches to reduce communication cost: 1) local methods: in each communication round, clients run multiple local update steps before communicating, in the hope of reducing the number of communication rounds; 2) compressed methods: clients send compressed communication messages, in the hope of reducing the communication cost per communication round.

Both categories have received significant attention in recent years. For local methods, a small sample of examples include FedAvg (McMahan et al., 2017), Local-SVRG (Gorbunov et al., 2020), SCAFFOLD (Karimireddy et al., 2020) and FedPAGE (Zhao et al., 2021b). On the other hand, many compressed methods are proposed recently such as Alistarh et al. (2017), Khirirat et al. (2018), Tang et al. (2018a), Stich et al. (2018), Koloskova et al. (2019a, 2020), Gorbunov et al. (2021), Li and Richtárik (2021), Richtárik et al. (2021), Fatkhullin et al. (2021), Zhao et al. (2021a). In this paper, we will adopt the second approach based on communication compression to enhance communication efficiency.
Decentralized nonconvex optimization with compression  As discussed earlier and summarized in Table 1, there have been limited existing works on decentralized nonconvex optimization with communication compression. In particular, SQARM-SGD [Singh et al., 2021] can be viewed as CHOCO-SGD with momentum, but its theoretical convergence rate is slower than the original CHOCO-SGD. Deepsqueeze [Tang et al., 2019] and CHOCO-SGD have a close relationship, where Deepsqueeze can be viewed as decentralized SGD (DSGD or D-PSGD) with the explicit error feedback framework, and CHOCO-SGD uses control variables to implicitly handle the compression error.

Notation  Throughout this paper, we use boldface letters to denote vectors, e.g., \( \mathbf{x} \in \mathbb{R}^d \). Let \([n]\) denote the set \( \{1, 2, \cdots, n\} \), \( \mathbf{1} \) be the all-one vector, \( \mathbf{I} \) be the identity matrix, \|\cdot\| denote the Euclidean norm of a vector, and \( \|\cdot\|_F \) denote the Frobenius norm of a matrix. Let \( \langle \mathbf{u}, \mathbf{v} \rangle \) denote the standard Euclidean inner product of two vectors \( \mathbf{u} \) and \( \mathbf{v} \). In addition, we use the standard order notation \( O(\cdot) \) to hide absolute constants.

2 Problem Setup

In this section, we formally define the decentralized optimization problem with communication compression, and introduce a few important quantities and assumptions that will be used in developing our algorithm and theory.

2.1 Decentralized optimization

The goal of decentralized optimization is to solve

\[
\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) \right\},
\]

where \( n \) is the number of clients, \( f(\mathbf{x}) \) is the global objective function, and \( f_i(\mathbf{x}) := f(\mathbf{x}; D_i) := \mathbb{E}_{\xi_i \sim D_i} f(\mathbf{x}; \xi_i) \) is the local objective function, with \( \mathbf{x} \) the parameter of interest, \( \xi_i \) a random data sample drawn from the local dataset \( D_i \).

In the decentralized setting, the clients can only communicate with their local neighbors over a prescribed network topology, which is specified by an undirected weighted graph \( \mathcal{G}([n], E) \). Here, each node in \([n]\) represents a client, and \( E \) is the set of possible communication links between different clients. Information sharing across the clients is implemented mathematically by the use of a mixing matrix \( \mathbf{W} = [w_{ij}] \in [0, 1]^{n \times n} \), which is defined in accordance with the network topology: we assign a positive weight \( w_{ij} \) for any \((i, j) \in E\) and \( w_{ij} = 0 \) for all \((i, j) \notin E\). We make the following standard assumption on the mixing matrix (Nedić et al., 2018).

Assumption 2.1 (Mixing matrix). The mixing matrix \( \mathbf{W} = w_{ij} \in [0, 1]^{n \times n} \) is symmetric (\( \mathbf{W}^\top = \mathbf{W} \)) and doubly stochastic (\( \mathbf{W} \mathbf{1} = \mathbf{1} \mathbf{1}^\top \mathbf{W} = \mathbf{1}^\top \)). Let its eigenvalues be \( 1 = |\lambda_1(\mathbf{W})| > |\lambda_2(\mathbf{W})| \geq \cdots \geq |\lambda_n(\mathbf{W})| \). The spectral gap is denoted by

\[
\rho := 1 - |\lambda_2(\mathbf{W})| \in (0, 1).
\]

The spectral gap of a mixing matrix is closely related to the network topology, see Nedić et al. (2018) for its scaling with respect to the network size (i.e. the number of clients \( n \)) for representative network topologies.

2.2 Compression operators

Compression, in the forms of quantization or sparsification, can be used to reduce the total communication cost. We now introduce the notion of a randomized compression operator, which is widely used in the decentralized/federated optimization literature, e.g. Tang et al. (2018a), Stich et al. (2018), Koloskova et al. (2019a); Richtárik et al. (2021).

Definition 2.2 (Compression operator). A randomized map \( \mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d \) is an \( \alpha \)-compression operator if for all \( \mathbf{x} \in \mathbb{R}^d \), it satisfies

\[
\mathbb{E} \left[ \|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2 \right] \leq (1 - \alpha) \|\mathbf{x}\|^2.
\]
In particular, no compression (\(C(x) \equiv x\)) implies \(\alpha = 1\).

Compared with the *unbiased compression operator* used in, e.g., Alistarh et al. (2017); Khirrat et al. (2018); Mishchenko et al. (2019), the compression operator in Definition 2.2 does not impose the additional constraint on the expectation such that \(E[C(x)] = x\). Besides, it is always possible to convert an *unbiased compression operator* into a biased one satisfying Definition 2.2 (Richtárik et al., 2021). In particular, for an unbiased compression operator \(C : \mathbb{R}^d \mapsto \mathbb{R}^d\) that satisfies
\[
E[C'(x)] = x, \quad E \left[ \|C'(x) - x\|^2 \right] \leq \omega \|x\|^2,
\]
we can construct a *biased* compression operator \(C : C(x) = \frac{C'(x)}{1+\omega}\) and the new compression operator satisfies Definition 2.2 with \(\alpha = \frac{1}{1+\omega}\). Thus, Definition 2.2 is a generalization of the unbiased compression operator that allows biased compression.

### 2.3 Assumptions on functions

We now state the assumptions on the functions \(\{f_i\}\) and \(f\). Throughout this paper, we assume that \(f^* = \min_x f(x)\) exists and \(f^* > -\infty\).

In the nonconvex setting, we assume that the functions \(\{f_i\}_{i \in [n]}\) are arbitrary functions that satisfy the following standard smoothness assumption.

**Assumption 2.3** (Smoothness). The function \(f\) is \(L\)-smooth if there exists \(L \geq 0\) such that
\[
\|\nabla f(x_1) - \nabla f(x_2)\| \leq L \|x_1 - x_2\|, \forall x_1, x_2 \in \mathbb{R}^d.
\]

In addition, we allow local computation to be performed via stochastic gradient updates, where \(\bar{\nabla} f_i(x) := \nabla f_i(x; \xi_i)\) denotes a local stochastic gradient computed via a sample \(\xi_i\) drawn i.i.d. from \(D_i\), and \(\bar{\nabla}_b f_i(x) := \frac{1}{b} \sum_{j=1}^b \nabla f_i(x; \xi_{i,j})\) denotes the stochastic gradient computed by a minibatch with size \(b\) drawn i.i.d. from \(D_i\).

We assume \(\bar{\nabla} f_i(x)\) and \(\bar{\nabla}_b f_i(x)\) have bounded variance, which is again standard in the decentralized/federated optimization literature (McMahan et al., 2017; Karimireddy et al., 2020; Koloskova et al., 2019a).

**Assumption 2.4** (Bounded variance). There exists a constant \(\sigma \geq 0\) such that for all \(i \in [n]\) and \(x \in \mathbb{R}^d\),
\[
E \left\| \bar{\nabla} f_i(x) - \nabla f_i(x) \right\|^2 \leq \sigma^2.
\]

For a stochastic gradient with minibatch size \(b\), we have
\[
E \left\| \bar{\nabla}_b f_i(x) - \nabla f_i(x) \right\|^2 \leq \frac{\sigma^2}{b}.
\]

In addition, we consider the setting when the function \(f\) additionally satisfies the following Polyak-Łojasiewicz (PL) condition (Polyak, 1963), which can lead to fast linear convergence even when the function is nonconvex.

**Assumption 2.5** (PL condition). There exists some constant \(\mu > 0\) such that for any \(x \in \mathbb{R}^d\),
\[
\|\nabla f(x)\|^2 \geq 2\mu (f(x) - f^*).
\]

Note that the PL condition is a weaker assumption than strong convexity, which means that if the objective function \(f\) is \(\mu\)-strongly convex, then the PL condition also holds with the parameter \(\mu\).

### 3 Proposed Algorithm

In this section, we introduce our proposed algorithm **BEER** for decentralized nonconvex optimization with compressed communication. Before embarking on the description of **BEER**, we introduce some convenient matrix notation. Since in a decentralized setting, the parameter estimates at different clients are typically
Algorithm 1 BEER: Better compression for dEcentRalized optimization

1: **Input:** Initial point $X^0 = x_0 1^T$, $G^0 = 0$, $H^0 = 0$, $V^0 = \nabla F(X_0)$, step size $\eta$, mixing step size $\gamma$, minibatch size $b$.
2: for $t = 0, 1, \ldots$ do
3: $X^{t+1} = X^t + \gamma H^t (W - I) - \eta V^t$
4: $Q^t_i = C(X^{t+1} - H^t)$  // client $i$ sends $Q^t_i$ to all its neighbors
5: $H^{t+1} = H^t + Q^t_i$
6: $V^{t+1} = V^t + \gamma G^t (W - I) + \tilde{\nabla}_b F(X^{t+1}) - \tilde{\nabla}_b F(X^t)$  // client $i$ sends $Q^t_i$ to all its neighbors
7: $G^{t+1} = G^t + Q^t_i$
8: end for

Different, we use $X = [x_1, x_2, \ldots, x_n]$ to denote the collection of parameter estimates from all clients, where $x_i$ is from client $i$. The average of $\{x_i\}_{i \in [n]}$ is denoted by $\bar{x} := \frac{1}{n} X 1$. Other quantities are defined similarly. With slight abuse of notation, we define

$$\nabla F(X) := [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)] \in \mathbb{R}^{d \times n},$$

which collects the local gradients computed at the local parameters. Similarly, the stochastic variant is defined as

$$\tilde{\nabla}_b F(X) := [\tilde{\nabla}_b f_1(x_1), \tilde{\nabla}_b f_2(x_2), \ldots, \tilde{\nabla}_b f_n(x_n)].$$

We now proceed to describe BEER, which is detailed in Algorithm 1. Because the matrix notations introduced above. At the $t$-th iteration, BEER maintains the current model estimates $X^t$ and the global gradient estimates $V^t$ across the clients. At the core of its design, BEER also tracks and maintains two control sequences $H^t$ and $G^t$ that serve as compressed surrogates of $X^t$ and $V^t$, respectively. In particular, these two control sequences are updated by aggregating the received compressed messages alone (cf. Line 4 and Line 5).

It then boils down to how to carefully update these quantities in each iteration with communication compression. To begin, note that for each client $i$, BEER not only maintains its own parameters $\{x^t_i, v^t_i, h^t_i, g^t_i\}$, but also the control variables from its neighbors, namely, $\{h^t_j\}_{j \in \mathcal{N}(i)}$ and $\{g^t_j\}_{j \in \mathcal{N}(i)}$.

**Update the model estimate** Each client $i$ first updates its model $x^{t+1}_i$ according to Line 3. By thinking of $\{h^t_j\}_{j \in \mathcal{N}(i)}$ as a surrogate of $\{x^t_j\}_{j \in \mathcal{N}(i)}$, the second term aims to achieve better consensus among the clients through mixing, while the last term performs a gradient descent update.

**Update the global gradient estimate** Each client $i$ updates the global gradient estimate $v^{t+1}_i$ according to Line 6, where the last correction term—based on the difference of the gradients at consecutive models—is known as a trick called gradient tracking (Qu and Li [2017]; Di Lorenzo and Scutari [2016]; Nedic et al. [2017]). The use of gradient tracking is critical: as will be seen momentarily, it contributes to the key difference from CHOCO-SGD that enables the fast rate of $O(1/T)$ without any bounded dissimilarity or bounded gradient assumptions. Indeed, if we remove the control sequence $G^t$ and substitute Lines 5 and 6 by $V^{t+1} = \tilde{\nabla}_b F(X^{t+1})$, we recover CHOCOSGD from BEER.

**Update the compressed surrogates with communication** To update $\{h^t_j\}_{j \in \mathcal{N}(i)}$, each client $i$ first computes a compressed message $q^{t+1}_{h,i}$ that encodes the difference $x^{t+1}_i - h^t_i$, and broadcasts it to its neighbors (cf. Line 4). Then, each client $i$ updates $\{h^t_j\}_{j \in \mathcal{N}(i)}$ by aggregating the received compressed messages $\{q^{t+1}_{h,j}\}_{j \in \mathcal{N}(i)}$ following Line 5. The updates of $\{g^t_j\}_{j \in \mathcal{N}(i)}$ can be performed similarly. Moreover, all the compressed messages can be sent in a single communication round at one iteration, i.e., the communications in Lines 4 and 5 can be performed at once.

The BEER algorithm leverages EF21 (Richtárik et al. [2021]) for communication compression, which is a better and simpler algorithm that deals with biased compression operators compared with the error feedback (or error compensation, EF/EC) framework (Karimireddy et al. [2019]; Stich and Karimireddy [2020]). BEER does not need to apply EF/EC explicitly and can deal with the error implicitly using the control sequence $G^t$. 


Therefore, BEER is also much simpler than EF-C-GT (Liao et al., 2021), which is another recently proposed algorithm (for the strongly convex setting) using EF/EC compression and gradient tracking simultaneously.

4 Convergence Guarantees

In this section, we show the convergence guarantees of BEER under different settings: the $O(1/T)$ rate in the nonconvex setting in Section 4.1, and the improved linear rate under the PL condition (Assumption 2.5) in Section 4.2. In Section 4.3, we briefly sketch the proof.

Our convergence guarantees are based on an appropriately designed Lyapunov function, given by

$$\Phi_t = \mathbb{E} f(\bar{x}^t) - f^* + \frac{c_1 L}{n} \Omega_1^t + \frac{c_2 \rho^2}{nL} \Omega_2^t + \frac{c_3 L}{n} \Omega_3^t + \frac{c_4 \rho^4}{nL} \Omega_4^t,$$

where the choice of constants $\{c_i\}_{i=1}^4$ might be different from theorem to theorem, $\mathbb{E} f(\bar{x}^t) - f^*$ represents the sub-optimality gap, and the errors $\{\Omega_i^t\}_{i=1}^4$ are defined by

$$\begin{align*}
\Omega_1^t &:= \mathbb{E} \left\| H^t - X^t \right\|_F^2, \\
\Omega_2^t &:= \mathbb{E} \left\| G^t - V^t \right\|_F^2, \\
\Omega_3^t &:= \mathbb{E} \left\| X^t - \bar{x}^t 1^\top \right\|_F^2, \\
\Omega_4^t &:= \mathbb{E} \left\| V^t - \bar{v}^t 1^\top \right\|_F^2.
\end{align*}$$

Here, $\Omega_1^t$ and $\Omega_2^t$ denote the compression errors for $X^t$ and $V^t$ when approximated using the compressed surrogates $H^t$ and $G^t$ respectively, and $\Omega_3^t$ and $\Omega_4^t$ denote the consensus errors of $X^t$ and $V^t$.

4.1 Convergence in the nonconvex setting

First, we present the following convergence result of BEER in the nonconvex setting where there is no local variance ($\sigma^2 = 0$), i.e., we can use the local full gradient $\nabla F(X^t)$ instead of $\nabla_b F(X^t)$ in Line 6 of Algorithm 1.

**Theorem 4.1** (Convergence in the nonconvex setting without local variance). Suppose Assumptions 2.1 and 2.3 hold, and we can compute the local full gradient $\nabla f_i(x)$ for any $x$. Then there exist absolute constants $c_1, c_2, c_3, c_4, c_5 > 0$, such that if we set $\gamma = c_4 \alpha \rho$, $\eta = c_5 \gamma \rho^2 / L$, then for the Lyapunov function $\Phi_t$ in (4), it holds

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(\bar{x}^t) \right\|^2 \leq \frac{2(\Phi_0 - \Phi_T)}{\eta T}.$$  

Theorem 4.1 shows that BEER converges at a rate of $O(1/T)$ when there is no local variance ($\sigma^2 = 0$), which is faster than the $O(1/T^{2/3})$ rate by CHOCO-SGD (Koloskova et al., 2019a) and DeepSqueeze (Tang et al., 2019), and the $O(1/\sqrt{T})$ rate by SQuARM-SGD (Singh et al., 2021); see also Table 1.

More specifically, to achieve $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(\bar{x}^t) \right\|^2 \leq \epsilon^2$, BEER needs $O \left( \frac{1}{\rho^2 \alpha^2} \right)$ iterations or communication rounds, where $\rho$ and $\alpha$ are the spectral gap (cf. [2]) and the compression parameter (cf. [3]), respectively. In comparison, the state-of-the-art algorithm CHOCO-SGD (Koloskova et al., 2019a) converges at a rate of $O((G/\rho^2 \alpha T)^{2/3})$, which translates to an iteration complexity of $O \left( \frac{G}{\rho^2 \alpha^2 T} \right)$, with $G$ being the bounded gradient parameter, namely, $\mathbb{E} \xi_i \left\| \nabla f(x_i) \right\|^2 \leq G^2$. Therefore, BEER improves over CHOCO-SGD not only in terms of a better dependency on $\epsilon$, but also removing the bounded gradient assumption, which is significant since in practice, $G$ can be excessively large due to data heterogeneity across the clients.

The dependency on $\alpha$ of BEER is consistent with other compression schemes, such as CHOCO-SGD, DeepSqueeze and SQuARM-SGD for the nonconvex setting, as well as LEAD (Liu et al., 2020) and EF-C-GT (Liao et al., 2021) for the strongly convex setting.

As for the dependency on $\rho$, BEER is slightly worse than CHOCO-SGD, where CHOCO-SGD has a dependency of $O(1/\rho^2)$ whereas BEER has a dependency of $O(1/\rho^3)$. This degeneration is also seen in the analysis of...
uncompressed decentralized algorithms using gradient tracking \cite{sun2020separation,xin2020convergent}, where the rate \(O(1/\rho^2)\) is worse than the rate of \(O(1/\rho)\) for basic decentralized SGD algorithms \cite{kempe2003convergence,liang2017coordinate} by a factor of \(\rho\). In addition, both BEER and CHOCO-SGD use small mixing step size \(\gamma\) to guarantee convergence, which makes the dependency on \(\rho\) worse than their uncompressed counterparts.

**Stochastic gradient oracles** BEER also supports the use of stochastic gradient oracles with bounded local variance (Assumption 2.4). More specifically, we have the following theorem, which generalizes Theorem 4.1.

**Theorem 4.2** (Convergence in the nonconvex setting). Suppose Assumptions 2.1, 2.3, and 2.4 hold. Then there exist absolute constants \(c_1, c_2, c_3, c_4, c_5, \gamma_0 > 0\), such that if we set \(\gamma = c_1 \alpha \rho, \eta = c_5 \gamma_0 \rho^2 / L\), then for the Lyapunov function \(\Phi_t\) in (4), it holds

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{x}_t) \|^2 \leq \frac{2(\Phi_0 - \Phi_T)}{\eta T} + \frac{36c_4 \sigma^2}{c_5 b \alpha L}.
\]

In the presence of local variance, the squared gradient norm of BEER has an additional term that scales on the order of \(O\left(\frac{\sigma^2}{\eta T}\right)\) (ignoring other parameters). By choosing a sufficiently large minibatch size \(b\), i.e. \(b \geq O\left(\frac{\sigma^2}{\eta T}\right)\), BEER maintains the iteration complexity

\[
O\left(\frac{1}{\rho^3 \alpha \epsilon^2}\right)
\]

to reach \(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{x}_t) \|^2 \leq \epsilon^2\), without the bounded gradient assumption, thus inheriting similar advantages over CHOCO-SGD as discussed earlier. In terms of local computation, the gradient oracle complexity on a single client of BEER is

\[
O\left(\frac{1}{\rho^3 \alpha \epsilon^2} + \frac{\sigma^2}{\rho^3 \alpha^2 \epsilon^4}\right).
\]

### 4.2 Linear convergence with PL condition

Now, we show that the convergence of BEER can be strengthened to a linear rate with the addition of the PL condition (Assumption 2.5). Similar to the nonconvex setting, we first show the convergence result without local gradient variance (\(\sigma^2 = 0\)).

**Theorem 4.3** (Convergence under the PL condition without local variance). Suppose Assumptions 2.1, 2.3, and 2.5 hold, and we can compute the local full gradient \(\nabla f_i(x)\) for any \(x\). Then there exist constants \(c_1, c_2, c_3, c_4, c_5, c_\gamma > 0\), such that if we set \(\gamma = c_1 \alpha \rho, \eta = c_5 \gamma \rho^2 / L\), then for the Lyapunov function \(\Phi_t\) in (4), it holds

\[
\Phi_T \leq (1 - \mu \eta)^T \Phi_0.
\]

Theorem 4.3 demonstrates that under the PL condition, BEER converges linearly to the global optimum \(f^*\), where it finds an \(\epsilon\)-optimal solution in \(O\left(\frac{b \log(\frac{1}{\epsilon})}{\mu \rho \alpha}\right)\) iterations.

**Stochastic gradient oracles** Under the PL condition, BEER also supports the use of stochastic gradient oracles with bounded local variance (Assumption 2.4). The following theorem shows that BEER linearly converges to a neighborhood of size \(O\left(\frac{\sigma^2}{\eta \rho^2}\right)\) around the global optimum.

**Theorem 4.4** (Convergence under PL condition). Suppose Assumptions 2.1, 2.3, 2.4, and 2.5 hold. Then there exist absolute constants \(c_1, c_2, c_3, c_4, c_\gamma, c_\eta > 0\), such that if we set \(\gamma = c_1 \alpha \rho, \eta = c_\eta \gamma \rho^2 / L\), then for the Lyapunov function \(\Phi_t\) in (4), it holds

\[
\Phi_T \leq (1 - \mu \eta)^T \Phi_0 + \frac{36c_4 \sigma^2}{c_\gamma L \rho \alpha}.
\]
4.3 Proof sketch

We now provide a proof sketch of Theorem 4.1, which establishes the $O(1/T)$ rate of BEER in the nonconvex setting using full gradient, highlighting the technical reason of the rate improvement of BEER over CHOCO-SGD.

Recalling the quantities $\Omega_1^t$ to $\Omega_4^t$ from (5), which capture the approximation errors using compression and the consensus errors of $X^t$ and $V^t$, we would like to control these errors by obtaining inequalities of the form:

$$\Omega_i^{t+1} \leq (1 - a_i)\Omega_i^t + b_i, \quad \forall i \in \{1, 2, 3, 4\},$$

where $0 < a_i < 1$ denotes the size of the contraction, and $b_i > 0$ wraps together other terms, which may depend on $\Omega_j^t$ for $j \neq i$, as well as the expected squared gradient norm of $\bar{v}^t$, i.e.,

$$\Omega_5^t = \mathbb{E} \|\bar{v}^t\|^2.$$  \hspace{1cm} (6)

Then, by choosing the Lyapunov function properly (cf. (4)), we can show that the Lyapunov function actually descends, and small manipulations lead to the claimed convergence results in Theorem 4.1.

We now explain briefly how gradient tracking helps in BEER. Note that CHOCO-SGD also has the control variable $H^t$ for the model $X^t$, therefore, in its analysis, it also deals with the quantities $\Omega_1^t$ and $\Omega_4^t$. However, CHOCO-SGD also needs to bound the term $\|V^t\|^2_F$, where $V^t = \nabla F(X^t)$ for CHOCO-SGD when using full gradients. Thus, CHOCO-SGD needs to assume the bounded gradient assumption and only obtain a slower $O(1/T^{2/3})$ convergence rate. In contrast, BEER deals with the term $\|V^t\|^2_F$ by decomposing it using Young’s inequality, leading to

$$\|V^t\|^2_F \leq (1 + \beta)\Omega_4^t + (1 + 1/\beta)\Omega_5^t$$

for some $\beta > 0$. Here, $\Omega_4^t$ can be controlled via the gradient tracking technique (see Line 6 in Algorithm 4) without the bounded gradient assumption, and $\Omega_5^t$ can be handled using the smoothness assumption (Assumption 2.3).

5 Numerical Experiments

This section presents numerical experiments on real-world datasets to showcase BEER’s superior ability to handle data heterogeneity across the clients, by running each experiment on unshuffled datasets and comparing the performances with the state-of-the-art baseline algorithms.

5.1 Experiment setup

We run experiments on two nonconvex problems: logistic regression with a nonconvex regularizer (Wang et al., 2018) on the a9a dataset (Chang and Lin, 2011), and training 1-hidden layer neural network on the MNIST dataset (LeCun et al., 1995).

For logistic regression with a nonconvex regularizer, following Wang et al. (2018), the objective function over a datum $(a, b)$ is defined as

$$f(x; (a, b)) = \log (1 + \exp(-ba^\top x)) + \alpha \sum_{j=1}^{d} \frac{x_j^2}{1 + x_j^2},$$

where the last term is the nonconvex regularizer and the regularization parameter is set to $\alpha = 0.05$.

For 1-hidden layer neural network training, we use 32 hidden neurons, sigmoid activation functions and cross-entropy loss. The objective function over a datum $(a, b)$ is defined as

$$f(x; (a, b)) = \ell(\text{softmax}(W_2 \text{sigmoid}(W_1 a + c_1) + c_2), b),$$

where $\ell(\cdot, \cdot)$ denotes the cross-entropy loss, the optimization variable is collectively denoted by $x = \text{vec}(W_1, c_1, W_2, c_2)$, where the dimensions of the network parameters $W_1, c_1, W_2, c_2$ are $64 \times 784, 64 \times 1, 10 \times 64$, and $10 \times 1$, respectively.

For both experiments, we split the unshuffled datasets evenly to 10 clients that are connected by a ring topology. By using unshuffled data, we can simulate the scenario with high data heterogeneity across clients.
Figure 1: The training gradient norm and testing accuracy against communication rounds (top two panels) and communication bits (bottom two panels) for logistic regression with nonconvex regularization on unshuffled a9a dataset. Both BEER and CHOCO-SGD employ the biased gsgd\(_b\) compression (Alistarh et al., 2017) with \(b = 5\).

Approximately, for the a9a dataset, 5 clients receive data with label 1 and others receive data with label 0; for the MNIST dataset, client \(i\) receives data with label \(i\). We use the FDLA matrix (Xiao and Boyd, 2004) as the mixing matrix to perform weighted information aggregation to accelerate convergence.

5.2 Comparisons with state-of-the-art algorithms

We compare BEER with 1) CHOCO-SGD (Koloskova et al., 2019), which is the state-of-the-art nonconvex decentralized optimizing algorithm using communication compression, and 2) DSGD (Lian et al., 2017) and \(D^2\) (Tang et al., 2018), which are decentralized optimization algorithms without compression. All algorithms are initialized in the same experiment by the same initial point. Moreover, we use the same best-tuned learning rate \(\eta = 0.1\), batch size \(b = 100\), and biased compression operator (gsgd\(_b\)) (Alistarh et al., 2017) for BEER and CHOCO-SGD on both experiments.

Figure 1 and Figure 2 plot the training gradient norm and testing accuracy against communication rounds and communication bits for logistic regression with nonconvex regularization and 1-hidden-layer neural network training, respectively.

In the nonconvex logistic regression experiment (cf. Figure 1), the algorithms with communication compression (BEER and CHOCO-SGD) converge faster than the uncompressed algorithms (DSGD and \(D^2\)) in terms of the communication bits. However, CHOCO-SGD fails to converge to a small gradient norm solution since it cannot tolerate a high level of data dissimilarity across different clients, and its performance is not...
5.3 Impacts of network topology and compression schemes on BEER

We further investigate the impact of network topology and compression schemes on the performance of BEER. We follow the setup in Section 5.1 and run logistic regression with nonconvex regularization ($\alpha = 0.05$) on the unshuffled a9a dataset, by splitting it evenly to 40 agents. All experiments use the same best-tuned step size $\eta = 0.5$, batch size $b = 100$, and $\gamma = 0.7$, to guarantee fast convergence.

comparable to $D^2$. In contrast, BEER can converge to a point with a relatively smaller gradient norm, which is comparable to $D^2$. The performance of testing accuracy is similar to that of the training gradient norm, where BEER achieves the best testing accuracy and also learns faster than the uncompressed algorithms.

Turning to the neural network experiment (cf. Figure 2), in terms of the final training gradient norm, BEER converges to a solution comparable to $D^2$, but at a lower communication cost, while CHOCO-SGD and DSGD cannot converge due to the data heterogeneity. In terms of testing accuracy, BEER and $D^2$ have very similar performance, and outperform CHOCO-SGD and DSGD.

In summary, BEER has much better performance in terms of communication efficiency than CHOCO-SGD in heterogeneous data scenario, which corroborates our theory. BEER also performs similarly or even better than the uncompressed baseline algorithm $D^2$, and much better than DSGD. In addition, by leveraging different communication compression schemes, BEER allows more flexible trade-offs between communication and computation, making it an appealing choice in practice.
Impacts of network topology  Figure 3 shows the training gradient norm and testing accuracy of BEER with respect to the communication rounds over different network topologies using the \textit{gsgd}_b compression (Alistarh et al. 2017): ring topology ($\rho = 0.022$), star topology ($\rho = 0.049$), grid topology ($\rho = 0.063$), and Erdős-Rényi topology with connectivity probability $p = 0.5$ and $p = 0.2$ ($\rho = 0.51$ and $\rho = 0.77$, respectively). Despite the huge difference of the spectral gaps $\rho$ of different topologies, BEER can use nearly the same hyper-parameters to obtain similar performance. The experiment results complement our theoretical analysis and show that BEER may converge way better in practice despite its cubic dependency of $1/\rho$ in Theorem 4.2.

Impacts of compression schemes  Figure 4 shows the training gradient norm and testing accuracy of BEER with respect to the communication rounds and communication bits on a ring topology using different compression schemes: no compression, \textit{gsgd}_5, \textit{top}_{10} and \textit{random}_{10} (see Appendix A for their formal definitions). The parameters within the compression operators are chosen such that BEER with different compression operators transfer similar amount of bits per communication round. All experiments use the same best-tuned step size $\eta = 0.5$, batch size $b = 100$ and $\gamma = 0.7$, except that we use $\eta = 0.005$ and $\gamma = 0.8$ for \textit{top}_{10} compression. In terms of communication bits, it can be seen that the use of compression operators improves over the uncompressed baseline, in the sense that, all algorithms with compression converge to a solution with lower gradient norm and higher testing accuracy at a lower communication cost. In terms of communication rounds, different compression operators can lead to significantly behaviors. For example, BEER with \textit{gsgd}_5 and \textit{random}_{10} compression operators converge faster than BEER without compression, but BEER with the \textit{top}_{10} compression operator converges slower than BEER without compression. Among all experiments, BEER with \textit{gsgd}_5 reaches the highest final testing accuracy, while behaves similar to BEER without compression in terms of communication rounds, which implies that \textit{gsgd}_b may be a more practical compression operator, at least more suitable for BEER.

6 Conclusion

This paper presents BEER, which achieves a faster $O(1/T)$ convergence rate for decentralized nonconvex optimization with communication compression, \textit{without} imposing the bounded dissimilarity or bounded gradient assumptions. In addition, a faster linear rate of convergence is established for BEER under the PL condition. Numerical experiments are provided to corroborate our theory on the advantage of BEER in the data heterogeneous scenario. An interesting direction of future work is to investigate the lower bounds for decentralized (nonconvex) optimization with communication compression.
Figure 4: The training gradient norm and testing accuracy against communication rounds (top two panels) and communication bits (bottom two panels) for BEER using different compression schemes for logistic regression with nonconvex regularization on unshuffled a9a dataset.

Acknowledgements

The work of H. Zhao is supported in part by NSF, ONR, Simons Foundation, DARPA and SRC through awards to S. Arora. The work of B. Li and Y. Chi is supported in part by ONR N00014-19-1-2404, by AFRL under FA8750-20-2-0504, and by NSF under CCF-1901199 and CCF-2007911. The work of Z. Li and P. Richtárik is supported by KAUST Baseline Research Fund.

References


A Examples of Compression Operators

We provide some examples of compression operators satisfying Definition 2.2 that are used in our experiments.
\( \text{gsgd}_b \) \cite{Alistarh2017} \( \text{gsd}_b : \mathbb{R}^d \rightarrow \mathbb{R}^d \) \((b > 1)\), or random dithering, is a compression operator satisfying the following formula

\[
\text{gsd}_b(x) := \frac{1}{\tau} \cdot |x| \cdot \text{sign}(x) \cdot \left( 2^{(b-1)|x|} + u \right),
\]

where \( \tau = 1 + \min \left\{ \frac{d}{2^{d-1} - 1}, \frac{\sqrt{d}}{2^{d-1} - 1} \right\} \), and \( u \) is the random dithering vector uniformly sampled from \([0,1]^d\). It follows that \( \text{gsd}_b \) satisfies Definition 2.2 with \( \alpha = 1/\tau \).

\text{random}_k \cite{Wangni2018} \( \text{random}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d \), or random sparsification, is a compression operator satisfying the following formula

\[
\text{random}_k(x) := x \odot u,
\]

where \( u \) is uniformly sampled from the following set \( \{ u' \in \{0,1\}^d \mid \|u\|_1 = k \} \) and \( \odot \) denotes the Hadamard product or element-wise product. In words, \( \text{random}_k(x) \) randomly keeps \( k \) coordinates of \( x \) and sets the remaining coordinates of \( x \) to 0. It follows that \( \text{random}_k(x) \) satisfies Definition 2.2 with \( \alpha = k/d \).

\text{top}_k \cite{Alistarh2018,Stich2018} \( \text{top}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is a compression operator satisfying the following formula

\[
\text{top}_k(x) := x \odot u(x),
\]

where \( u(x) \in \{0,1\}^d \) that satisfies \( \|u\|_1 = k \) and \( u_i = 1 \) for all \( i \in I \) such that \( |x_i| > |x_j| \) for any \( i \in I \) and \( j \in [d] \setminus I \). In words, \( \text{top}_k \) keeps the coordinates of \( x \) with the \( k \) largest absolute values, and sets the other coordinates to 0. It follows that \( \text{top}_k(x) \) satisfies Definition 2.2 with \( \alpha = k/d \).

\section*{B Proof of Main Theorems}

\subsection*{B.1 Technical preparation}

We first recall some classical inequalities that helps our derivation.

\textbf{Proposition B.1.} Let \( \{v_1, \ldots, v_\tau\} \) be a set of \( \tau \) vectors in \( \mathbb{R}^d \). Then, \( \forall \beta > 0 \), we have

\[
\langle v_i, v_j \rangle \leq \frac{\beta}{2} \|u\|^2 + \frac{1}{2\beta} \|v\|^2,
\]

(7)

\[
\|v_i + v_j\|^2 \leq (1 + \beta)\|v_i\|^2 + \left( 1 + \frac{1}{\beta} \right) \|v_j\|^2;
\]

(8)

\[
\left\| \sum_{i=1}^{\tau} v_i \right\|^2 \leq \tau \sum_{i=1}^{\tau} \|v_i\|^2.
\]

(9)

Here, (7) is referred as the Cauchy-Schwarz inequality, (8) and (9) are referred as Young’s inequality.

\textbf{Additional notation} \ The following notation will be used throughout our proof:

\[
\nabla F(X) := [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)], \quad \nabla F_b(X) := [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)],
\]

\[
\nabla F(x) := [\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_n(x)], \quad \nabla F_b(x) := [\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_n(x)],
\]

where \( \bar{x} := \frac{1}{n}X1 \) with \( X = [x_1, x_2, \ldots, x_n] \).
Properties of the mixing matrix

We make note of several useful properties of the mixing matrix in the following lemma.

**Lemma B.2.** Let $W$ be a mixing matrix satisfying Assumption 2.1 and has spectral gap $\rho$, then for any matrix $M \in \mathbb{R}^{d \times n}$ and $\bar{m} = \frac{1}{n}M\mathbf{1}$, we have

$$\|MW - \bar{m}\mathbf{1}\mathbf{1}^\top\|_F = \|MW - \bar{m}\mathbf{1}\mathbf{1}W\|_F \leq (1 - \rho)\|M - \bar{m}\mathbf{1}\mathbf{1}\mathbf{1}^\top\|_F.$$ (10)

In addition, for any $\gamma \in (0, 1]$, the matrix $\tilde{W} = I + \gamma(W - I)$ satisfies Assumption 2.1 with a spectral gap at least $\gamma \rho$.

**Proof.** The first claim follows from the spectral decomposition of $W$. Since $W$ is a doubly stochastic matrix, the largest absolute eigenvalue of $W$ is 1 and the corresponding eigenvector is $\mathbf{1}$. Let $v_2, \ldots, v_n$ be the eigenvectors of $W$ corresponding to the remaining eigenvalues. Then, we have

$$\|MW - \bar{m}\mathbf{1}\mathbf{1}^\top\|_F = \|MW - \bar{m}\mathbf{1}\mathbf{1}W\|_F = \sum_{i=1}^{n}\|W(m_i - \bar{m}_i\mathbf{1})\|^2,$$

where the first equality follows from $\mathbf{1}^\top W = \mathbf{1}^\top$, $m_i$ denotes the transpose of the $i$-th row of matrix $M$, and $\bar{m}_i$ denotes the average of $m_i$. Now we decompose $m_i - \bar{m}_i\mathbf{1}$ using the eigenvectors of $W$. Noting that $\mathbf{1}^\top (m_i - \bar{m}_i\mathbf{1}) = \mathbf{1}^\top m_i - 1\mathbf{1}^\top \frac{1}{n}\mathbf{1}^\top m_i = 0$,

and thus we can write

$$m_i - \bar{m}_i\mathbf{1} = \sum_{j=2}^{n} c_j v_j$$

for some $\{c_j\}_{j=2}^{n}$. Then, we have

$$\|W(m_i - \bar{m}_i\mathbf{1})\|^2 = \left\|W \sum_{j=2}^{n} c_j v_j\right\|^2 \leq (1 - \rho)^2 \sum_{j=2}^{n} c_j^2 \leq (1 - \rho) \sum_{j=2}^{n} c_j^2 = (1 - \rho)\|m_i - \bar{m}_i\mathbf{1}\|^2,$$

and we conclude the proof of this claim.

For the second claim, recall the fact that if $v$ is an eigenvector of $W$ corresponding to the eigenvalue $\lambda$, then $v$ is also an eigenvector of $\tilde{W}$ with the corresponding eigenvalue $(1 - \gamma) + \gamma \lambda$. This claim follows from simple computation based on this relation.

A key consequence of gradient tracking

Before diving in the proofs of the main theorems, we record a key property of gradient tracking. Specifically, we have the following lemma.

**Lemma B.3.** If $\bar{v}^0 = \frac{1}{n} \tilde{\nabla}_b F(X^0)\mathbf{1}$, then for any $t \geq 1$, we have

$$\bar{v}^t = \frac{1}{n} \tilde{\nabla}_b F(X^t)\mathbf{1},$$ (11)

and

$$\bar{x}^{t+1} = \bar{x}^t - \frac{n}{n} \tilde{\nabla}_b F(X^t)\mathbf{1}.$$ (12)

**Proof.** We first prove (11) by induction. For the base case ($t = 0$), the relation (11) is obviously true by the means of initialization. Now suppose that at the $t$-th iteration, the relation (11) is true, i.e.,

$$\bar{v}^t = \frac{1}{n} \tilde{\nabla}_b F(X^t)\mathbf{1},$$

and
then at the \((t+1)\)-th iteration, we have

\[
\hat{\theta}^{t+1} = \frac{1}{n} V^{t+1} \mathbf{1}
\]

\[
= \frac{1}{n} V^t \mathbf{1} + \frac{1}{n} \gamma \mathbf{G}'(W - I) \mathbf{1} + \frac{1}{n} \left( \nabla_b F(X^{t+1}) - \nabla_b F(X^t) \right) \mathbf{1}
\]

\[
= \frac{1}{n} V^t \mathbf{1} + \frac{1}{n} \left( \nabla_b F(X^{t+1}) - \nabla_b F(X^t) \right) \mathbf{1}
\]

\[
= \frac{1}{n} \nabla_b F(X^{t+1}) \mathbf{1}.
\]

where (13) follows from the update rule of BEER (cf. Line 3), the penultimate line follows from \(W \mathbf{1} = \mathbf{1}\), and the last line follows from the induction hypothesis at the \(t\)-th iteration. Thus the induction hypothesis is also true at the \((t+1)\)-th iteration, and we complete the proof of (11).

For (12), it follows from the update rule of BEER (cf. Line 3) that

\[
\tilde{x}^{t+1} = \tilde{x}^t + \frac{\gamma}{n} H'(W - I) \mathbf{1} - \frac{\eta}{n} V^t \mathbf{1}
\]

\[
= \tilde{x}^t - \eta T^t = \tilde{x}^t - \frac{\eta}{n} \nabla_b F(X^t) \mathbf{1},
\]

where the second line uses \(W \mathbf{1} = \mathbf{1}\) and (11).

\[\square\]

### B.2 Recursive relations of main errors

As mentioned previously, the proof is centered around controlling the following set of errors which we repeat below for convenience (cf. (3)),

- compression approximation error: \(\Omega_1 = \mathbb{E} \left\| H^t - X^t \right\|_F^2\), \(\Omega_2 = \mathbb{E} \left\| G^t - V^t \right\|_F^2\)
- consensus error: \(\Omega_3 = \mathbb{E} \left\| X^t - \bar{x}^t \mathbf{1}^\top \right\|_F^2\), \(\Omega_4 = \mathbb{E} \left\| V^t - \tilde{\theta}^t \mathbf{1}^\top \right\|_F^2\)
- gradient norm: \(\Omega_5 = \mathbb{E} \left\| \theta^t \right\|^2\)

In particular, we aim to build a set of recursive relations of \(\Omega_1^t\) to \(\Omega_4^t\), which will be specified in the following lemma.

**Lemma B.4.** Suppose Assumptions \[2.1\], \[2.3\] and \[2.4\] hold, then for any \(t \geq 0\), we have

\[
\Omega_1^{t+1} \leq \left(1 - \frac{\alpha}{2} + \frac{6\gamma^2 C}{\alpha}\right) \Omega_1^t + \frac{6\gamma^2 C}{\alpha} \Omega_2^t + \frac{6\eta^2}{\alpha} \Omega_3^t + \frac{6n\eta^2}{\alpha} \Omega_5^t,
\]

\[
\Omega_2^{t+1} \leq \frac{18\gamma^2 CL^2}{\alpha} \Omega_1^t + \left(1 - \frac{\alpha}{2} + \frac{6\gamma^2 C}{\alpha}\right) \Omega_2^t + \frac{18\gamma^2 CL^2}{\alpha} \Omega_3^t + \frac{18L^2 \eta^2}{\alpha} \Omega_4^t + \frac{18L^2 \eta^2 n}{\alpha} \Omega_5^t + \frac{12n\sigma^2}{b\alpha},
\]

\[
\Omega_3^{t+1} \leq \frac{6\gamma C}{\rho} \Omega_1^t + \frac{6\eta^2}{\gamma \rho} \Omega_4^t + \left(1 - \frac{\gamma \rho}{2}\right) \Omega_3^t + \frac{6n^2}{\gamma \rho} \Omega_5^t + 0 \cdot \Omega_5^t,
\]

\[
\Omega_4^{t+1} \leq \frac{18\gamma^2 CL^2}{\rho} \Omega_1^t + \frac{6\gamma C}{\rho} \Omega_2^t + \frac{18\gamma^2 CL^2}{\rho} \Omega_3^t + \left(1 - \frac{\gamma \rho}{2} + \frac{18L^2 \eta^2}{\gamma \rho}\right) \Omega_4^t + \frac{18L^2 \eta^2 L^2}{\gamma \rho} \Omega_5^t + \frac{12n^2 \sigma^2}{b^2 \rho},
\]

where

\[
C = \left\| W - I \right\|^2 = \sigma_{\max}(W - I)^2
\]

is the square of the maximum singular value of the matrix \(W - I\).

Note that the eigenvalues of \(W\) and \(I\) all lies in \([-1, 1]\), and thus clearly \(C \leq 4\).

**Proof.** We will establish the inequalities in (14) one by one.
Bounding $\Omega_1^t$ in (14a) First from the update rule of BEER (cf. Line 3), we have
\[
\|H^{t+1} - X^{t+1}\|_F^2 = \|H^t + C(X^{t+1} - H^t) - X^{t+1}\|_F^2 \\
\leq (1 - \alpha) \|X^{t+1} - H^t\|_F^2 \\
\leq \left(1 - \frac{\alpha}{2}\right) \|X^t - H^t\|_F^2 + \frac{2}{\alpha} \|X^{t+1} - X^t\|_F^2,
\]
where the first inequality comes from the definition of compression operators (Definition 2.2) and the second inequality comes from Young’s inequality. It then boils down to bound $\|X^{t+1} - X^t\|_F^2$, for which we have
\[
\|X^{t+1} - X^t\|_F^2 = \|\gamma H^t(W - I) - \eta V^t\|_F^2 \\
= \|\gamma(H^t - X^t)(W - I) + \gamma(X^t - \bar{x}^t 1^\top)(W - I) - \eta V^t\|_F^2 \\
\leq 3\gamma^2 C \|X^t - H^t\|_F + 3\gamma^2 C \|X^t - \bar{x}^t 1^\top\|_F^2 + 3\eta^2 \|V^t\|_F^2 \\
= 3\gamma^2 C \|X^t - H^t\|_F + 3\gamma^2 C \|X^t - \bar{x}^t 1^\top\|_F^2 + 3\eta^2 \|V^t - \bar{v}^t 1^\top\|_F^2 + 3\eta^2 n \|\bar{v}^t\|^2,
\]
where in the first line we use the update rule of BEER (cf. Line 3), in the second line we use the property of the mixing matrix $1^\top W = 1^\top$, and in the third line, we apply Young’s inequality (cf. (9)). In the fourth line, we use $\|v\|_2^2 = \|v - \bar{v} 1\|_2^2 + n \bar{v}^2$ for any vector $v$ with an average $\bar{v}$. Plugging this back into (16), we get
\[
\|H^{t+1} - X^{t+1}\|_F^2 \leq \left(1 - \frac{\alpha}{2}\right) \|X^t - H^t\|_F^2 + \frac{6\gamma^2 C}{\alpha} \|X^t - \bar{x}^t 1^\top\|_F^2 \\
+ \frac{6\eta^2}{\alpha} \|V^t - \bar{v}^t 1^\top\|_F^2 + \frac{6n\eta^2}{\alpha} \|\bar{v}^t\|^2.
\]
Plugging in the definitions of $\Omega_1^t$, we obtain (14a).

Bounding $\Omega_2^t$ in (14b) Similar to the derivation of (14a), by applying the update rule of $G^t$ in BEER (Line 5), the definition of compression operators (Definition 2.2), and Young’s inequality, we have
\[
\|V^{t+1} - G^{t+1}\|_F^2 = \|G^t + C(V^{t+1} - G^t) - V^{t+1}\|_F^2 \\
\leq (1 - \alpha) \|G^t - V^{t+1}\|_F^2 \\
\leq \left(1 - \frac{\alpha}{2}\right) \|G^t - V^t\|_F^2 + \frac{2}{\alpha} \|V^{t+1} - V^t\|_F^2.
\]
It then boils down to bound $\|V^{t+1} - V^t\|_F^2$. By the update rule of BEER (cf. Line 6), we have
\[
\|V^{t+1} - V^t\|_F^2 = \|\gamma G^t(W - I) + (\bar{\nabla}_b F(X^{t+1}) - \bar{\nabla}_b F(X^t))\|_F^2 \\
= \|\gamma(G^t - V^t)(W - I) + (\bar{\nabla}_b F(X^{t+1}) - \bar{\nabla}_b F(X^t))\|_F^2 \\
\leq 3\gamma^2 C \|G^t - V^t\|_F^2 + 3\gamma^2 C \|V^t - \bar{v}^t 1^\top\|_F^2 + 3 \|\bar{\nabla}_b F(X^{t+1}) - \bar{\nabla}_b F(X^t)\|_F^2 \\
\leq 3\gamma^2 C \|G^t - V^t\|_F^2 + 3\gamma^2 C \|V^t - \bar{v}^t 1^\top\|_F^2 + 3 \|\nabla F(X^{t+1}) - \nabla F(X^t)\|_F^2 + \frac{6n\sigma^2}{b} \\
\leq 3\gamma^2 C \|G^t - V^t\|_F^2 + 3\gamma^2 C \|V^t - \bar{v}^t 1^\top\|_F^2 + 3 \|\nabla F(X^{t+1}) - \nabla F(X^t)\|_F^2 + \frac{6n\sigma^2}{b} \\
+ 9\gamma^2 C L^2 \|X^t - H^t\|_F^2 + 9\gamma^2 C L^2 \|X^t - \bar{x}^t 1^\top\|_F^2 + 9L^2 \eta^2 n \|\bar{v}^t\|^2 + \frac{6n\sigma^2}{b},
\]
where (i) comes from Young’s inequality (cf. (9)) and basic facts of matrix norm (cf. (15)), (ii) comes from the bounded variance assumption (Assumption 2.4), (iii) comes from the smoothness assumption (Assumption 2.3), and (iv) follows from (18). Combining the above inequality with (19), we have

$$
\| V^{t+1} - G^{t+1} \|_F^2 \leq \left( 1 - \frac{\alpha}{2} \right) \| G^t - V^t \|_F^2 + \frac{2}{\alpha} \| V^{t+1} - V^t \|_F^2
\leq \left( 1 - \frac{\alpha}{2} \right) \frac{6\gamma^2 C}{\alpha} \| G^t - V^t \|_F^2 + \frac{6\gamma^2 C + 18L\eta^2}{\alpha} \| V^t - \bar{v}^t 1^\top \|_F^2
+ \frac{18\gamma^2 CL^2}{\alpha} \| X^t - H^t \|_F^2 + \frac{18\gamma^2 CL^2}{\alpha} \| X^t - \bar{x}^t 1^\top \|_F^2 + \frac{18L^2 \eta^2 n}{\alpha} \| \bar{v}^t \|_F^2 + \frac{12na^2}{\beta \alpha}.
$$

Plugging in the definitions of $\Omega_t^p$, we obtain (14b).

**Bounding $\Omega_t^p$ in (14c)** To bound the consensus error $\| X^{t+1} - \bar{x}^{t+1} 1^\top \|_F^2$, by the update rule of BEER (cf. Line 3), we have

$$
\| X^{t+1} - \bar{x}^{t+1} 1^\top \|_F^2
= \| X^t + \gamma H^t (W - I) - \eta V^t - \bar{x}^t 1^\top - \bar{v}^t 1^\top \|_F^2
\leq (1 + \beta)(1 - \gamma \rho) \| X^t - \bar{x}^t 1^\top \|_F^2 + \left( 1 + \frac{1}{\beta} \right) \left( 2\gamma^2 \| (H^t - X^t)(W - I) \|_F^2 + 2\eta^2 \| V^t - \bar{v}^t 1^\top \|_F^2 \right)
\leq \left( 1 - \frac{\gamma \rho}{2} \right) \| X^t - \bar{x}^t 1^\top \|_F^2 + \left( 1 + \frac{2}{\gamma \rho} \right) \left( 2\gamma^2 \| (H^t - X^t)(W - I) \|_F^2 + 2\eta^2 \| V^t - \bar{v}^t 1^\top \|_F^2 \right)
\leq \left( 1 - \frac{\gamma \rho}{2} \right) \| X^t - \bar{x}^t 1^\top \|_F^2 + \left( 1 + \frac{2}{\gamma \rho} \right) \left( 2\gamma^2 \| H^t - X^t \|_F^2 + 2\eta^2 \| V^t - \bar{v}^t 1^\top \|_F^2 \right)
\leq \left( 1 - \frac{\gamma \rho}{2} \right) \| X^t - \bar{x}^t 1^\top \|_F^2 + \frac{6\gamma C}{\rho} \| H^t - X^t \|_F^2 + \frac{18\gamma^2 C}{\rho} \| V^t - \bar{v}^t 1^\top \|_F^2,
$$

where (i) follows from the definition $\bar{W} = I + \gamma (W - I)$, (ii) follows from applying Young’s inequality twice and Lemma B.2 i.e.

$$
\| X^t \bar{W} - \bar{x}^t 1^\top \|_F \leq (1 - \gamma \rho) \| X^t - \bar{x}^t 1^\top \|_F^2,
$$

(iii) follows by choosing $\beta = \gamma \rho/2$, and (iv) uses the definition of $C$ (cf. (15)). Plugging in the definitions of $\Omega_t^p$, we obtain (14c).

**Bounding $\Omega_t^p$ in (14d)** First, note that

$$
\| V^{t+1} - \bar{v}^{t+1} 1^\top \|_F^2 = \| V^{t+1} - \bar{v}^t 1^\top + \bar{v}^t 1^\top - \bar{v}^{t+1} 1^\top \|_F^2
= \| V^{t+1} - \bar{v}^t 1^\top \|_F^2 - n\| \bar{v}^{t+1} - \bar{v}^t \|_F^2
\leq \| V^{t+1} - \bar{v}^t 1^\top \|_F^2.
$$

Thus by the update rule of BEER (cf. Line 6), we have

$$
\| V^{t+1} - \bar{v}^{t+1} 1^\top \|_F^2
\leq \| V^{t+1} - \bar{v}^t 1^\top \|_F^2
= \| V^t + \gamma G^{t+1} (W - I) + \nabla_b F(X^{t+1}) - \nabla_b F(X^t) - \bar{v}^t 1^\top \|_F^2
= \| (V^t \bar{W} - \bar{v}^t 1^\top) + \gamma (G^{t+1} - V^t)(W - I) + (\nabla_b F(X^{t+1}) - \nabla_b F(X^t)) \|_F^2
$$
we can get (14d) by plugging in the definitions of stochastic samples, we get

\[
\begin{align*}
F(x^{t+1}) &\leq F(x^t) - \frac{nL^2}{2} \|\nabla F(x^t)\|^2 + \frac{\eta L^2}{2n} \|\nabla F(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \|\nabla F(x^t)\|^2 + \frac{\eta \sigma^2}{2n}
\end{align*}
\]

where (i) and (ii) are obtained similarly as the derivation of (14c), and the last line follows from (18). Thus, we can get (14d) by plugging in the definitions of \(\Omega^b_t\) and conclude the proof. \(\square\)

B.3 Proof of Theorem 4.1 and 4.2

Note that Theorem 4.2 is a strict generalization of Theorem 4.1, and thus we will directly prove Theorem 4.2. This proof makes use of Lemma B.3 and Lemma B.4 by constructing some proper Lyapunov function and demonstrate its descending property using a linear system argument, which is also used in, e.g., Li et al. (2021a); Liao et al. (2021).

Step 1: establishing a “descent” property of the function value First, we have the following inequality captures the “descent” of the function value.

\[
\begin{align*}
 f(x^{t+1}) &\leq f(x^t) - \eta \langle \nabla f(x^t), \nabla f(x^t) \rangle + \frac{\eta^2 L^2}{2} \|\nabla f(x^t)\|^2 \\
 &\leq f(x^t) - \frac{\eta}{2} \|\nabla f(x^t)\|^2 - \frac{\eta}{2} \|\nabla f(x^t)\|^2 + \frac{\eta}{2} \|\nabla f(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \|\nabla f(x^t)\|^2 \\
 &\leq f(x^t) - \frac{\eta}{2} \|\nabla f(x^t)\|^2 + \frac{\eta}{2} \|\nabla f(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \|\nabla f(x^t)\|^2 \\
 &\leq f(x^t) - \frac{\eta}{2} \|\nabla F(x^t)\|^2 + \frac{\eta}{2} \|\nabla F(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \|\nabla f(x^t)\|^2 \\
 &\leq f(x^t) - \frac{\eta}{2} \|\nabla F(x^t)\|^2 + \frac{\eta L^2}{2n} \|\nabla F(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \|\nabla f(x^t)\|^2 + \frac{\eta \sigma^2}{2n}
\end{align*}
\]

where (i) comes from the L-smooth assumption (Assumption 2.3), (ii) comes from Lemma B.3. Take expectation on both sides, and using the bounded variance assumption (Assumption 2.4) and independence of stochastic samples, we get

\[
\begin{align*}
 \mathbb{E} f(x^{t+1}) &\leq \mathbb{E} f(x^t) - \frac{\eta}{2} \mathbb{E} \|\nabla f(x^t)\|^2 + \frac{\eta}{2} \mathbb{E} \|\nabla F(x^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \mathbb{E} \|\nabla f(x^t)\|^2 + \frac{\eta \sigma^2}{2n}
\end{align*}
\]

where (i) comes from Young’s inequality, and (ii) comes from the L-smooth assumption (Assumption 2.3) again. Finally, by substituting definitions of \(\Omega^b_t\) and \(\Omega^b_t\), we reach

\[
\begin{align*}
 \mathbb{E} f(x^{t+1}) &\leq \mathbb{E} f(x^t) - \frac{\eta}{2} \mathbb{E} \|\nabla f(x^t)\|^2 + \frac{\eta L^2}{2n} \Omega^b_t - \left(\frac{\eta}{2} - \frac{\eta L^2}{2n}\right) \Omega^b_t + \frac{\eta \sigma^2}{2n}.
\end{align*}
\]
Step 2: constructing the Lyapunov function

By representing

\[ \Omega^t = \begin{bmatrix} \Omega_1^t & \Omega_2^t & \Omega_3^t & \Omega_4^t \end{bmatrix}^T, \]

Lemma B.4 can be written more compactly as

\[
\Omega^{t+1} \leq \begin{bmatrix}
1 - \frac{\eta^2 L^2}{2n} + \frac{6\eta^2 C}{\alpha} & 0 & \frac{6\eta^2 C}{\rho} & 0 \\
\frac{6\eta^2 C}{\alpha} & 1 - \frac{\eta^2 L^2}{2n} + \frac{6\eta^2 C + 18L^2 \rho^2}{\rho} & 0 & \frac{6\eta^2 C + 18L^2 \rho^2}{\rho} \\
\frac{6\eta^2 C}{\rho} & 0 & 1 - \frac{\eta^2 L^2}{2n} + \frac{6\eta^2 C + 18L^2 \rho^2}{\rho} & 0 \\
\frac{6\eta^2 C}{\rho} & 0 & 0 & 1 - \frac{\eta^2 L^2}{2n} + \frac{6\eta^2 C + 18L^2 \rho^2}{\rho}
\end{bmatrix} \Omega^t + \begin{bmatrix}
0 \\
\frac{6\eta^2 L^2}{\alpha} \\
\frac{6\eta^2 C}{\rho} \\
\frac{6\eta^2 C + 18L^2 \rho^2}{\rho}
\end{bmatrix} \sigma^2 + \begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix} \frac{\sigma^2}{b}.
\]

Define the Lyapunov function

\[
\Phi_t = \mathbb{E} f(\bar{x}^t) - f^* + \frac{c_1 L}{n} \cdot \Omega_1^t + \frac{c_2 \rho^2}{nL} \cdot \Omega_2^t + \frac{c_3 L}{n} \cdot \Omega_3^t + \frac{c_4 \rho^2}{nL} \Omega_4^t
\]

\[
= \mathbb{E} f(\bar{x}^t) - f^* + s^T \Omega^t,
\]

where

\[
s = \begin{bmatrix}
c_1 L \\
c_2 \rho^2 \\
c_3 L \\
c_4 \rho^2
\end{bmatrix}
\]

for some constants \(c_1, c_2, c_3, c_4\) that will be specified later.

By (22) from Lemma B.4 and the descent property (20), we have

\[
\Phi_{t+1} = \mathbb{E} f(\bar{x}^{t+1}) - f^* + s^T \Omega^{t+1}
\]

\[
\leq \mathbb{E} f(\bar{x}^t) - f^* - \frac{\eta}{2} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 - \frac{\eta^2 L^2}{2n} \Omega_3^t - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2}\right) \Omega_5^t + \frac{\eta \sigma^2}{2b} + s^T \left(A \Omega^t + \Omega_5^t b_1 + \frac{\sigma^2}{b} b_2\right)
\]

\[
\leq \Phi_t - \frac{\eta}{2} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2}\right) \Omega_5^t + \frac{\eta \sigma^2}{2b} + (s^T A - s^T + q^T) \Omega^t + s^T \left(\Omega_5^t b_1 + \frac{\sigma^2}{b} b_2\right)
\]

\[
= \Phi_t - \frac{\eta}{2} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + (s^T A - s^T + q^T) \Omega^t - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2} - s^T b_1\right) \Omega_5^t + \left(\frac{\eta}{2} + s^T b_2\right) \frac{\sigma^2}{b},
\]

where \(q = [0 \ 0 \ \frac{nL^2}{2n} \ 0]^T\). For a moment we assume that there exist some constants \(c_1, c_2, c_3, c_4 > 0\) such that

\[
s^T (A - I) + q^T \leq 0,
\]

\[
\frac{\eta}{2} - \frac{\eta^2 L}{2} - s^T b_1 \geq 0,
\]

leading to

\[
\Phi_{t+1} \leq \Phi_t - \frac{\eta}{2} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + \left(\frac{\eta}{2n} + s^T b_2\right) \frac{\sigma^2}{b} \leq \Phi_t - \frac{\eta}{2} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + \frac{36c_4 \sigma^2}{c_7 L b_0}.
\]

The proof is thus completed by recursing the above relation over \(t = 0, \ldots, T - 1\).

Step 3: verifying (25) It boils down to verify (25) is feasible, and it is equivalent to verify there exist parameters \(c_1, c_2, c_3, c_4, \gamma, \eta > 0\) satisfying the following matrix inequality:

\[
\begin{bmatrix}
I - A^T \\
-b_1
\end{bmatrix} \diag \begin{bmatrix}
L/n \rho^2 & L/n \rho^2 \\
L/n \rho^2 & L/n \rho^2
\end{bmatrix} \begin{bmatrix}
c_1 \\
c_2 \\
c_3 \\
c_4
\end{bmatrix} \geq \begin{bmatrix}
\gamma^2 L^2 - \frac{q}{2} \\
\frac{q}{2}
\end{bmatrix}.
\]
Note that by choosing \( \gamma = c_\eta \rho \alpha, \eta = c_\eta \gamma \rho^2 / L \), and setting \( c_\gamma \leq \frac{1}{6 \sqrt{C}} \) and \( c_\eta \leq \frac{1}{9} \), we get

\[
1 - \frac{\alpha}{2} + \frac{6 c_\gamma^2 C}{\alpha} \leq 1 - \frac{\alpha}{4}, \quad 1 - \frac{\gamma \rho}{2} + \frac{18 L^2 \eta^2}{\gamma \rho} \leq 1 - \frac{\gamma \rho}{4}.
\]

(26)

Now, it suffices to show that there exist \( c_1, c_2, c_3, c_4, c_\gamma, c_\eta > 0 \) such that the following inequalities are satisfied:

\[
\begin{bmatrix}
\frac{aL}{4n} & -\frac{2\sigma^2}{n} & 0 & -\frac{18 \sigma^2}{n} & -\frac{6 \sigma^2}{n} & -\frac{18 \sigma^2}{n} \\
0 & \frac{\sigma^2}{n} & \frac{\sigma^2}{n} & \frac{\sigma^2}{n} & \frac{\sigma^2}{n} & \frac{\sigma^2}{n} \\
-\frac{6 \sigma^2}{n} & 0 & 0 & 0 & 0 & 0 \\
-\frac{6 \sigma^2}{n} & 0 & 0 & 0 & 0 & 0 \\
-12 c_\eta c_\gamma \rho^2 \frac{2}{L} & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_2 \\
c_3 \\
c_4 \\
\end{bmatrix}
\geq
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
\end{bmatrix}.
\]

Given \( \alpha \leq 1, \rho \leq 1 \), this can be further reduced to show the existence of \( c_1, c_2, c_3, c_4, c_\gamma, c_\eta > 0 \) such that

\[
\begin{bmatrix}
1 & -72C c_\gamma^2 & -24C c_\gamma & -72C c_\gamma \\
0 & 1 & 0 & -24C c_\gamma \\
-12C c_\gamma & -35C c_\gamma & 1 & -36C \\
-24 c_\eta c_\gamma & -24 c_\gamma (1 + 3 c_\gamma^2) & -24 c_\eta^2 & 1 \\
-12 c_\eta c_\gamma & -36 c_\eta c_\gamma & 0 & -36 c_\eta
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_2 \\
c_3 \\
c_4 \\
\end{bmatrix}
\geq
\begin{bmatrix}
0 \\
0 \\
0 \\
-1 + c_\eta c_\gamma
\end{bmatrix}.
\]

This can be easily verified by noting that as long as \( c_\eta \) and \( c_\gamma \) are set sufficiently small, it is straightforward to find feasible \( c_1, c_2, c_3, c_4 \).

**B.4 Proof of Theorem 4.3 and 4.4**

Since Theorem 4.4 is a generalization of Theorem 4.3, it suffices to directly prove Theorem 4.4. The proof strategy of Theorem 4.4 is similar to that of Theorem 4.2. However, in order to achieve the advertised linear convergence rate under the PL condition (Assumption 2.5), we need to use a slightly different linear system. Denote \( \kappa := L / \mu \). Taking the same Lyapunov function \( \Phi_t \) in \ref{eq:Lyapunov}, by the same argument of Section B.3 up to \ref{eq:gamma convergence}, we have

\[
\Phi_{t+1} \leq \mathbb{E}_t f(x_t) - f^* - \eta \mathbb{E}_t \| \nabla f(x_t) \|^2 + \frac{\eta L^2}{2n} \Omega_3^t - \left( \frac{\eta}{2} - \frac{\eta^2 L}{2} \right) \Omega_3^t + \frac{\eta^2}{2n} + s^T \left( A \Omega_t^t + b_1 \right) + \frac{\sigma^2}{b} b_2^t
\]

\[
\leq (1 - \eta \mu) \mathbb{E}_t f(x_t) - f^* + \left( s^T A - \eta \mu s^T + q^T \right) \Omega_t^t - \left( \frac{\eta}{2} - \frac{\eta^2 L}{2} - s^T b_1 \right) \Omega_t^t + \left( \frac{\eta}{2} + s^T b_2 \right) \frac{\sigma^2}{b}
\]

\[
= (1 - \eta \mu) \Phi_t + \left( s^T A - (1 - \eta \mu) s^T + q^T \right) \Omega_t^t - \left( \frac{\eta}{2} - \frac{\eta^2 L}{2} - s^T b_1 \right) \Omega_t^t + \left( \frac{\eta}{2} + s^T b_2 \right) \frac{\sigma^2}{b},
\]

where \( q = [0 \ 0 \ \frac{n L^2}{2n} \ 0]^T \), and the second inequality follows from the PL condition (Assumption 2.5). If we can establish that there exist there exist some constants \( c_1, c_2, c_3, c_4 \) such that

\[
\begin{align}
\eta \frac{\eta^2 L}{2} - s^T b_1 & \geq 0, \quad (27a) \\
\eta \frac{\eta^2 L}{2} - s^T b_1 & \geq 0, \quad (27b)
\end{align}
\]

we arrive at

\[
\Phi_{t+1} \leq (1 - \eta \mu) \Phi_t + \left( \frac{\eta}{2} + s^T b_2 \right) \frac{\sigma^2}{b} \leq (1 - \eta \mu) \Phi_t + \frac{36 c_4 \sigma^2}{c_\eta L b \alpha}.
\]

Recurring the above relation then complete the proof.
It then boils down to establish (27). By similar arguments as Section B.3 in view of (26) and $\alpha \leq 1, \rho \leq 1$, it is sufficient to show there exist constants $c_1, c_2, c_3, c_4, c_\eta, c_\gamma > 0$ such that

$$
\begin{bmatrix}
1 - \frac{4c_\eta c_\gamma}{\kappa} & -72C c_\gamma^2 & -24Cc_\gamma & -72Cc_\gamma \\
0 & 1 - \frac{4c_\eta c_\gamma}{\kappa} & 0 & -24Cc_\gamma \\
-12Cc_\gamma & -35Cc_\gamma & 1 - \frac{2c_\eta}{\kappa} & -36C \\
-24c_\eta^2c_\gamma & -24c_\gamma(1 + 3c_\eta^2) & -24c_\eta^2 & 1 - \frac{4c_\eta}{\kappa} \\
-12c_\eta c_\gamma & -36c_\eta c_\gamma & 0 & -36c_\eta
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_2 \\
c_3 \\
c_4
\end{bmatrix}
\geq
\begin{bmatrix}
0 \\
0 \\
c_\eta \\
0
\end{bmatrix}
-1 + c_\eta c_\gamma
$$

This can be easily verified by noting that as long as $c_\eta$ and $c_\gamma$ are set sufficiently small, it is straightforward to find feasible $c_1, c_2, c_3, c_4$. 

25