Basis Matters: Better Communication-Efficient Second Order Methods for Federated Learning

Xun Qian KAUST[1] JD Explore Academy, Beijing
Rustem Islamov Institut Polytechnique de Paris KAUST
Mher Safaryan KAUST
Peter Richtárik KAUST

Abstract

Recent advances in distributed optimization have shown that Newton-type methods with proper communication compression mechanisms can guarantee fast local rates and low communication cost compared to first order methods. We discover that the communication cost of these methods can be further reduced, sometimes dramatically so, with a surprisingly simple trick: *Basis Learn (BL)*. The idea is to transform the usual representation of the local Hessians via a change of basis in the space of matrices and apply compression tools to the new representation. To demonstrate the potential of using custom bases, we design a new Newton-type method (BL1), which reduces communication cost via both *BL* technique and bidirectional compression mechanism. Furthermore, we present two alternative extensions (BL2 and BL3) to partial participation to accommodate federated learning applications. We prove local linear and superlinear rates independent of the condition number. Finally, we support our claims with numerical experiments by comparing several first and second order methods.

1 INTRODUCTION

We consider federated optimization problem of the form

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where each function $f_i : \mathbb{R}^d \to \mathbb{R}$ represents the local loss associated with the data owned by device or client $i \in [n] := \{1, 2, \ldots, n\}$ only. This formulation aims to train a single machine learning model $x \in \mathbb{R}^d$ composed of d parameters by minimizing empirical loss $f(x)$ using all n clients' data. We assume f is μ-strongly convex and problem (1) has the unique optimal solution x^* throughout the paper.

Due to the increasing size of the model and the amount of the training data, in practical deployments, methods of choice for solving the problem (1) have been *distributed first-order gradient methods* so far (Liu and Zhang, 2020; Xu et al., 2020). Among other things, two key considerations in the design of efficient distributed optimization method are *iteration complexity*, which measures how many iterations the method should take to achieve some prescribed accuracy, and *communication cost* per iteration, which measures the number of bits that clients communicate to each other or some parameter server (Bekkerman et al., 2011). Expectedly, these two quantities are in a trade-off: reducing the size of communicated messages per iteration, potentially increases the total number of iterations. This trade-off then interacts with the problem structure (training data and model) and network properties (bandwidth and latency) to find the best configuration for final deployment.

However, despite their wide applicability, all first-order methods inevitably suffer from a dependence of suitably defined *condition number*. To overcome the curse of the condition number, Newton-type or second-order optimization methods have been gaining considerable attention recently since (at least local) convergence properties of these algorithms are not affected by the condition number of the problem (Dennis and Moré, 1974; Dembo et al., 1982; Griewank, 1981; Nesterov and Polyak, 2006). On the other hand, the caveat of this approach is that, although it greatly improves iteration complexity, the cost of naively communicating second-order information, such as Hessian matrices, is extremely high and infeasible in practice (Bekkerman et al., 2011). In this work, we argue that with proper care of second-order information and for ill-conditioned

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

problems, distributed second-order algorithms can offer essentially better trade-offs than first-order algorithms.

1.1 Distributed Second Order Methods

The straightforward implementation of classical Newton's method in the distributed environment includes communication of local Hessian matrices $\nabla^2 f_i(x^k)$ with d^2 entries in all iterations $k \geq 0$. Consider this naive implementation of Newton's method as the baseline algorithm for distributed second-order optimization, just like the distributed gradient descent algorithm for first-order methods. Below we discuss the main algorithmic designs to reduce the quadratic dependence d^2 of the dimension in per-iteration communication cost and make second-order methods communication efficient for distributed optimization.

One stream of works avoids sending full Hessian matrices and uses second-order information locally to compute Hessian-vector products $\nabla^2 f_i(x^k) v_i^k \in \mathbb{R}^d$ for some (adaptively defined) vectors $v_i^k \in \mathbb{R}^d$. With this approach, second order information is imparted only through such products, which cost d floats of communication instead of d^2. The computation side of this approach is also efficient since the Hessian matrices are not computed directly but Hessian-vector products only, which are as cheap to compute as gradients $\nabla f_i(x^k)$ (Pearlmutter, 1993). Methods following this approach are DiSCO (Zhang and Xiao, 2015) (see also (Zhuang et al., 2015; Lin et al., 2014; Roosta et al., 2019)), GIANT (Wang et al., 2018) (see also (Shamir et al., 2014; Reddi et al., 2016)) and DINGO (Crane and Roosta, 2019) (see also (Ghosh et al., 2020)). Inspired by the local first order methods (Gorbunov et al., 2021; Stich, 2020; Khaled et al., 2020; Konečný et al., 2016), local variant of Newton's method was studied in (Gupta et al., 2021). Typically these methods either guarantee fast rates under stronger assumptions, such as quadratic problems or/and homogeneous data distribution or guarantee only linear rates attainable by first-order methods.

Alternatively, the high cost of Hessian communication can be reduced by compressing second-order information via lossy compression operators acting on matrices (such as low-rank approximations, randomly or greedily sparsifying entries). Again, this idea was originated from the first-order methods employing communication compression (Wangni et al., 2018; Alistarh et al., 2018, 2017; Wen et al., 2017; Chen et al., 2021). Recently developed second-order methods DAN-LA (Zhang et al., 2020), Quantized Newton (Alimisis et al., 2021), NewtonLearn (Islamov et al., 2021) and FedNL (Safaryan et al., 2021) are based on this idea of properly incorporating compression strategies for second-order information. In contrast to the previous approach, this strategy relies on the computation of full Hessian matrices, which might be computationally intensive for some applications. However, the optimization problem these methods address is quite generic (general finite sum structure (1) over arbitrarily heterogeneous data), and theoretical guarantees (global linear with local superlinear convergence rates) are far beyond the reach of all first-order methods.

Motivated by these recent developments on distributed second-order methods with communication compression, we extend the results of FedNL (Safaryan et al., 2021) allowing even more aggressive compression for some applications.

2 MOTIVATION AND CONTRIBUTIONS

To motivate our approach properly and illustrate the potential of our technique, we discuss three different implementations of *classical Newton's method* for solving the problem (1).

2.1 Naive Implementation

For general finite sums (1), Newton's method requires each device $i \in [n]$ to compute gradient vector $\nabla f_i(x)$ and Hessian matrix $\nabla^2 f_i(x)$ at the current point and send them to the parameter server to do the model update. While the convergence of Newton's method is locally quadratic, $\mathcal{O}(d^2)$ communication costs are high due to quadratic dependence from the dimension d. However, we can devise more efficient implementations, given some prior knowledge of the problem or/and data structure.

2.2 Utilizing the Problem Structure

Suppose the problem (1) models the training of *Generalized Linear Model (GLM)*, such as ridge regression or logistic regression. Then each local loss function has the form[2]

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x), \tag{2}$$

where $f_{ij}(x) := \varphi_{ij}(a_{ij}^\top x)$ is the loss associated with j^{th} data-vector $a_{ij} \in \mathbb{R}^d$ stored on i^{th} device and $\varphi_{ij} \colon \mathbb{R} \to \mathbb{R}$ is the corresponding loss function. A better implementation of Newton's method taking advantage of the Hessian structure

$$\nabla^2 f_i(x) \quad = \quad \frac{1}{m} \sum_{j=1}^{m} \varphi_{ij}''(a_{ij}^\top x) a_{ij} a_{ij}^\top \tag{3}$$

[2]For simplicity we assume that each device has the same number of local data points denoted by m. Generally, it could be different for different clients.

is described in Section 2.2 of (Islamov et al., 2021). In this implementation, the server is assumed to have *access to all training dataset* $\{a_{ij}\}_{ij}$. Then to communicate Hessian matrix of the form (3), it is enough to send m coefficients $\{\varphi''_{ij}(a_{ij}^\top x)\colon j \in [m]\}$ instead of d^2 entries. Hence, in cases when $m \ll d^2$, we can run Newton's method much efficiently with $\mathcal{O}(m+d)$ communication cost.

However, there are two limitations here. First, this approach fails to benefit when local datasets are too big (i.e., $m > d^2$), which is often the case in practice. Second, all devices reveal their local training data to the server, making the approach inapplicable to federated learning applications, where data privacy is crucial.

2.3 Utilizing the Data Structure

We now describe a strategy that additionally takes advantage of the data structure and dramatically reduces communication costs regardless of the data size and without revealing any local data.

The imposed structural assumption on the data is that local data points $\{a_{ij}\colon j \in [m]\}$ of i^{th} client belong to some linear subspace $G_i \subseteq \mathbb{R}^d$ of dimension $r \in [d]$.[3] Note that this condition is trivially satisfied for $r = d$ for any data. However, in practice, training data points have much smaller intrinsic dimensionality $r \ll d$. Notice that once we fix some basis $\{v_{it}\}_{t=1}^r$ of G_i, we can represent data points a_{ij} as linear combinations

$$a_{ij} = \sum_{t=1}^{r} \alpha_{ijt} v_{it}, \quad j \in [m]. \tag{4}$$

Instead of directly revealing actual data points a_{ij}, each device sends the basis $\{v_{it}\}_{t=1}^r$ to the server *only once* (before the training) with the cost of sending rd floats. Based on the representations (3) from the problem structure and (4) from the data structure, the Hessian of $f_i(x)$ can be transformed into

$$\begin{aligned}\nabla^2 f_i(x) &\overset{(3),(4)}{=} \frac{1}{m}\sum_{j=1}^{m} \varphi''_{ij}(a_{ij}^\top x) \sum_{t,l=1}^{r} \alpha_{ijt}\alpha_{ijl} v_{it} v_{il}^\top \qquad (5)\\ &= \sum_{t,l=1}^{r} \underbrace{\left[\frac{1}{m}\sum_{j=1}^{m} \varphi''_{ij}(a_{ij}^\top x)\alpha_{ijt}\alpha_{ijl}\right]}_{\gamma_{itl}(x)} \underbrace{v_{it}v_{il}^\top}_{\mathbf{V}_{itl}},\end{aligned}$$

where outer products $\mathbf{V}_{itl} := v_{it}v_{il}^\top$ are linearly independent matrices (see the Appendix) and $\gamma_{itl}(x)$ are coefficients in the brackets. The key observation from (5) is that to communicate $\nabla^2 f_i(x)$ we need to send only r^2 coefficients $\{\gamma_{itl}(x)\colon t,l \in [r]\}$ instead of d^2 entries as the server already knows matrices $\mathbf{V}_{itl}$ through the basis $\{v_{it}\}_{t=1}^r$. The takeaway from this observations is that the standard basis of $\mathbb{R}^{d\times d}$ is not always optimal. Indeed, in this case, any basis of $\mathbb{R}^{d\times d}$ containing r^2 matrices $\{\mathbf{V}_{itl}\}_{t,l=1}^r$ is better choice for encoding Hessians $\nabla^2 f_i(x)$ without any loss in precision. Thus, $\mathcal{O}(d^2)$ communication cost is reduced to $\mathcal{O}(r^2+d)$. In case of $r = \mathcal{O}(\sqrt{d})$, we get Newton's method with $\mathcal{O}(d)$ communication cost and local quadratic convergence.

Analogous to (5), similar representation can be derived for gradients too, namely $\nabla f_i(x) \in G_i$. Hence, we can send $\nabla f_i(x)$ by its r basis coefficients instead of d coordinates. This way we further reduce communication cost up to $\mathcal{O}(r^2)$. In the extreme case of $r = \mathcal{O}(1)$, we run Newton's method with $\mathcal{O}(1)$ cost per iteration!

Note that (5) is a special case of more general Hessian representation $\nabla^2 f_i(x) = \mathbf{Q}_i \Lambda_i(x) \mathbf{Q}_i^\top$, where $\mathbf{Q}_i$ is a fixed invertible matrix (known to the server) and $\Lambda_i(x)$ is a sparse matrix with much less than d^2 (e.g., r^2 for (5)) non-zero entries. Changing the standard basis of $\mathbb{R}^{d\times d}$ via the transition matrix $\mathbf{Q}_i$, we transform potentially dense Hessian $\nabla^2 f_i(x)$ (in the standard basis) into sparse $\Lambda_i(x)$ in the new basis. *Thus, we save in communication for free just by changing the basis in the beginning of the training.* Motivated by this idea, we propose a new approach: *Basis Learn*.

2.4 Contributions

Our goal is to further investigate the benefits and possible pitfalls of using custom bases in second-order optimization for general finite sums (1) with arbitrarily heterogeneous data. As, by choosing a suitable basis, we can transform the Hessian into a sparser matrix in a lossless way, we propose and design three new methods, which apply lossy compression strategies afterwards to get even better performance in terms of communication complexity.

2.4.1 Basis Learn with Bidirectional Compression.

Our first contribution is the new method BL1, which successfully integrates bidirectional compression with any predefined basis for Hessians. In BL1, both client-to-server and server-to-client communications are compressed via careful application of compression operators. We allow both unbiased compressors, such as random sparsification (Rand-K) or random dithering, and contractive compressors, such as greedy sparsification (Top-K) or low-rank approximations (Rank-R). In the special case of choosing the standard basis, our method recovers FedNL (Safaryan et al., 2021). Thus,

[3] To make notations simpler, r is the same for all clients.

basis learn can be viewed as a generalization of FedNL.

2.4.2 Extensions to Partial Device Participation.

For massively distributed trainings, such as in federated learning, with too many clients, we propose two extensions, BL2 and BL3, to accommodate partial participation of devices. Thus, we unify bidirectional compression and partial participation under basis learn. Furthermore, within these two extensions we propose two options to guarantee the positive definiteness of accumulated Hessian estimator at the server avoiding matrix projection steps of BL1: first option (implemented in BL2) is based on compression error trick of (Safaryan et al., 2021), while the other option (realized in BL3) is to choose bases with positive semidefinite matrices in the symmetric matrix space.

2.4.3 Fast Local Rates.

For all our methods we prove local linear and superlinear rates independent of the condition number and the size of local dataset.

2.4.4 Experiments.

By composing low-rank approximation and unbiased compression operators, we propose more efficient compressors for matrices leading to better performance in the experiments.

3 MATRIX COMPRESSION

Here we adopt two classes of vector compressions to matrices. A (possibly) randomized map $\mathcal{C}:\mathbb{R}^{d\times d}\to\mathbb{R}^{d\times d}$ is called a *contraction compressor* if there exists a constant $0<\delta\leq 1$ such that

$$\mathbb{E}\left[\|\mathbf{A}-\mathcal{C}(\mathbf{A})\|_{\mathrm{F}}^2\right]\leq(1-\delta)\|\mathbf{A}\|_{\mathrm{F}}^2,\quad\forall\mathbf{A}\in\mathbb{R}^{d\times d}.\quad(6)$$

Further, we say that a randomized map $\mathcal{C}:\mathbb{R}^{d\times d}\to\mathbb{R}^{d\times d}$ is an *unbiased compressor* if there exists a constant $\omega\geq 0$ such that for any $\mathbf{A}\in\mathbb{R}^{d\times d}$

$$\mathbb{E}\left[\mathcal{C}(\mathbf{A})\right]=\mathbf{A}\text{ and }\mathbb{E}\left[\|\mathcal{C}(\mathbf{A})\|_{\mathrm{F}}^2\right]\leq(\omega+1)\|\mathbf{A}\|_{\mathrm{F}}^2.\quad(7)$$

The contraction compressor and unbiased compressor on $\mathbb{R}^d$ can be defined in the same way where the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ is replaced by the Euclidean norm $\|\cdot\|$. For more examples of contraction and unbiased compressors, we refer the reader to (Safaryan et al., 2021; Beznosikov et al., 2020). On the other hand, the compressor on $\mathbb{R}^{d\times d}$ can be regarded as a compressor on $\mathbb{R}^{d^2}$. Hence, compressors on the vector space $\mathbb{R}^{d^2}$ can be applied to the matrix in $\mathbb{R}^{d\times d}$. One can combine two compressors from different classes to get new ones (Qian et al., 2021). In particular, we consider composition of Rank-R (Safaryan et al., 2021) and unbiased compressors below.

Suppose $\mathcal{Q}_1^i$ and $\mathcal{Q}_2^i$, $i\in[d]$ are unbiased compressors on $\mathbb{R}^d$ with parameter ω_1 and ω_2 respectively. For any $\mathbf{A}\in\mathbb{R}^{d\times d}$, let $\mathbf{A}=\sum_{i=1}^d\sigma_i u_i v_i^\top$ be the singular value decomposition of $\mathbf{A}$ with singular values $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_d\geq 0$. For $R\leq d$, define

$$\mathcal{C}_1(\mathbf{A}):=\textstyle\sum_{i=1}^R\frac{\sigma_i\mathcal{Q}_1^i(a_iu_i)\mathcal{Q}_2^i(b_iv_i)^\top}{a_ib_i(\omega_1+1)(\omega_2+1)},$$

where $a_i,b_i>0$ are independent of $\mathcal{Q}_1^i$ and $\mathcal{Q}_2^i$ for $1\leq i\leq R$. For example, we can set $a_i\equiv b_i\equiv 1$, or $a_i=b_i=\sqrt{\sigma_i}$. Notice that if $\mathbf{A}$ is symmetric, $\mathcal{C}_1(\mathbf{A})$ is not necessarily symmetric. However, we can symmetrize the output matrix by defining

$$\mathcal{C}_2(\mathbf{A}):=\begin{cases}\mathcal{C}_1(\mathbf{A}) & \text{if }\mathbf{A}\text{ is not symmetric}\\ \frac{\mathcal{C}_1(\mathbf{A})+\mathcal{C}_1(\mathbf{A})^\top}{2} & \text{if }\mathbf{A}\text{ is symmetric}\end{cases}$$

Lemma 3.1. *(i) For any $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$, if $\mathbf{A}$ is symmetric, then we have $\left\|{}^{(\mathbf{B}+\mathbf{B}^\top)}/_2-\mathbf{A}\right\|_{\mathrm{F}}\leq\|\mathbf{B}-\mathbf{A}\|_{\mathrm{F}}$. (ii) For any contraction compressor $\mathcal{C}$ in $\mathbb{R}^{d\times d}$ with parameter δ,*

$$\tilde{\mathcal{C}}:=\begin{cases}\mathcal{C}(\mathbf{A}) & \text{if }\mathbf{A}\text{ is not symmetric}\\ \frac{\mathcal{C}(\mathbf{A})+\mathcal{C}(\mathbf{A})^\top}{2} & \text{if }\mathbf{A}\text{ is symmetric}\end{cases}$$

is also a contraction compressor with parameter δ.

Proposition 3.2. *$\mathcal{C}_1$ and $\mathcal{C}_2$ are both contraction compressors with parameter $\frac{R}{d(\omega_1+1)(\omega_2+1)}$.*

4 BASIS LEARN

4.1 Basis Learn in $\mathbb{R}^{d\times d}$

Let $\{\mathbf{B}_i^{jl}\mid j,l\in[d]\}$ be a basis in the space of matrices $\mathbb{R}^{d\times d}$ and $N:=d^2$ be the number of matrices in the basis for any $i\in[n]$. Then any matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$ has a unique basis representation $\mathbf{A}=\sum_{j,l}h_{jl}^i(\mathbf{A})\mathbf{B}_i^{jl}$, where $h_{jl}^i(\mathbf{A})\in\mathbb{R}$ is the coefficient corresponding to $\mathbf{B}_i^{jl}$.

Define $h^i(\mathbf{A})\in\mathbb{R}^{d\times d}$ to be the matrix of basis coefficients such that $h^i(\mathbf{A})_{jl}:=h_{jl}^i(\mathbf{A})$ for all $j,l\in[d]$. Let $vec:\mathbb{R}^{d\times d}\to\mathbb{R}^{d^2}$ be the map that vectorizes a given matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$ into the vector $vec(\mathbf{A})=(\mathbf{A}_{11},...,\mathbf{A}_{d1},...,\mathbf{A}_{1i},...,\mathbf{A}_{di},...,\mathbf{A}_{1d},...,\mathbf{A}_{dd})^\top\in\mathbb{R}^N$ by stacking all entries together. Besides, define $\mathcal{M}_i:=(\mathbf{B}_i^{11},...,\mathbf{B}_i^{d1},...,\mathbf{B}_i^{1j},...,\mathbf{B}_i^{dj},...,\mathbf{B}_i^{1d},...,\mathbf{B}_i^{dd})$ by stacking all basis matrices $\mathbf{B}_i^{jl}$, $j,l\in[d]$. Then we have $\mathbf{A}=\mathcal{M}_i vec(h^i(\mathbf{A}))$, which is equivalent to

$$vec(\mathbf{A})=\mathcal{B}_i\cdot vec(h^i(\mathbf{A})),\quad(8)$$

for any matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$, where

$$\mathcal{B}_i:=(vec(\mathbf{B}_i^{11}),...,vec(\mathbf{B}_i^{d1}),...,vec(\mathbf{B}_i^{1j}),...,vec(\mathbf{B}_i^{dj}),...,vec(\mathbf{B}_i^{1d}),...,vec(\mathbf{B}_i^{dd})).$$

As $vec(\mathbf{B}_i^{jl}) \in \mathbb{R}^N$ for all $j,l \in [d]$ and the number of matrices in the basis $\{\mathbf{B}_i^{jl} \mid j,l \in [d]\}$ is also N, we conclude $\mathcal{B}_i \in \mathbb{R}^{N\times N}$. Since the representation (8) is unique, we know $\mathcal{B}_i$ is invertible, and thus

$$vec(h^i(\mathbf{A})) = \mathcal{B}_i^{-1} vec(\mathbf{A}). \tag{9}$$

Next we provide some examples of the basis in $\mathbb{R}^{d\times d}$.

Example 4.1. *The $(j,l)^{th}$ entry of $\mathbf{B}_i^{jl}$ is 1 and the others are 0 for $j,l \in [d]$. Then $\mathbf{A} = h^i(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{R}^{d\times d}$.*

Example 4.2. *The $(j,l)^{th}$ and $(l,j)^{th}$ entries of $\mathbf{B}_i^{jl}$ are 1 and the others are 0 for $d \geq j \geq l \geq 1$. The $(j,l)^{th}$ entry of $\mathbf{B}_i^{jl}$ is 1, the $(l,j)^{th}$ entry of $\mathbf{B}_i^{jl}$ is -1, and the others are 0 for $1 \leq j < l \leq d$. Then for any symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $h^i(\mathbf{A})$ is the lower triangular part of $\mathbf{A}$.*

4.1.1 Basis Learn with Bidirectional Compression

We have Algorithm 1 (BL1) as an extension of FedNL-BC in (Safaryan et al., 2021). BL1 mainly has two differences from FedNL-BC: (i) We use $\mathbf{L}_i^k$ to learn the coefficient matrix $h^i(\nabla^2 f_i(z^k))$ rather than the Hessian; (ii) When $\xi^k = 0$, we use $\left[\mathbf{H}^k\right]_\mu$ rather than $\mathbf{H}^k$ to construct the gradient estimator g^k, where $[\cdot]_\mu$ represents the projection on the set $\{\mathbf{A} \in \mathbb{R}^{d\times d} \mid \mathbf{A} = \mathbf{A}^\top, \mathbf{A} \succeq \mu\mathbf{I}\}$. The notation $\xi^k \sim$ Bernoulli(p) in BL1 means that $\xi^k = 1$ with probability p and $\xi^k = 0$ otherwise. The main update in line 16 of Algorithm 1 is based on the Newton method, where $\left[\mathbf{H}^k\right]_\mu$ is an estimator of the Hessian matrix, and g^k is an estimator of $\nabla f(z^k)$. The rest steps are the same as FedNL-BC, hence we omit the description.

For the theory, we utilize the following assumptions commonly posed on the compression operators.

Assumption 4.3. *(i)$\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is an unbiased compressor with parameter ω_{M} and $0 < \eta \leq 1/(\omega_{\mathrm{M}}+1)$. (ii) For all $j \in [d]$, $(z^k)_j$ in Algorithm 1 ($(z_i^k)_j$ in Algorithm 2 or Algorithm 3) is a convex combination of $\{(x^t)_j\}_{t=0}^k$ for $k \geq 0$.*

Assumption 4.4. *(i) $\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is a contraction compressor with parameter δ_{M} and $\eta = 1$. (ii) $\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is deterministic, i.e., $\mathbb{E}[\mathcal{Q}^k(x)] = \mathcal{Q}^k(x)$ for any $x \in \mathbb{R}^d$.*

Assumption 4.5. *(i) $\mathcal{C}_i^k$ is an unbiased compressor with parameter ω and $0 < \alpha \leq 1/(\omega+1)$.*
(ii) For all $i \in [n]$ and $j,l \in [d]$, $(\mathbf{L}_i^k)_{jl}$ is a convex combination of $\{h^i(\nabla^2 f_i(z^t))_{jl}\}_{t=0}^k$ in Algorithm 1 ($\{h^i(\nabla^2 f_i(z_i^t))_{jl}\}_{t=0}^k$ in Algorithm 2 or $\{\tilde{h}^i(\nabla^2 f_i(z_i^t))_{jl}\}_{t=0}^k$ in Algorithm 3) for $k \geq 0$.

Assumption 4.6. *(i) $\mathcal{C}_i^k$ is a contraction compressor with parameter δ and $\alpha = 1$. (ii) $\mathcal{C}_i^k$ is deterministic, i.e., $\mathbb{E}[\mathcal{C}_i^k(\mathbf{A})] = \mathcal{C}_i^k(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{R}^{d\times d}$.*

Algorithm 1 BL1 (Basis Learn with Bidirectional Compression)

1: **Parameters:** Hessian learning rate $\alpha \geq 0$; model learning rate $\eta \geq 0$; gradient compression probability $p \in (0,1]$; compression operators $\{\mathcal{C}_1^k, \ldots, \mathcal{C}_n^k\}$ and $\mathcal{Q}^k$
2: **Initialization:** $x^0 = w^0 = z^0 \in \mathbb{R}^d$; $\mathbf{L}_i^0 \in \mathbb{R}^{d\times d}$, $\mathbf{H}_i^0 = \sum_{jl}(\mathbf{L}_i^0)_{jl}\mathbf{B}_i^{jl}$, and $\mathbf{H}^0 := \frac{1}{n}\sum_{i=1}^n \mathbf{H}_i^0$; $\xi^0 = 1$

3: **for** each device $i = 1, \ldots, n$ in parallel **do**
4: **if** $\xi^k = 1$
5: $w^{k+1} = z^k$, compute local gradient $\nabla f_i(z^k)$ and send to the server
6: **if** $\xi^k = 0$
7: $w^{k+1} = w^k$
8: Compute local Hessian $\nabla^2 f_i(z^k)$ and send $\mathbf{S}_i^k := \mathcal{C}_i^k(h^i(\nabla^2 f_i(z^k)) - \mathbf{L}_i^k)$ to the server
9: Update local Hessian shifts $\mathbf{L}_i^{k+1} = \mathbf{L}_i^k + \alpha\mathbf{S}_i^k$, $\mathbf{H}_i^{k+1} = \mathbf{H}_i^k + \alpha\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$
10: **end for**
11: **on** server
12: **if** $\xi^k = 1$
13: $w^{k+1} = z^k$, $g^k = \nabla f(z^k)$
14: **if** $\xi^k = 0$
15: $w^{k+1} = w^k$, $g^k = \left[\mathbf{H}^k\right]_\mu (z^k - w^k) + \nabla f(w^k)$
16: $x^{k+1} = z^k - \left[\mathbf{H}^k\right]_\mu^{-1} g^k$
17: $\mathbf{H}^{k+1} = \mathbf{H}^k + \frac{\alpha}{n}\sum_{i=1}^n \sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$
18: Send $v^k := \mathcal{Q}^k(x^{k+1} - z^k)$ to all devices $i \in [n]$
19: Update the model $z^{k+1} = z^k + \eta v^k$
20: Send $\xi^{k+1} \sim$ Bernoulli(p) to all devices $i \in [n]$
21: **for** each device $i = 1, \ldots, n$ in parallel **do**
22: Update the model $z^{k+1} = z^k + \eta v^k$
23: **end for**

One can easily check that the condition of convex combination in Assumptions 4.3 and 4.5 is satisfied when random sparsification is used. Next assumption is mainly related to the smoothness of the Hessian matrix and the property of the basis.

Assumption 4.7. *We have $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq H\|x - y\|$, $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x - y\|$, $\|h^i(\nabla^2 f_i(x)) - h^i(\nabla^2 f_i(y))\|_{\mathrm{F}} \leq M_1\|x - y\|$, $\max_{jl}\{|h^i(\nabla^2 f_i(x))_{jl} - h^i(\nabla^2 f_i(y))_{jl}|\} \leq M_2\|x - y\|$, $\max_{jl}\{\|\mathbf{B}_i^{jl}\|_{\mathrm{F}}\} \leq R$ for any $x, y \in \mathbb{R}^d$ and $i \in [n]$. For Algorithm 2, we assume each f_i is μ-strongly convex.*

We estimate the parameters M_1 and M_2 of Assumption 4.7 in the following lemma.

Lemma 4.8. *Assume $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x - y\|$ and $\max_{jl}\{|(\nabla^2 f_i(x))_{jl} - (\nabla^2 f_i(y))_{jl}|\} \leq \nu\|x - y\|$ for any $x, y \in \mathbb{R}^d$, and $i \in [n]$. Then we have $M_1 \leq \max_i\{\|\mathcal{B}_i^{-1}\|\}H_1$ and $M_2 \leq \nu\max_i\{\|\mathcal{B}_i^{-1}\|_\infty\}$.*

To present our theory in a unified manner, we define some constants depending on the structure of bases, the properties of compressors, and the choice of stepsize.

$$N_{\mathrm{B}} := \begin{cases} 1 & \text{if the bases } \{\mathbf{B}_i^{jl}\}_{j,l\in[d]} \text{ are all orthogonal} \\ d^2 & \text{otherwise} \end{cases} \tag{10}$$

$$(A_{\mathrm{M}}, B_{\mathrm{M}}) := \begin{cases} (\eta,\eta) & \text{if Asm. 4.3(i) holds} \\ \left(\frac{\delta_{\mathrm{M}}}{4}, \frac{6}{\delta_{\mathrm{M}}} - \frac{7}{2}\right) & \text{if Asm. 4.4(i) holds} \end{cases} \tag{11}$$

$$(A, B) := \begin{cases} (\alpha,\alpha) & \text{if Asm. 4.5(i) holds} \\ \left(\frac{\delta}{4}, \frac{6}{\delta} - \frac{7}{2}\right) & \text{if Asm. 4.6(i) holds} \end{cases} \tag{12}$$

For any $k \geq 0$, denote $\mathcal{H}^k := \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2$, $\Phi_1^k := \|z^k - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{2p}\|w^k - x^*\|^2$, where $\mathbf{L}_i^* := h^i(\nabla^2 f_i(x^*))$.

Theorem 4.9 (Linear convergence of BL1)**.** *Let Assumption 4.7 hold. Let Assumption 4.3 (i) or Assumption 4.4 (i) hold. Assume $\|z^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4H^2 B_{\mathrm{M}}}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16 N_{\mathrm{B}} R^2 B_{\mathrm{M}}}$ for $k \geq 0$. Then we have*

$$\mathbb{E}[\Phi_1^k] \leq \left(1 - \frac{\min\{A_{\mathrm{M}}, p\}}{2}\right)^k \Phi_1^0,$$

for $k \geq 0$.

Theorem 4.9 shows that the linear convergence is obtainable when $\|z^k - x^*\|$ and $\mathcal{H}^k$ are small enough, and the linear rate depends on A_{M} and p only, which indicates that we should choose A_{M} and p in the same order in BL1.

Define $\Phi_2^k := \mathcal{H}^k + \frac{4BM_1^2}{A_{\mathrm{M}}}\|x^k - x^*\|^2$ for $k \geq 0$. We prove a local superlinear convergence in the following theorem if there is no compression applied to the model and full gradients are calculated in each iteration. The convergence rate depends on parameter θ_1 which is also independent of the condition number of the problem.

Theorem 4.10 (Superlinear convergence of BL1)**.** *Let $\eta = 1$, $\xi^k \equiv 1$ and $\mathcal{Q}^k(x) \equiv x$ for any $x \in \mathbb{R}^d$ and $k \geq 0$. Let Assumption 4.7 hold. Let Assumption 4.5 (i) or Assumption 4.6 (i) hold. Assume $\|z^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4H^2 B_{\mathrm{M}}}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16 N_{\mathrm{B}} R^2 B_{\mathrm{M}}}$ for $k \geq 0$. Then we have*

$$\mathbb{E}[\Phi_2^k] \leq \theta_1^k \Phi_2^0,$$

$$\mathbb{E}\left[\frac{\|x^{k+1} - x^*\|^2}{\|x^k - x^*\|^2}\right] \leq \theta_1^k \left(\frac{A_{\mathrm{M}} H^2}{8BM_1^2\mu^2} + \frac{2N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_2^0,$$

for $k \geq 0$, where $\theta_1 := \left(1 - \frac{\min\{4A, A_{\mathrm{M}}\}}{4}\right)$.

Theorems 4.9 and 4.10 also show the trade-off between the communication cost at each iteration and the convergence rate, i.e., the less we communicate at each iteration (corresponding to higher level of compression and smaller value of p), the slower the convergence will be. Next, we explore under what conditions we can guarantee the boundedness of $\|z^k - x^*\|$ and $\mathcal{H}^k$.

Theorem 4.11. *Let Assumption 4.7 hold. Then we have the following results.*
(i) Let Assumption 4.3 and Assumption 4.5 (ii) hold. If $\|x^0 - x^*\|^2 \leq \tilde{c}_1 := \min\left\{\frac{\mu^2}{4d^2H^2}, \frac{\mu^2}{16d^4 N_{\mathrm{B}} R^2 M_2^2}\right\}$, then $\|z^k - x^*\|^2 \leq d\tilde{c}_1$ and $\mathcal{H}^k \leq \frac{\mu^2}{16dN_{\mathrm{B}}R^2}$ for $k \geq 0$.
(ii) Let Assumption 4.4 and Assumption 4.6 hold. If $\|z^0 - x^*\|^2 \leq \tilde{c}_2 := \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2 B_{\mathrm{M}}}, \frac{AA_{\mathrm{M}}\mu^2}{16 N_{\mathrm{B}} R^2 B_{\mathrm{M}} B M_1^2}\right\}$ and $\mathcal{H}^0 \leq \frac{A_{\mathrm{M}}\mu^2}{16 N_{\mathrm{B}} R^2 B_{\mathrm{M}}}$, then $\|z^k - x^*\|^2 \leq \tilde{c}_2$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16 N_{\mathrm{B}} R^2 B_{\mathrm{M}}}$ for $k \geq 0$.

4.1.2 Basis Learn with Bidirectional Compression and Partial Participation

In this section, we present the extension of BL1 where only a few selected clients are active at each iteration. In other words, we unify the bidirectional compression and partial participation in our BL2 (Algorithm 2).

As we deal with symmetric matrices such as Hessians, we introduce the operator $[\cdot]_s$ on the space of matrices $\mathbb{R}^{d\times d}$, which symmetrizes its input $\mathbf{A} \in \mathbb{R}^{d\times d}$ as $[\mathbf{A}]_s = (\mathbf{A}+\mathbf{A}^\top)/2$. The main update of the global model x^k is based on the Stochastic Newton method (Kovalev et al., 2019), where the update has the form of

$$x^{k+1} = \left[\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(w_i^k)\right]^{-1} \cdot \left[\frac{1}{n}\sum_{i=1}^n \left(\nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right)\right].$$

We use $[\mathbf{H}_i^k]_s + l_i^k \mathbf{I}$ to estimate $\nabla^2 f_i(w_i^k)$, and g_i^k to estimate $\nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)$, where $l_i^k = \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(z_i^k)\|_{\mathrm{F}}$ is adopted to guarantee the positive definiteness of $[\mathbf{H}_i^k]_s + l_i^k \mathbf{I}$. Thus, like in FedNL-PP (Safaryan et al., 2021), the key relation

$$g_i^k = ([\mathbf{H}_i^k]_s + l_i^k \mathbf{I}) w_i^k - \nabla f_i(w_i^k) \tag{13}$$

need to be maintained in the design of BL2. Since each node has a local model w_i^k, we introduce z_i^k to apply the bidirectional compression, and $\mathbf{L}_i^k$ is expected to learn $h^i(\nabla^2 f_i(z_i^k))$ iteratively. For the update of g_i^k on the server when $\xi_i^k = 0$, from (13), it is natural to make $g_i^{k+1} - g_i^k = ([\mathbf{H}_i^{k+1}]_s - [\mathbf{H}_i^k]_s + l_i^{k+1}\mathbf{I} - l_i^k \mathbf{I}) w_i^{k+1}$ since $w_i^{k+1} = w_i^k$.

We give the convergence results of BL2 in the following two theorems. Let $\Phi_3^k := \mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k$, where $\mathcal{Z}^k := \frac{1}{n}\sum_{i=1}^n \|z_i^k - x^*\|^2$, for $k \geq 0$.

Algorithm 2 BL2 (Basis Learn with Bidirectional Compression and Partial Participation)

Parameters: $\alpha > 0$; $\eta > 0$; matrix compression operators $\{\mathcal{C}_1^k, \ldots, \mathcal{C}_n^k\}$; $p \in (0,1]$; $0 < \tau \leq n$
Initialization: $w_i^0 = z_i^0 = x^0 \in \mathbb{R}^d$; $\mathbf{L}_i^0 \in \mathbb{R}^{d\times d}$; $\mathbf{H}_i^0 = \sum_{jl}(\mathbf{L}_i^0)_{jl}\mathbf{B}_i^{jl}$; $l_i^0 = \|[\mathbf{H}_i^0]_s - \nabla^2 f_i(w_i^0)\|_{\mathrm{F}}$; $g_i^0 = ([\mathbf{H}_i^0]_s + l_i^0\mathbf{I})w_i^0 - \nabla f_i(w_i^0)$; Moreover: $\mathbf{H}^0 = \frac{1}{n}\sum_{i=1}^n \mathbf{H}_i^0$; $l^0 = \frac{1}{n}\sum_{i=1}^n l_i^0$; $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$
on server
  $x^{k+1} = \left([\mathbf{H}^k]_s + l^k\mathbf{I}\right)^{-1} g^k$, choose a subset $S^k \subseteq [n]$ such that $\mathbb{P}[i \in S^k] = \tau/n$ for all $i \in [n]$
  $v_i^k = \mathcal{Q}_i^k(x^{k+1} - z_i^k)$, $z_i^{k+1} = z_i^k + \eta v_i^k$ for $i \in S^k$
  $z_i^{k+1} = z_i^k$, $w_i^{k+1} = w_i^k$ for $i \notin S^k$
  Send v_i^k to the selected devices $i \in S^k$
for each device $i = 1, \ldots, n$ in parallel **do**
  for participating devices $i \in S^k$ **do**
  $z_i^{k+1} = z_i^k + \eta v_i^k$, $\mathbf{S}_i^k := \mathcal{C}_i^k(h^i(\nabla^2 f_i(z_i^{k+1})) - \mathbf{L}_i^k)$
  $\mathbf{L}_i^{k+1} = \mathbf{L}_i^k + \alpha\mathbf{S}_i^k$, $\mathbf{H}_i^{k+1} = \mathbf{H}_i^k + \alpha\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$
  $l_i^{k+1} = \|[\mathbf{H}_i^{k+1}]_s - \nabla^2 f_i(z_i^{k+1})\|_{\mathrm{F}}$
  Sample $\xi_i^{k+1} \sim$ Bernoulli(p)
  if $\xi_i^k = 1$
  $w_i^{k+1} = z_i^{k+1}$, $g_i^{k+1} = ([\mathbf{H}_i^{k+1}]_s + l_i^{k+1}\mathbf{I})w_i^{k+1} - \nabla f_i(w_i^{k+1})$, send $g_i^{k+1} - g_i^k$ to server
  if $\xi_i^k = 0$
  $w_i^{k+1} = w_i^k$, $g_i^{k+1} = ([\mathbf{H}_i^{k+1}]_s + l_i^{k+1}\mathbf{I})w_i^{k+1} - \nabla f_i(w_i^{k+1})$
  Send $\mathbf{S}_i^k$, $l_i^{k+1} - l_i^k$, and ξ_i^k to server
  for non-participating devices $i \notin S^k$ **do**
  $z_i^{k+1} = z_i^k$, $w_i^{k+1} = w_i^k$, $\mathbf{L}_i^{k+1} = \mathbf{L}_i^k$, $\mathbf{H}_i^{k+1} = \mathbf{H}_i^k$, $l_i^{k+1} = l_i^k$, $g_i^{k+1} = g_i^k$
end for
on server
  if $\xi_i^k = 1$
  $w_i^{k+1} = z_i^{k+1}$, receive $g_i^{k+1} - g_i^k$
  if $\xi_i^k = 0$
  $w_i^{k+1} = w_i^k$, $g_i^{k+1} - g_i^k = \alpha\left[\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}\right]_s w_i^{k+1} + (l_i^{k+1} - l_i^k)w_i^{k+1}$
  $g^{k+1} = g^k + \frac{1}{n}\sum_{i\in S^k}\left(g_i^{k+1} - g_i^k\right)$
  $\mathbf{H}^{k+1} = \mathbf{H}^k + \frac{\alpha}{n}\sum_{i\in S^k}\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$
  $l^{k+1} = l^k + \frac{1}{n}\sum_{i\in S^k}\left(l_i^{k+1} - l_i^k\right)$

Theorem 4.12 (Linear convergence of BL2). *Let Assumption 4.7 hold. Let Assumption 4.3 (i) or Assumption 4.4 (i) hold. Assume* $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{(6H^2+24H_1^2)B_{\mathrm{M}}}$ *and* $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$. *Then*

$$\mathbb{E}[\Phi_3^k] \leq \left(1 - \frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)^k \Phi_3^0,$$

for $k \geq 0$.

The result in Theorem 4.12 is similar to that in Theorem 4.9, and τ represents the expected number of participating devices at each iteration. Define $\Phi_4^k := \mathcal{H}^k + \frac{4BM_1^2}{A_{\mathrm{M}}}\|x^k - x^*\|^2$ for $k \geq 0$. We can obtain the following local superlinear rate.

Theorem 4.13 (Superlinear convergence of BL2). *Let* $\eta = 1$, $\xi^k \equiv 1$, $S^k \equiv [n]$, *and* $\mathcal{Q}_i^k(x) \equiv x$ *for any* $x \in \mathbb{R}^d$ *and* $k \geq 0$. *Let Assumption 4.7 hold. Let Assumption 4.5 (i) or Assumption 4.6 (i) hold. Assume* $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{(6H^2+24H_1^2)B_{\mathrm{M}}}$ *and* $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$. *Then we have*

$$\mathbb{E}[\Phi_4^k] \leq \theta_2^k\Phi_4^0,$$

$$\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] \leq \theta_2^k\left(\frac{A_{\mathrm{M}}(3H^2+12H_1^2)}{16BM_1^2\mu^2} + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_4^0,$$

for $k \geq 0$, *where* $\theta_2 := \left(1 - \frac{\min\{2A, A_{\mathrm{M}}\}}{2}\right)$.

Now, we explore under what conditions we can guarantee the boundedness of $\|z_i^k - x^*\|$ and $\mathcal{H}^k$.

Theorem 4.14. *Let Assumption 4.7 hold. Then we have the following results.*
(i) Let Assumption 4.3 and Assumption 4.5 (ii) hold. If $\|x^0 - x^*\|^2 \leq \tilde{c}_3 := \min\left\{\frac{\mu^2}{d^2(6H^2+24H_1^2)}, \frac{\mu^2}{96d^4N_{\mathrm{B}}R^2M_2^2}\right\}$, *then* $\|z_i^k - x^*\|^2 \leq d\tilde{c}_3$ *and* $\mathcal{H}^k \leq \frac{\mu^2}{96dN_{\mathrm{B}}R^2}$ *for* $i \in [n]$ *and* $k \geq 0$.
(ii) Let Assumption 4.4 and Assumption 4.6 hold. If $\|z_i^0 - x^*\|^2 \leq \tilde{c}_4$, *where* $\tilde{c}_4 := \min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, \frac{AA_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}BM_1^2}\right\}$, *and* $\|\mathbf{L}_i^0 - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ *for all* $i \in [n]$, *then* $\|z_i^k - x^*\|^2 \leq \tilde{c}_4$ *and* $\|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$.

4.2 Basis Learn in $\mathcal{S}^d$

Let $\{\mathbf{B}_i^{jl} \mid j, l \in [d], j \geq l\}$ be a basis in the symmetric subspace $\mathcal{S}^d$ of $\mathbb{R}^{d\times d}$ that consists of all the symmetric matrices for $i \in [n]$. In this case, the number of symmetric matrices in the basis is $\tilde{N} := \frac{d(d+1)}{2}$. Then any symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ can be uniquely represented as

$$\mathbf{A} = \sum_{j\geq l}\tilde{h}_{jl}^i(\mathbf{A})\mathbf{B}^{jl},$$

where $\tilde{h}_{jl}^i(\mathbf{A}) \in \mathbb{R}$ is the coefficient corresponding to $\mathbf{B}_i^{jl}$. Let $\mathbf{B}_i^{lj} := \mathbf{B}_i^{jl}$ for $j > l$ and define $\tilde{h}^i(\mathbf{A})$ as a symmetric matrix in $\mathbb{R}^{d\times d}$ such that $\tilde{h}^i(\mathbf{A})_{jl} := \frac{1}{2}\tilde{h}_{jl}^i$ for $j > l$ and $\tilde{h}^i(\mathbf{A})_{jl} := \tilde{h}_{jl}^i$ for $j = l$. Let $svec : \mathcal{S}^d \to \mathbb{R}^{\tilde{N}}$ be defined as $svec(\mathbf{A}) :=$

$$(\mathbf{A}_{11}, 2\mathbf{A}_{21}..., 2\mathbf{A}_{d1}, ..., \mathbf{A}_{jj}, ..., 2\mathbf{A}_{dj}, ..., \mathbf{A}_{dd})^\top,$$

and $\tilde{\mathcal{M}}_i := (\mathbf{B}_i^{11}, ..., \mathbf{B}_i^{d1}, ..., \mathbf{B}_i^{jj}, ..., \mathbf{B}_i^{dj}, ..., \mathbf{B}_i^{dd})$. Then we have $\mathbf{A} = \tilde{\mathcal{M}}_i svec(\tilde{h}^i(\mathbf{A}))$, which is equivalent to

$$svec(\mathbf{A}) = \tilde{\mathcal{B}}_i \cdot svec(\tilde{h}^i(\mathbf{A})), \tag{14}$$

for any symmetric matrix $\mathbf{A}$, where

$$\tilde{\mathcal{B}}_i := (svec(\mathbf{B}_i^{11}), ..., svec(\mathbf{B}_i^{d1}), ..., svec(\mathbf{B}_i^{jj}), ..., svec(\mathbf{B}_i^{dj}), ..., svec(\mathbf{B}_i^{dd})) \in \mathbb{R}^{\tilde{N}\times\tilde{N}}.$$

Since the representation (14) is unique, we know $\tilde{\mathcal{B}}_i$ is invertible, and thus

$$svec(\tilde{h}^i(\mathbf{A})) = (\tilde{\mathcal{B}}_i)^{-1} svec(\mathbf{A}). \tag{15}$$

Example 4.15. *We choose $\mathbf{B}_i^{jl} \in \mathcal{S}^d$ such that for $j \neq l$, $(\mathbf{B}_i^{jl})_{jl} = (\mathbf{B}_i^{jl})_{lj} = (\mathbf{B}_i^{jl})_{jj} = (\mathbf{B}_i^{jl})_{ll} = 1$ and the other entries are 0; for $j = l$, $(\mathbf{B}_i^{jj})_{jj} = 1$ and the other entries are 0. It is easy to verify it is a basis in $\mathcal{S}^d$, and we also have $\mathbf{B}_i^{jl} \succeq 0$.*

Example 4.16. *For the basis in Example 4.15 and any invertible matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$, it is easy to see $\{\mathbf{Q}\mathbf{B}_i^{jl}\mathbf{Q}^\top\}$ is also a basis in $\mathcal{S}^d$ with $\mathbf{Q}\mathbf{B}_i^{jl}\mathbf{Q}^\top \succeq 0$.*

We choose a basis $\{\mathbf{B}_i^{jl}\}$ in $\mathcal{S}^d$ such that $\mathbf{B}_i^{jl} \succeq 0$ for BL3 (Algorithm 3). To save space, BL3 is put in the Appendix. The way to guarantee the positive definiteness of the Hessian estimator is similar to that in (Islamov et al., 2021). From the definition of γ_i^k, we know $(\mathbf{L}_i^k)_{jl} + 2\gamma_i^k \geq c > 0$. Noticing that $\nabla^2 f_i(z_i^k)$ can be expressed in the form

$$\sum_{jl} \left(\frac{\tilde{h}^i(\nabla^2 f_i(z_i^k))_{jl} + 2\gamma_i^k}{(\mathbf{L}_i^k)_{jl} + 2\gamma_i^k} \cdot ((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - 2\gamma_i^k \right) \mathbf{B}_i^{jl},$$

for β_i^k in Option 2, we have the inequality

$$\begin{aligned}
&\sum_{jl} \left(\beta^k((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - 2\gamma_i^k\right) \mathbf{B}_i^{jl} - \nabla^2 f_i(z_i^k) \\
&= \sum_{jl} \left(\beta^k - \frac{\tilde{h}^i(\nabla^2 f_i(z_i^k))_{jl} + 2\gamma_i^k}{(\mathbf{L}_i^k)_{jl} + 2\gamma_i^k}\right) \cdot ((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k)\mathbf{B}_i^{jl} \\
&\succeq \mathbf{0}.
\end{aligned}$$

Thus, if we can maintain the Hessian estimator in the form $\mathbf{H}_i^k := \sum_{jl} \left(\beta^k((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - 2\gamma_i^k\right) \mathbf{B}_i^{jl}$, then $\mathbf{H}_i^k \succeq \nabla^2 f_i(z_i^k)$ (we can get $\mathbf{H}_i^k \succeq \nabla^2 f_i(z_i^{k-1})$ for Option 1 similarly). To achieve this goal, we use two auxiliary matrices $\mathbf{A}_i^k$, $\mathbf{C}_i^k$, and maintain them as $\mathbf{A}_i^k = \sum_{jl}((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k)\mathbf{B}_i^{jl}$, $\mathbf{C}_i^k = \sum_{jl} 2\gamma_i^k \mathbf{B}_i^{jl}$, and $\mathbf{H}_i^k = \beta^k \mathbf{A}_i^k - \mathbf{C}_i^k$. BL3 follows the same structure of BL2, thus we also need to keep the relation $g_i^k = \mathbf{H}_i^k w_i^k - \nabla f_i(w_i^k)$, which is actually $g_i^k = \beta^k \mathbf{A}_i^k w_i^k - \mathbf{C}_i^k w_i^k - \nabla f_i(w_i^k)$. Since for non-participating devices, β^k usually changes at each step, we split g_i^k into two parts by using two auxiliary vectors $g_{i,1}^k$, $g_{i,2}^k$, and assign $g_{i,1}^k = \mathbf{A}_i^k w_i^k$, $g_{i,2}^k = \mathbf{C}_i^k w_i^k - \nabla f_i(w_i^k)$, and $g_i^k = \beta^k g_{i,1}^k - g_{i,2}^k$. The rest of BL3 is the same as BL2. The convergence results of BL3 are similar to that of BL2 and are listed in the Appendix.

5 Experiments

We conduct numerical experiments to compare the performance of BL methods with various efficient methods in federated learning. We consider regularized logistic regression problem

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) + \frac{\lambda}{2}\|x\|^2 \right\}, \tag{16}$$

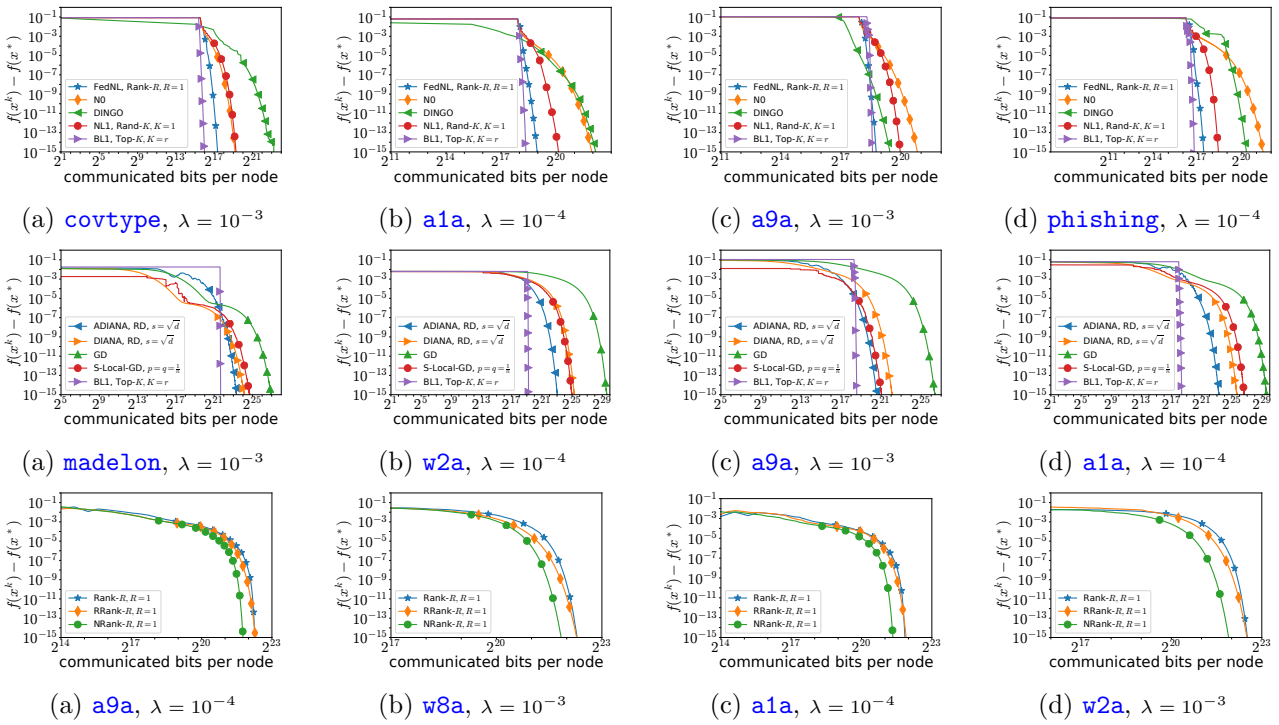
where $f_i(x) = \frac{1}{m}\sum_{j=1}^m \log\left(1 + \exp\left(-b_{ij}a_{ij}^\top x\right)\right)$. Here $\{a_{ij}, b_{ij}\}_{j\in[m]}$ are data points storing at the i-th device. LibSVM data sets were used in the experiments (Chang and Lin, 2011): a1a, a9a, phishing, w2a, w8a, covtype, madelon. We use two values of the regularization parameter: $\lambda \in \{10^{-3}, 10^{-4}\}$. In the figures we plot the relation of the optimality gap $f(x^k) - f(x^*)$ and the number of communicated bits per node. The optimal value $f(x^*)$ is chosen as the function value at the 20-th iterate of standard Newton's method.

5.1 Basis Computation for BL

One of the most popular types of data preprocessing in classical machine learing is dimension reduction. One of such techniques is based on SVD of the feature matrix. We would like to point out that SVD could also be used to find a basis for each client. In our experiments we use linalg.orth function from SciPy module (Jones et al., 2001). In other words, such data preprocessing could be used not only for stability of certain machine learning model, but also for the improvement of optimization process.

5.2 Comparison with Second-order Methods

We compare the performance of BL1 with DINGO (Crane and Roosta, 2019), FedNL (Safaryan et al., 2021), NL1 (Islamov et al., 2021), N0 (Safaryan et al., 2021) in terms of communication complexity. For FedNL, NL1, and BL1 we use $\nabla^2 f_i(x^0)$ as the initialization of $\mathbf{H}_i^0$. Besides, the stepsize $\alpha = 1$, Rank-1 compression for matrices, and option 1 (projection) were used for FedNL. For NL1 compression mechanism is Rand-1 with stepsize $\alpha = 1/(\omega+1)$. Backtracking linesearch for DINGO selects the largest stepsize from $\{1, 2^{-1}, 2^{-2}, \ldots, 2^{-10}\}$. We set the authors' choice for other parameters of the method: $\theta = 10^{-4}, \phi = 10^{-6}, \rho = 10^{-4}$. Compression operator $\mathcal{C}_i^k$ in BL1 is Top-K, where $K = r$ (r is the dimension of the local data). We set $p = 1$ and use identity compression for $\mathcal{Q}^k$ with stepsize $\eta = 1$ for models (backside compression is not used). According the results in Figure 1 (1^{st} row), BL1 is the most efficient method in all cases.



(a) covtype, $\lambda = 10^{-3}$ (b) a1a, $\lambda = 10^{-4}$ (c) a9a, $\lambda = 10^{-3}$ (d) phishing, $\lambda = 10^{-4}$

(a) madelon, $\lambda = 10^{-3}$ (b) w2a, $\lambda = 10^{-4}$ (c) a9a, $\lambda = 10^{-3}$ (d) a1a, $\lambda = 10^{-4}$

(a) a9a, $\lambda = 10^{-4}$ (b) w8a, $\lambda = 10^{-3}$ (c) a1a, $\lambda = 10^{-4}$ (d) w2a, $\lambda = 10^{-3}$

Figure 1: Comparison of BL1 with N0, FedNl, NL1, DINGO (**first row**), DIANA, ADIANA, GD, S-Local-GD (**second row**) and the performance of BL2 with compressors Rank-R, RRank-R, and NRank-R (**third row**) in terms of communication complexity.

5.3 Comparison with First-order Methods

Next we compare the performance of BL1 with vanilla gradient descent GD, DIANA (Mishchenko et al., 2019), ADIANA (Li et al., 2020), and shifted local gradient descent (S-Local-GD) (Gorbunov et al., 2021) in terms of communication complexity. Theoretical stepsizes were chosen for first-order methods. For DIANA and ADIANA we use random dithering compression (Alistarh et al., 2017; Horváth et al., 2019) with $s = \sqrt{d}$ levels. Probabilities p and q are equal to $1/n$ for S-Local-GD. Parameters of BL1 are the same as in the previous section. We clearly see in Figure 1 (2^{nd} row) that BL1 is more communication efficient than all gradient type methods by *several orders in magnitude.*

5.4 Composition of Compressors

In our next experiment we analyse the composition of Rank-R and unbiased compression operators; see Section 3 for more details. We consider BL2 with 3 compression mechanisms: Rank-R, RRank-R (composition of Rank-R and random dithering with $s = \sqrt{d}$ levels), and NRank-R (composition of Rank-R and natural compression). For all three compressors $R = 1$, and initializaion is $\mathbf{H}^0 = \nabla^2 f(x^0)$. Besides, the parameters of BL2 are the following: $\tau = n$, $p = \frac{1}{10}$. Finally, we use Top-K with $K = \lfloor \frac{d}{10} \rfloor$ for $\mathcal{Q}_i^k$. In this experiment we use standard basis in the space of matrices which means that BL2 turns to be FedNL. According to numerical results presented in Figure 1 (3^{rd} row), composition is indeed useful.

6 MORE DISCUSSIONS

In this paper, we consider the basis in $\mathbb{R}^{d\times d}$ and $\mathcal{S}^d$. It is actually possible to extend Basis Learn to the case where $\{\mathbf{B}_i^{jl}\}$ is not necessarily a basis in some space. More precisely, if there exist a set $\{\mathbf{B}_i^j\}_{j\in S^i}$ and a map $h^i : \mathbb{R}^d \to \mathbb{R}^{|S^i|}$ such that for any $x \in \mathbb{R}^d$, $\nabla^2 f_i(x)$ can be represented by $\sum_j h^i(x)_j \mathbf{B}_i^j$ and h^i is L-Lipschitz continuous, i.e., $\|h^i(x) - h^i(y)\| \leq L\|x - y\|$ for any $x, y \in \mathbb{R}^d$, then we can get the corresponding algorithm and convergence results in the same way.

References

Foivos Alimisis, Peter Davies, and Dan Alistarh. Communication-efficient distributed optimization with quantized preconditioners. In *International Conference on Machine Learning (ICML)*, 2021.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient

SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5977–5987, 2018.

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

Chih-Chung Chang and Chih-Jen Lin. LibSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Congliang Chen, Li Shen, Haozhi Huang, and Wei Liu. Quantized adam with error feedback. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–26, 2021.

Rixon Crane and Fred Roosta. Dingo: Distributed newton-type method for gradient-norm optimization. In *Advances in Neural Information Processing Systems*, volume 32, pages 9498–9508, 2019.

Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug. Inexact newton methods. *SIAM J. Numer. Anal., 19(2)*, pages 400–408, 1982.

J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of Computation, 28(126)*, page 549–560, 1974.

Avishek Ghosh, Raj Kumar Maity, Arya Mazumdar, and Kannan Ramchandran. Communication efficient distributed approximate newton method. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2539–2544, 2020. doi: 10.1109/ISIT44484.2020.9174216.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Andreas Griewank. The modification of newton's method for unconstrained optimization by bounding cubic terms. *Technical report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Technical Report NA/12*, 1981.

Vipul Gupta, Avishek Ghosh, Michal Derezinski, Rajiv Khanna, Kannan Ramchandran, and Michael Mahoney. Localnewton: Reducing communication bottleneck for distributed learning. In *37th Conference on Uncertainty in Artificial Intelligence (UAI 2021)*, 2021.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. *International Conference on Machine Learning (ICML)*, 2021.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python. *http://www.scipy.org/*, 2001.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

Dmitry Kovalev, Konstanting Mishchenko, and Peter Richtárik. Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. In *NeurIPS Beyond First Order Methods Workshop*, 2019.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, 2020.

Chieh-Yen Lin, Cheng-Hao Tsai, Ching pei Lee, and Chih-Jen Lin. Large-scale logistic regression and linear support vector machines using spark. *2014 IEEE International Conference on Big Data (Big Data)*, pages 519–528, 2014.

Ji Liu and Ce Zhang. *Distributed Learning Systems with First-Order Methods*, volume 9. Foundations and Trends in Databases, 2020. doi: 10.1561/1900000062.

Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 1993.

Constantin Philippenko and Aymeric Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2021.

Xun Qian, Hanze Dong, Peter Richtárik, and Tong Zhang. Error compensated loopless svrg, quartz, and sdca for distributed optimization. *arXiv preprint arXiv:2109.10049*, 2021.

Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczos, and Alexander J. Smola. AIDE: Fast and communication efficient distributed optimization. *CoRR*, abs/1608.06879, 2016.

Fred Roosta, Yang Liu, Peng Xu, and Michael W. Mahoney. Newton-MR: Newton's Method Without Smoothness or Convexity. *arXiv preprint arXiv:1810.00303*, 2019.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making Newton-Type Methods Applicable to Federated Learning. *arXiv preprint arXiv:2106.02969*, 2021.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-effcient distributed optimization using an approximate newton-type method. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, pages 1000–1008, 2014.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2020.

Shusen Wang, Fred Roosta abd Peng Xu, and Michael W Mahoney. GIANT: Globally improved approximate Newton method for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1306–1316, 2018.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

Hang Xu, Chen-Yu Ho, Ahmed M. Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos Karatsenidis, Marco Canini, and Panos Kalnis. Compressed Communication for Distributed Deep Learning: Survey and Quantitative Evaluation. Technical report, KAUST, Apr 2020. URL `http://hdl.handle.net/10754/662495`.

Jiaqi Zhang, Keyou You, and Tamer Başar. Achieving globally superlinear convergence for distributed optimization with adaptive newton method. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2329–2334, 2020. doi: 10.1109/CDC42340.2020.9304321.

Yuchen Zhang and Lin Xiao. DiSCO: Distributed optimization for self-concordant empirical loss. In *In Proceedings of the 32nd International Conference on Machine Learning, PMLR, volume 37, pages 362–370*, 2015.

Yong Zhuang, Wei-Sheng Chin, Yu-Chin Juan, and Chih-Jen Lin. Distributed newton methods for regularized logistic regression. In Tru Cao, Ee-Peng Lim, Zhi-Hua Zhou, Tu-Bao Ho, David Cheung, and Hiroshi Motoda, editors, *Advances in Knowledge Discovery and Data Mining*, pages 690–703, Cham, 2015. Springer International Publishing. ISBN 978-3-319-18032-8.

Supplementary Material: Basis Matters: Better Communication-Efficient Second Order Methods for Federated Learning

Contents

A CONVERGENCE RESULTS FOR BL3

Algorithm 3 BL3

Parameters: learning rate $\alpha > 0$, positive constant $c > 0$, minibatch size $\tau \in \{1, 2, \ldots, n\}$

Initialization: $\mathbf{B}_i^{jl} \succeq 0$; $w_i^0 = z_i^0 = x^0$ for $i \in [n]$; $\mathbf{L}_i^0 \in \mathbb{R}^{d\times d}$; $\gamma_i^0 = \max_{jl}\{c, |(\mathbf{L}_i^0)_{jl}|\}$; $\mathbf{A}_i^0 = \sum_{jl}((\mathbf{L}_i^0)_{jl} + 2\gamma_i^0)\mathbf{B}_i^{jl}$; $\mathbf{C}_i^0 = \sum_{jl} 2\gamma_i^0 \mathbf{B}_i^{jl}$; $\mathbf{A}^0 = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i^0$; $\mathbf{C}^0 = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^0$; $\beta_i^0 = \max_{jl} \frac{\tilde{h}^i(\nabla^2 f_i(w_i^0))_{jl} + 2\gamma_i^0}{(\mathbf{L}_i^0)_{jl} + 2\gamma_i^0}$; $\beta^0 = \max_i\{\beta_i^0\}$; $g_{i,1}^0 = \mathbf{A}_i^0 w_i^0$; $g_{i,2}^0 = \mathbf{C}_i^0 w_i^0 + \nabla f_i(w_i^0)$; $g_1^0 = \frac{1}{n}\sum_{i=1}^n g_{i,1}^0$; $g_2^0 = \frac{1}{n}\sum_{i=1}^n g_{i,2}^0$; $\mathbf{H}^0 = \beta^0 \mathbf{A}^0 - \mathbf{C}^0$; $g^0 = \beta^0 g_1^0 - g_2^0$

for $k = 0, 1, 2, \ldots$ **do**

on server

$x^{k+1} = \left(\mathbf{H}^k\right)^{-1} g^k$ — Main step: Update the global model

Choose a subset $S^k \subseteq \{1, \ldots, n\}$ such that $\mathbb{P}[i \in S^k] = \tau/n$ for all $i \in [n]$

$v_i^k = \mathcal{Q}_i^k(x^{k+1} - z_i^k), \quad z_i^{k+1} = z_i^k + \eta v_i^k$ for $i \in S^k$

$z_i^{k+1} = z_i^k, \quad w_i^{k+1} = w_i^k$ for $i \notin S^k$

Send v_i^k to the selected devices $i \in S^k$ — Communicate to selected clients

for each node $i = 1, \ldots, n$ **do**

for participating devices $i \in S^k$ **do**

$z_i^{k+1} = z_i^k + \eta v_i^k$, $\mathbf{L}_i^{k+1} = \mathbf{L}_i^k + \alpha \mathcal{C}_i^k\left(\tilde{h}^i(\nabla^2 f_i(z_i^{k+1})) - \mathbf{L}_i^k\right), \quad \gamma_i^{k+1} = \max_{jl}\{c, |(\mathbf{L}_i^{k+1})_{jl}|\}$

Option 1: $\beta_i^{k+1} = \max_{jl} \frac{\tilde{h}^i(\nabla^2 f_i(z_i^k))_{jl} + 2\gamma_i^{k+1}}{(\mathbf{L}_i^{k+1})_{jl} + 2\gamma_i^{k+1}}$

Option 2: $\beta_i^{k+1} = \max_{jl} \frac{\tilde{h}^i(\nabla^2 f_i(z_i^{k+1}))_{jl} + 2\gamma_i^{k+1}}{(\mathbf{L}_i^{k+1})_{jl} + 2\gamma_i^{k+1}}$

$\mathbf{A}_i^{k+1} = \mathbf{A}_i^k + \sum_{jl}\left((\mathbf{L}_i^{k+1})_{jl} - (\mathbf{L}_i^k)_{jl} + 2\gamma_i^{k+1} - 2\gamma_i^k\right)\mathbf{B}_i^{jl}, \quad \mathbf{C}_i^{k+1} = \mathbf{C}_i^k + \sum_{jl}\left(2\gamma_i^{k+1} - 2\gamma_i^k\right)\mathbf{B}_i^{jl}$

$\xi_i^k = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

if $\xi_i^k = 1$

$w_i^{k+1} = z_i^{k+1}$, $g_{i,1}^{k+1} = \mathbf{A}_i^{k+1} w_i^{k+1}$, $g_{i,2}^{k+1} = \mathbf{C}_i^{k+1} w_i^{k+1} + \nabla f_i(w_i^{k+1})$

Send $g_{i,1}^{k+1} - g_{i,1}^k$, $g_{i,2}^{k+1} - g_{i,2}^k$ to server

if $\xi_i^k = 0$

$w_i^{k+1} = w_i^k$, $g_{i,1}^{k+1} = \mathbf{A}_i^{k+1} w_i^{k+1}$, $g_{i,2}^{k+1} = \mathbf{C}_i^{k+1} w_i^{k+1} + \nabla f_i(w_i^{k+1})$

Send $\mathbf{L}_i^{k+1} - \mathbf{L}_i^k$, β_i^{k+1}, ξ_i^k, $\gamma_i^{k+1} - \gamma_i^k$ to server

for non-participating devices $i \notin S^k$ **do**

$z_i^{k+1} = z_i^k$, $w_i^{k+1} = w_i^k$, $\mathbf{L}_i^{k+1} = \mathbf{L}_i^k$, $\gamma_i^{k+1} = \gamma_i^k$, $\beta_i^{k+1} = \beta_i^k$, $\mathbf{A}_i^{k+1} = \mathbf{A}_i^k$, $\mathbf{C}_i^{k+1} = \mathbf{C}_i^k$, $g_{i,1}^{k+1} = g_{i,1}^k$, $g_{i,2}^{k+1} = g_{i,2}^k$

end for

on server

if $\xi_i^k = 1$

$w_i^{k+1} = z_i^{k+1}$, Receive $g_{i,1}^{k+1} - g_{i,1}^k$, $g_{i,2}^{k+1} - g_{i,2}^k$,

if $\xi_i^k = 0$

$w_i^{k+1} = w_i^k$, $g_{i,1}^{k+1} - g_{i,1}^k = \sum_{jl}(\mathbf{L}_i^{k+1} - \mathbf{L}_i^k)_{jl}\mathbf{B}_i^{jl} w_i^{k+1} + 2(\gamma_i^{k+1} - \gamma_i^k) w_i^{k+1}$

$g_{i,2}^{k+1} - g_{i,2}^k = \sum_{jl} 2(\gamma_i^{k+1} - 2\gamma_i^k)\mathbf{B}_i^{jl} w_i^{k+1}$

$g_1^{k+1} = g_1^k + \frac{1}{n}\sum_{i\in S^k}\left(g_{i,1}^{k+1} - g_{i,1}^k\right)$, $g_2^{k+1} = g_2^k + \frac{1}{n}\sum_{i\in S^k}\left(g_{i,2}^{k+1} - g_{i,2}^k\right)$

$\beta^{k+1} = \max_i\{\beta_i^{k+1}\}, \quad g^{k+1} = \beta^{k+1} g_1^{k+1} - g_2^{k+1}$

$\mathbf{A}^{k+1} = \mathbf{A}^k + \frac{1}{n}\sum_{i\in S^k}\sum_{jl}\left((\mathbf{L}_i^{k+1})_{jl} - (\mathbf{L}_i^k)_{jl} + 2\gamma_i^{k+1} - 2\gamma_i^k\right)\mathbf{B}_i^{jl}$

$\mathbf{C}^{k+1} = \mathbf{C}^k + \frac{1}{n}\sum_{i\in S^k}\sum_{jl}\left(2\gamma_i^{k+1} - 2\gamma_i^k\right)\mathbf{B}_i^{jl}, \quad \mathbf{H}^{k+1} = \beta^{k+1}\mathbf{A}^{k+1} - \mathbf{C}^{k+1}$

end for

Assumption A.1. *Assume* $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \le H\|x - y\|$ *for any* $x, y \in \mathbb{R}^d$ *and* $i \in [n]$. *Assume* $\max_{jl}\{|(\mathbf{L}_i^k)_{jl}|\} \le M_3$ *for all* $i \in [n]$ *and* $k \ge 0$. *Assume* $\|\tilde{h}^i(\nabla^2 f_i(x)) - \tilde{h}^i(\nabla^2 f_i(y))\|_{\mathrm{F}} \le M_4\|x - y\|$, $\max_{jl}\{|\tilde{h}^i(\nabla^2 f_i(x))_{jl} - \tilde{h}^i(\nabla^2 f_i(y))_{jl}|\} \le M_5\|x - y\|$ *for any* $x, y \in \mathbb{R}^d$ *and* $i \in [n]$, *and* $\max_{jl}\{\|\mathbf{B}_i^{jl}\|_{\mathrm{F}}\} \le R$ *for* $i \in [n]$. *Assume each* f_i *is* μ*-strongly convex.*

We estimate M_3, M_4, and M_5 in Assumption A.1 in the following lemma.

Lemma A.2. *(i) Assume* $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x-y\|$ *and* $\max_{jl}\{|(\nabla^2 f_i(x))_{jl} - (\nabla^2 f_i(y))_{jl}|\} \leq \nu\|x-y\|$ *for any* $x, y \in \mathbb{R}^d$*, and* $i \in [n]$*. Then we have* $M_4 \leq \sqrt{2}\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|\}H_1$ *and* $M_5 \leq 2\nu\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty\}$.
(ii) Assume $\max_{jl}\{|(\nabla^2 f_i(x))_{jl}|\} \leq \gamma$ *for any* $x \in \mathbb{R}^d$ *and* $i \in [n]$*. If Assumption 4.5 (ii) holds, then* $M_3 \leq 2\gamma\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty\}$.
(iii) Assume $\|\nabla^2 f_i(x)\|_{\mathrm{F}} \leq \tilde{\gamma}$ *for any* $x \in \mathbb{R}^d$ *and* $i \in [n]$*. If Assumption 4.6 holds and* $\|\mathbf{L}_i^0\|_{\mathrm{F}} \leq \frac{\sqrt{2B}}{\sqrt{A}}\|(\tilde{\mathcal{B}}_i)^{-1}\|\tilde{\gamma}$ *for all* $i \in [n]$*, then we have* $\|\mathbf{L}_i^k\|_{\mathrm{F}} \leq \frac{\sqrt{2B}}{\sqrt{A}}\|(\tilde{\mathcal{B}}_i)^{-1}\|\tilde{\gamma}$ *for* $k \geq 0$ *and* $i \in [n]$*, and* $M_3 \leq \frac{\sqrt{2B}\tilde{\gamma}}{\sqrt{A}}\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|\}$.

Let $\Phi_5^k := \mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k$, where $\mathcal{Z}^k := \frac{1}{n}\sum_{i=1}^n \|z_i^k - x^*\|^2$, for $k \geq 0$.

Theorem A.3 (Linear convergence of BL3)**.** *Let Assumption A.1 hold. Let Assumption 4.3 (i) or Assumption 4.4 (i) hold. Assume* $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4(H^2+4c_1)B_{\mathrm{M}}}$ *and* $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$*, where* $c_1 := \frac{4N^2R^2M_5^2(M_3+2\max\{c,M_3\})^2}{c^2}$ *and* $c_2 := 2NR^2\left(1 + \frac{2N(M_3+2\max\{c,M_3\})^2}{c^2}\right)$*. Then we have*

$$\mathbb{E}[\Phi_5^k] \leq \left(1 - \frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)^k \Phi_5^0,$$

for $k \geq 0$.

Define $\Phi_6^k := \mathcal{H}^k + \frac{4BM_4^2}{A_{\mathrm{M}}}\|x^k - x^*\|^2$ for $k \geq 0$.

Theorem A.4 (Superlinear convergence of BL3)**.** *Let* $\eta = 1$*,* $\xi^k \equiv 1$*,* $S^k \equiv [n]$*, and* $\mathcal{Q}_i^k(x) \equiv x$ *for any* $x \in \mathbb{R}^d$ *and* $k \geq 0$*. Let Assumption A.1 hold. Let Assumption 4.5 (i) or Assumption 4.6 (i) hold. Assume* $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4(H^2+4c_1)B_{\mathrm{M}}}$ *and* $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$*. Then we have*

$$\mathbb{E}[\Phi_6^k] \leq \theta_3^k\Phi_6^0,$$

for $k \geq 0$*, where* $\theta_3 := \left(1 - \frac{\min\{2A, A_{\mathrm{M}}\}}{2}\right)$*. Moreover, for Option 1, we have*

$$\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] \leq \theta_3^k\left(\frac{A_{\mathrm{M}}(H^2\theta_3+4c_1)}{8BM_4^2\mu^2\theta_3} + \frac{2c_2}{\mu^2}\right)\Phi_6^0,$$

and for Option 2, we have

$$\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] \leq \theta_3^k\left(\frac{A_{\mathrm{M}}(H^2+4c_1)}{8BM_4^2\mu^2} + \frac{2c_2}{\mu^2}\right)\Phi_6^0,$$

for $k \geq 0$.

Next, we explore under what conditions we can guarantee the boundedness of $\|z_i^k - x^*\|^2$ and $\mathcal{H}^k$.

Theorem A.5. *Let Assumption A.1 hold. Then we have the following results.*
(i) Let Assumption 4.3 and Assumption 4.5 (ii) hold. If

$$\|x^0 - x^*\|^2 \leq \min\left\{\frac{\mu^2}{4d^2(H^2+4c_1)}, \frac{\mu^2}{16d^4c_2M_5^2}\right\},$$

then $\|z_i^k - x^*\|^2 \leq \min\left\{\frac{\mu^2}{4d(H^2+4c_1)}, \frac{\mu^2}{16d^3c_2M_5^2}\right\}$ *and* $\mathcal{H}^k \leq \frac{\mu^2}{16dc_2}$ *for* $i \in [n]$ *and* $k \geq 0$.
(ii) Let Assumption 4.4 and Assumption 4.6 hold. If $\|z_i^0 - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, \frac{AA_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}BM_4^2}\right\}$ *and* $\|\mathbf{L}_i^0 - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$ *for all* $i \in [n]$*, then* $\|z_i^k - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, \frac{AA_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}BM_4^2}\right\}$ *and* $\|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$ *for all* $i \in [n]$ *and* $k \geq 0$.

B EXTRA EXPERIMENTS

In this section we demonstrate additional numerical experiments comparing BL with relevant benchmarks and with state-of-the-art methods. We consider regularized logistic regression problem

$$\min_{x\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{i=1}^{n} f_i(x)+\frac{\lambda}{2}\|x\|^2\right\}, \quad f_i(x)=\frac{1}{m}\sum_{j=1}^{m}\log\left(1+\exp(-b_{ij}a_{ij}^\top x)\right),$$

where $\{a_{ij}, b_{ij}\}_{j\in[m]}$ are data samples belonging i-th node.

B.1 Parameters Setting and Data Sets

The data sets were taken from LibSVM library (Chang and Lin, 2011): a1a, a9a, phishing, covtype, madelon, w2a, w8a. Each data set was partitioned across several nodes to cover a variety of scenarios. See Table 1 for more detailed description.

Table 1: Data sets used in the experiments with the number of worker nodes n used in each case.

data set	# workers n	# data points $(=nm)$	# features d	average dimension r
a1a	16	1600	123	64
a9a	80	32560	123	82
phishing	100	110	68	35
covtype	200	581000	54	24
madelon	10	2000	500	200
w2a	50	3450	300	59
w8a	142	49700	300	133

Theoretical parameters were used for gradient type methods: vanilla gradient descent (GD), DIANA (Mishchenko et al., 2019), ADIANA (Li et al., 2020), and local gradient descent (Local-GD). The parameter constants for DINGO (Crane and Roosta, 2019) were chosen following authors' choice: $\theta=10^{-4}, \phi=10^{-6}, \rho=10^{-4}$. Backtracking line search was used for DINGO to find the largest stepsize from $\{1, 2^{-1}, \ldots, 2^{-10}\}$. The initialization of $\mathbf{H}_i^0$ for NL1 (Islamov et al., 2021) and vanilla FedNL (Safaryan et al., 2021) is $\nabla^2 f_i(x^0)$. Besides, for NL1 we use Rand-K compressor with $K=1$ and the stepsize $\alpha=\frac{1}{\omega+1}$, where $\omega=\frac{m}{K}-1$. For FedNL we use option 1 to make the Hessian approximation to be positive definite (projection onto the cone of positive definite matrices), stepsize $\alpha=1$, and compression operator Rank-R with $R=1$. For BL3, we use option 2.

We carry out experiments for two values of regularization parameter $\lambda\in\{10^{-3}, 10^{-4}, 10^{-5}\}$. In the figures we plot the optimality gap $f(x^k)-f(x^*)$ versus the number of communicated bits per node. The optimal value $f(x^*)$ is chosen as the function value at the 20-th iterate of standard Newton's method.

B.2 Compression Operators

B.2.1 Unbiased Compression Operator: Random Dithering

In all experiments with ADIANA and DIANA the compression operator applied on gradient differences is random dithering (Alistarh et al., 2017; Horváth et al., 2019). This compressor has the parameter s (number of levels) and can be defined via the formula

$$\mathcal{C}(x) := \operatorname{sign}(x)\cdot\|x\|_q\cdot\frac{\xi_s}{s}, \tag{17}$$

where $\|x\|_q := \left(\sum_i |x_i|^q\right)^{1/q}$ and $\xi_s\in\mathbb{R}^d$ is a random vector whose i-th entire defind as follows

$$(\xi_s)_i=\begin{cases} l+1 & \text{with probability } \frac{|x_i|}{\|x\|_q}s-l,\\ l & \text{otherwise.}\end{cases} \tag{18}$$

Here $s \in \mathbb{N}_+$ denotes the levels of rounding, and l satisfies $\frac{|x_i|}{\|x\|_q} \in \left[\frac{l}{s}, \frac{l+1}{s}\right]$. This compressor has the variance parameter $\omega \leq 2 + \frac{d^{1/2}+d^{1/q}}{s}$ (Horváth et al., 2019). However, for Euclidean norm ($q = 2$) one can improve the bound by $\omega \leq \min\left\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right\}$ (Alistarh et al., 2017).

B.2.2 Examples of Contractive Compression Operators for Matrices

One of the examples of contractive compression operators is low-rank approximation or Rank-R compressor. This compression operator is based on singular value decomposition of the matrix and belongs to the class of contractive compressors with $\delta = \frac{R}{d}$ (Safaryan et al., 2021). Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ and singular value decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \sum_{i=1}^{d} \sigma_i u_i v_i^\top, \tag{19}$$

where the singular values σ_i are sorted in non-increasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$. Then, the Rank-$R$ compressor, for $R \leq d$, is defined by

$$\mathcal{C}(\mathbf{X}) := \sum_{i=1}^{R} \sigma_i u_i v_i^\top. \tag{20}$$

Note that if the input of Rank-R compressor is a symmetric matrix, then its output is automatically symmetric matrix.

Another popular choice of contractive compressors in practice is Top-K. This compressor applied on matrices sorts the entires of input in non-increasing order by magnitude, and then selects K maximal elements. Top-K compressor belongs to the class of contractive compressors with $\delta = \frac{d^2}{K}$. For arbitrary matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$ let sort its entires in non-increasing order by magnitude, i.e., $X_{i_k j_k}$ is the k-th maximal element of $\mathbf{X}$ by magnitude. Let $\{\mathbf{B}_{ij}\}_{i,j=1}^{d}$ be a standard basis in the space of matrices. Then, the Top-K compression operator can be defined via

$$\mathcal{C}(\mathbf{X}) := \sum_{k=1}^{K} X_{i_k j_k} \cdot \mathbf{B}_{i_k j_k}. \tag{21}$$

One way how to make the output of this compressor to be a symmetric matrix is to apply Top-K on upper triangular part of the input.

B.3 Example of Unbiased Compression Operators for Matrices

The simplest example of unbiased compressor which could be applied on matrices is random sparsification operator or Rand-K. This compressor belongs to the class of unbiased compressors with $\omega = \frac{d^2}{K} - 1$. For the input matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$ we choose a set $\mathcal{S}_K$ of indexes (i, j) of cardinality K uniformly at random. Then Rand-K compressor can be defined via

$$\mathcal{C}\left(\mathbf{X}\right)_{ij} := \begin{cases} \frac{d^2}{K} X_{ij} & \text{if } (i,j) \in \mathcal{S}_K, \\ 0 & \text{if } (i,j) \notin \mathcal{S}_K. \end{cases} \tag{22}$$

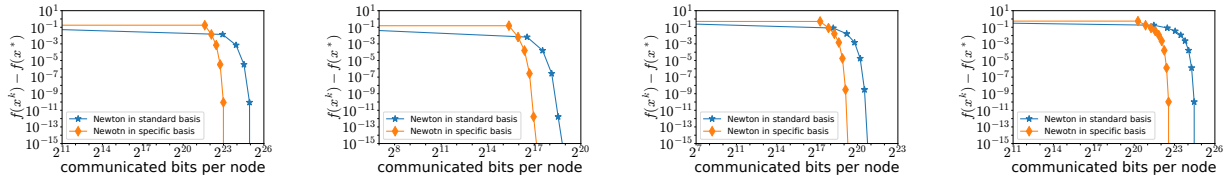
The way how to make the output of Rand-K to be a symmetric matrix is exaclty the same as for Top-K.

B.4 The Performance of Newton's Method in Different Basis

First, we investigate how the performance of Newton's method is influenced by the choice of the basis. We compare the efficiency of Newton's method on two bases: the one that was described in Section 2.3 and the standard one. The results are presented in Figure 2. We clearly see that Newton's method in the specific basis is approximately 4 times more communication-efficient than in standard one.

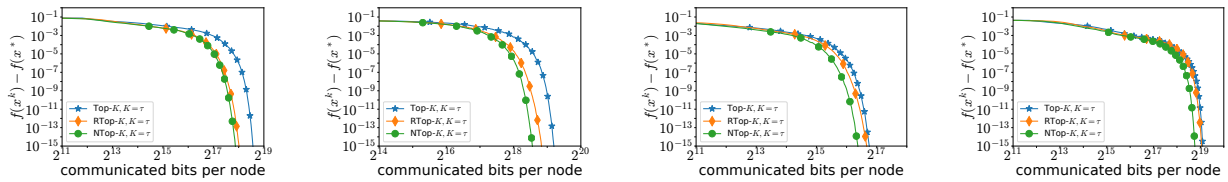
B.5 Composition of Top-K and Unbiased Compressor

Next, we study other type of composition of compression operators. We investigate how composition of Top-K and unbiased compression operator (Qian et al., 2021) influences the performance of BL2. We compare the performance of BL2 with Top-K ($K = r$), RTop-K ($K = r$) (composition of Top-K and random dithering with



madelon, $\lambda = 10^{-3}$ covtype, $\lambda = 10^{-3}$ phishing, $\lambda = 10^{-4}$ w8a, $\lambda = 10^{-4}$

Figure 2: The performance of Newton's method in different basis in terms of communication complexity.

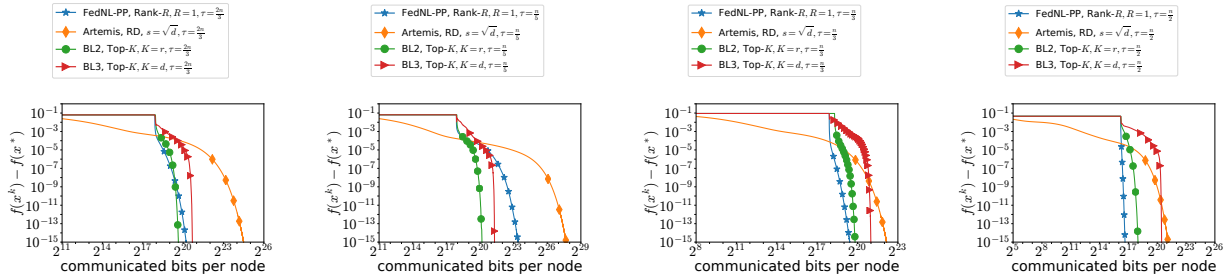


a9a, $\lambda = 10^{-3}$ w2a, $\lambda = 10^{-3}$ phishing, $\lambda = 10^{-4}$ a1a, $\lambda = 10^{-4}$

Figure 3: The performance of BL2 with different types of compression operators: Top-K, RTop-K (composition of Top-K and random dithering with $s = \sqrt{K}$), and NTop-K (composition of Top-K and natrual compression).

$s = \sqrt{K}$), and NTop-K ($K = r$) (composition of Top-K and natural compression). The initialization of $\mathbf{H}^0$ is $\nabla^2 f(x^0)$. Besides, we use the basis that was decribed in Section 2.3. We set the following parameters for BL2: $p = \frac{r}{2d}$, $\tau = n$, and Top-K with $K = \lfloor \frac{r}{2} \rfloor$ for models in the experiments on w2a, a1a data sets. In the experiments on a9a, phishing data sets, these parameters are $p = \frac{r}{4d}$, $\tau = n$, and Top-K, ($K = \lfloor \frac{r}{4} \rfloor$) compressor for models. The results are presented in Figure 3. According to numerical results, we can conclude that composition of Top-K and natural compression is the most efficient compressor in all cases. However, RTop-K have almost the same performance as Top-K on data sets a1a, a9a.

B.6 The Effect of Partial Participation



a1a, $\lambda = 10^{-4}$ a1a, $\lambda = 10^{-5}$ a9a, $\lambda = 10^{-4}$ phishing, $\lambda = 10^{-5}$

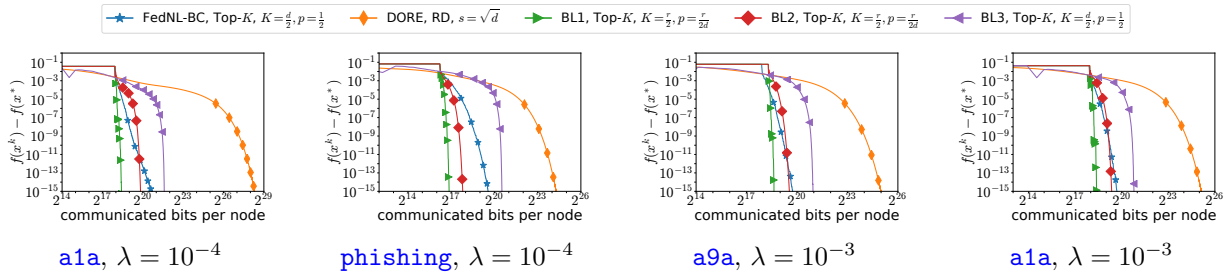
Figure 4: The comparison of FedNL-PP, BL2, BL3, and Artemis with partial device participation in terms of communication complexity.

In this section we study the effect of partial participation. For FedNL-PP (Safaryan et al., 2021) we use stepsize $\alpha = 1$ and Rank-R ($R = 1$) compression operator. The specific basis described in Section 2.3 were used for BL2. Besides, the parameters of this method are the following: compression operator $\mathcal{C}_i^k$ is Top-K with $K = r$, $p = 1$. The basis for BL3 were chosen from the Example 4.15. We use Top-K compressor with $K = d$ for $\mathcal{C}_i^k$, and set $p = 1$ for this method. Both for BL2 and BL3 stepsizes are $\alpha = \eta = 1$, model comressor $\mathcal{Q}_i^k$ is identity. Random dithering with $s = \sqrt{d}$ levels was used for Artemis (Philippenko and Dieuleveut, 2021). In different cases we set the number of active devices τ equal to various fractions of n. The results of the experiment are presented in

Figure 4. According to the plots, BL2 and FedNL-PP are the best methods, they outperform each depending on data set. BL3 also outperform FedNL-PP on a1a ($\lambda = 10^{-5}$) data set. In almost all cases FedNL-PP and BL2 outperform Artemis be *many orders in magnitude*. We can conclude that specific for the problem basis could be beneficial.

B.7 Bidirectional Compression

In our next test we compare FedNL-BC (Safaryan et al., 2021), BL1, BL2, BL3, and DORE (Liu et al., 2020). The parameters of FedNL-BC are the following: matrix compression operator is Top-K, $K = \lfloor \frac{d}{2} \rfloor$; model compression operator is Top-K, $K = \lfloor \frac{d}{2} \rfloor$; stepsizes are $\alpha = \eta = 1$; probability $p = 1$. We use option 1 (projection) to make Hessian approximation to be positive definite. Next, we use the basis described in Section 2.3 for BL1 and BL2. We use Top-K, $K = \lfloor \frac{r}{2} \rfloor$, for matrices and models compression, probability $p = \frac{r}{2d}$, and stepsizes $\alpha = \eta = 1$. The basis for BL3 is described in Example 4.15 in the main paper. Besides, this method has the following parameters: Top-K, $K = \lfloor \frac{d}{2} \rfloor$ for models and Hessians compression; stepsize $\alpha = \eta = 1$; probability $p = \frac{1}{2}$. Finally, all devices are active for BL2 and BL3, i.e. $\tau = n$. The results of this test can be found in Figure 5.



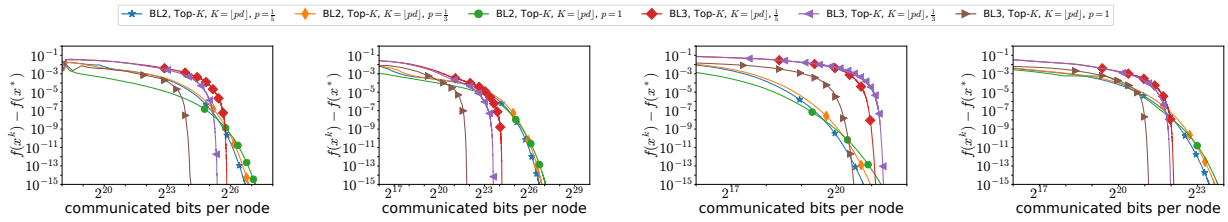
a1a, $\lambda = 10^{-4}$ | phishing, $\lambda = 10^{-4}$ | a9a, $\lambda = 10^{-3}$ | a1a, $\lambda = 10^{-3}$

Figure 5: The comparison of FedNL-BC, BL1, BL2, BL3 and DORE with bidirectional compression in terms of communication complexity.

We see that all second-order methods outperform DORE in terms of communication complexity by *many orders in magnitude*. Moreover, we can conclude that specific to the problem basis is helpful since BL1 and BL2 outperform FedNL-BC.

B.8 Comparison of BL2 and BL3

Finally, we compare BL2 and BL3 with bidirectional compression and partial participation simultaneously. We set the number of active devices to $\frac{n}{2}$. For BL2 we use standard basis in the space of matrices, for BL3 the basis is one that was given in the example 4.15. For both methods the compression operator is Top-K, $K = \lfloor pd \rfloor$, both for models and matrices. The gradient compressor is lazy Bernoulli compressor with parameter p. We set $p \in \{1, 1/3, 1/5\}$. In the Figure 6 we plot the optimality gap $f(x^k) - f(x^*)$ versus the average number of communicated bits per node.



a1a, $\lambda = 10^{-3}$ | a1a, $\lambda = 10^{-4}$ | phishing, $\lambda = 10^{-3}$ | phishing, $\lambda = 10^{-4}$

Figure 6: The comparison of BL2 and BL3 with bidirectional compression and partial participation in terms of communication complexity.

The first observation from the numerical results is that BL2 is less communication-efficient method than BL3. However, if we use specific basis for BL2, then it improves the performance of the method; in Figure 5 BL2 is better

than BL3. Besides, we clearly see that bicompression improves the performance of BL2 in partial participation setting. However, this is not the case for BL3.

C PROOFS OF LEMMA 3.1 AND PROPOSITION 3.2

C.1 Proof of Lemma 3.1

(i) We have

$$\begin{aligned}
&\left\|\frac{(\mathbf{B}+\mathbf{B}^\top)}{2}-\mathbf{A}\right\|_{\mathrm{F}}^2-\|\mathbf{B}-\mathbf{A}\|_{\mathrm{F}}^2\\
&=\frac{1}{4}\|\mathbf{B}+\mathbf{B}^\top\|_{\mathrm{F}}^2+\|\mathbf{A}\|_{\mathrm{F}}^2-\langle\mathbf{B}+\mathbf{B}^\top,\mathbf{A}\rangle-\|\mathbf{B}\|_{\mathrm{F}}^2-\|\mathbf{A}\|_{\mathrm{F}}^2+2\langle\mathbf{B},\mathbf{A}\rangle\\
&=\frac{1}{4}\|\mathbf{B}\|_{\mathrm{F}}^2+\frac{1}{4}\|\mathbf{B}^\top\|_{\mathrm{F}}^2+\frac{1}{2}\langle\mathbf{B},\mathbf{B}^\top\rangle-\|\mathbf{B}\|_{\mathrm{F}}^2+\langle\mathbf{B}-\mathbf{B}^\top,\mathbf{A}\rangle\\
&=\frac{1}{2}\langle\mathbf{B},\mathbf{B}^\top\rangle-\frac{1}{2}\|\mathbf{B}\|_{\mathrm{F}}^2+\langle\mathbf{B}-\mathbf{B}^\top,\mathbf{A}\rangle\\
&\leq\langle\mathbf{B}-\mathbf{B}^\top,\mathbf{A}\rangle\\
&=0,
\end{aligned}$$

where the first inequality comes from the Cauchy-Schwartz inequality, and the last equality comes from the fact that $\mathbf{A}$ is symmetric.

(ii) From (i), for any $\mathbf{A}\in\mathbb{R}^{d\times d}$ we have

$$\mathbb{E}\|\tilde{\mathcal{C}}(\mathbf{A})-\mathbf{A}\|_{\mathrm{F}}^2\leq\mathbb{E}\|\mathcal{C}(\mathbf{A})-\mathbf{A}\|_{\mathrm{F}}^2\leq(1-\delta)\|\mathbf{A}\|_{\mathrm{F}}^2.$$

C.2 Proof of Proposition 3.2

From the definition of $\mathcal{C}_1$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{C}_1(\mathbf{A})-\mathbf{A}\|_{\mathrm{F}}^2]&=\mathbb{E}\|\mathcal{C}_1(\mathbf{A})\|_{\mathrm{F}}^2+\|\mathbf{A}\|_{\mathrm{F}}^2-2\mathbb{E}[\langle\mathcal{C}_1(\mathbf{A}),\mathbf{A}\rangle]\\
&=\|\mathbf{A}\|_{\mathrm{F}}^2+\sum_{i=1}^R\mathbb{E}\left[\frac{\sigma_i^2\mathcal{Q}_2^i(b_iv_i)^\top\mathcal{Q}_2^i(b_iv_i)\mathcal{Q}_1^i(a_iu_i)^\top\mathcal{Q}_1^i(a_iu_i)}{a_i^2b_i^2(\omega_1+1)^2(\omega_2+1)^2}\right]\\
&\quad+\sum_{i,j\in[R],i\neq j}\mathbb{E}\left[\frac{\sigma_i\sigma_j\mathcal{Q}_2^i(b_iv_i)^\top\mathcal{Q}_2^j(b_jv_j)\mathcal{Q}_1^j(a_ju_j)^\top\mathcal{Q}_1^i(a_iu_i)}{a_ia_jb_ib_j(\omega_1+1)^2(\omega_2+1)^2}\right]\\
&\quad-2\left\langle\sum_{i=1}^R\frac{\sigma_iu_iv_i^\top}{(\omega_1+1)(\omega_2+1)},\mathbf{A}\right\rangle\\
&=\|\mathbf{A}\|_{\mathrm{F}}^2+\sum_{i=1}^R\frac{\sigma_i^2\mathbb{E}\|\mathcal{Q}_2^i(b_iv_i)\|^2\cdot\mathbb{E}\|\mathcal{Q}_1^i(a_iu_i)\|^2}{a_i^2b_i^2(\omega_1+1)^2(\omega_2+1)^2}-2\left\langle\sum_{i=1}^R\frac{\sigma_iu_iv_i^\top}{(\omega_1+1)(\omega_2+1)},\mathbf{A}\right\rangle,
\end{aligned}$$

where in the last two equalities, we use the independence of each $\mathcal{Q}_1^i, \mathcal{Q}_2^i$, and the fact that $u_j^\top u_i = 0$ and $v_i^\top v_j = 0$ for $i \neq j$. From the definition of unbiased compressors, we further have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{C}_1(\mathbf{A}) - \mathbf{A}\|_{\mathrm{F}}^2] &\leq \|\mathbf{A}\|_{\mathrm{F}}^2 + \sum_{i=1}^{R} \frac{\sigma_i^2 \|u_i\|^2 \|v_i\|^2}{(\omega_1+1)(\omega_2+1)} - 2\left\langle \sum_{i=1}^{R} \frac{\sigma_i u_i v_i^\top}{(\omega_1+1)(\omega_2+1)}, \mathbf{A} \right\rangle \\
&= \left(1 - \frac{1}{(\omega_1+1)(\omega_2+1)}\right) \|\mathbf{A}\|_{\mathrm{F}}^2 \\
&\quad + \frac{1}{(\omega_1+1)(\omega_2+1)} \left(\|\mathbf{A}\|_{\mathrm{F}}^2 + \sum_{i=1}^{R} \sigma_i^2 \|u_i\|^2 \|v_i\|^2 - 2\left\langle \sum_{i=1}^{R} \sigma_i u_i v_i^\top, \mathbf{A} \right\rangle \right) \\
&= \left(1 - \frac{1}{(\omega_1+1)(\omega_2+1)}\right) \|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{1}{(\omega_1+1)(\omega_2+1)} \left\| \sum_{i=1}^{R} \sigma_i u_i v_i^\top - \mathbf{A} \right\|_{\mathrm{F}}^2 \\
&\leq \left(1 - \frac{1}{(\omega_1+1)(\omega_2+1)}\right) \|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{(1 - R/d)}{(\omega_1+1)(\omega_2+1)} \|\mathbf{A}\|_{\mathrm{F}}^2 \\
&= \left(1 - \frac{R}{d(\omega_1+1)(\omega_2+1)}\right) \|\mathbf{A}\|_{\mathrm{F}}^2,
\end{aligned}$$

where in the last inequality we use the fact that Rank-R is a contraction compressor with parameter R/d (Safaryan et al., 2021).

For $\mathcal{C}_2$, the result follows from Lemma 3.1 (ii).

C.3 Linear Independence of Outer Products

Lemma C.1. *Let vectors $\{v_1, v_2, \ldots, v_r\} \subset \mathbb{R}^d$ are linearly independent. Then outer products $\{v_i v_j^\top : i, j = 1, 2, \ldots, r\}$ are linearly independent matrices in $\mathbb{R}^{d \times d}$.*

Proof. Let $\{e_1, e_2, \ldots, e_d\}$ be the standard basis in $\mathbb{R}^d$. Then, for all $i \in [r]$

$$v_i = \sum_{t=1}^{r} v_{it} e_t.$$

Denote $\mathbf{E}_{tl} = e_t e_l^\top$. Suppose linear combination of matrices $\{v_i v_j^\top : i, j = 1, 2, \ldots, r\}$ with some coefficients c_{ij} is zero matrix. After simple transformations, we get

$$\mathbf{0} = \sum_{i,j=1}^{r} c_{ij} v_i v_j^\top = \sum_{i,j=1}^{r} c_{ij} \sum_{t,l=1}^{d} v_{it} v_{jl} \mathbf{E}_{tl} = \sum_{t,l=1}^{d} \left[\sum_{i,j=1}^{r} c_{ij} v_{it} v_{jl} \right] \mathbf{E}_{tl},$$

which implies that

$$\sum_{i,j=1}^{r} c_{ij} v_{it} v_{jl} = 0, \quad \text{for all } t, l \in [d].$$

Then notice that

$$0 = \sum_{i,j=1}^{r} c_{ij} v_{it} v_{jl} = \sum_{i=1}^{r} \left[\sum_{j=1}^{r} c_{ij} v_{jl} \right] v_{it} = \sum_{i=1}^{r} c'_{il} v_{it}$$

holds for all $t \in [d]$, which implies that $\sum_{i=1}^{r} c'_{il} v_i = 0$ (where that last 0 is a vector of size d). Since v_i's are linearly independent, we get $c'_{il} = 0$ for all $i \in [d]$ and $l \in [r]$. By definition $c'_{il} = \sum_{j=1}^{r} c_{ij} v_{jl}$, hence $\sum_{j=1}^{r} c_{ij} v_j = 0$. Again using linear independence of v_i's, we get $c_{ij} = 0$ for all $i, j \in [d]$. Therefore outer products $v_i v_j^\top$ are also independent. □

Table 2: Key features of different implementation of classical Newton's method in distributed systems. Here m is the number of local training data, r is the intrinsic dimensionality of local data vectors (see Section 2).

Implementation of Newton's method	Standard/Naive	(Islamov et al., 2021)	**Ours**
Problem	General Finite Sum	General Finite Sum	Generalized Linear Model
Data	Arbitrary	Arbitrary	Low Intrinsic Dimensional
Gradient communication cost per iteration (floats)	d	$\min(m, d)$	r
Hessian communication cost per iteration (floats)	d^2	$\min(m, d^2)$	r^2
Initial communication cost (floats)	-	md	rd
Reveals local training data ?	No	Yes	No

D PROOFS OF BL1

We denote $\mathbb{E}_k[\cdot]$ as the conditional expectation on z^k, w^k, and $\mathbf{H}_i^k$.

D.1 Proof of Lemma 4.8

If $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x-y\|$ for any $x, y \in \mathbb{R}^d$, and $i \in [n]$, then from (9) we have

$$\begin{aligned}\|h^i(\nabla^2 f_i(x)) - h^i(\nabla^2 f_i(y))\|_{\mathrm{F}} &= \|vec(h^i(\nabla^2 f_i(x))) - vec(h^i(\nabla^2 f_i(y)))\| \\ &\leq \|\mathcal{B}_i^{-1}\| \cdot \|vec(\nabla^2 f_i(x)) - vec(\nabla^2 f_i(y))\| \\ &= \|\mathcal{B}_i^{-1}\| \cdot \|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \\ &\leq \|\mathcal{B}_i^{-1}\| H_1 \|x-y\|,\end{aligned}$$

which implies that M_1 in Assumption 4.7 satisfies $M_1 \leq \max_i\{\|\mathcal{B}_i^{-1}\|\}H_1$.

If $|(\nabla^2 f_i(x))_{jl} - (\nabla^2 f_i(y))_{jl}| \leq \nu\|x-y\|$ for any $x, y \in \mathbb{R}^d$, $i \in [n]$, and $j, l \in [d]$, then from (9), every entry of $h^i(\nabla^2 f_i(x)) - h^i(\nabla^2 f_i(y))$ will be bounded by $\nu\|\mathcal{B}_i^{-1}\|_\infty\|x-y\|$. Hence M_2 in Assumption 4.7 satisfies $M_2 \leq \nu \max_i\{\|\mathcal{B}_i^{-1}\|_\infty\}$.

D.2 Lemmas

The proofs of Lemma D.1 and Lemma D.2 are the same as that of Lemma B.1 in (Safaryan et al., 2021). Thus we omit them.

Lemma D.1. *Let $\mathcal{Q}$ be a compressor and $\eta > 0$. For any $x, y, z \in \mathbb{R}^d$, we have following results.*
(i) If $\mathcal{Q}$ is an unbiased compressor with parameter ω_{M} and $\eta \leq 1/(\omega_{\mathrm{M}}+1)$, then

$$\mathbb{E}\|z + \eta\mathcal{Q}(x-z) - y\|^2 \leq (1-\eta)\|z-y\|^2 + \eta\|x-y\|^2,$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to $\mathcal{Q}$.
(ii) If $\mathcal{Q}$ is a contraction compressor with parameter δ_{M} and $\eta = 1$, then

$$\mathbb{E}\|z + \eta\mathcal{Q}(x-z) - y\|^2 \leq \left(1 - \frac{\delta_{\mathrm{M}}}{4}\right)\|z-y\|^2 + \left(\frac{6}{\delta_{\mathrm{M}}} - \frac{7}{2}\right)\|x-y\|^2,$$

Lemma D.2. *Let $\mathcal{C}$ be a compressor and $\alpha > 0$. For any matrix $\mathbf{L} \in \mathbb{R}^{d\times d}$ and $y, z \in \mathbb{R}^d$, we have the following results.*
(i) If $\mathcal{C}$ is an unbiased compressor with parameter ω and $\alpha \leq 1/\omega+1$, then

$$\mathbb{E}\|\mathbf{L} + \alpha\mathcal{C}(h^i(\nabla^2 f_i(y)) - \mathbf{L}) - h^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 \leq (1-\alpha)\|\mathbf{L} - h^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 + \alpha M_1^2\|y-z\|^2,$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to $\mathcal{C}$.
(ii) If $\mathcal{C}$ is a contraction compressor with parameter δ and $\alpha = 1$, then

$$\mathbb{E}\|\mathbf{L} + \alpha\mathcal{C}(h^i(\nabla^2 f_i(y)) - \mathbf{L}) - h^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 \leq \left(1 - \frac{\delta}{4}\right)\|\mathbf{L} - h^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 + \left(\frac{6}{\delta} - \frac{7}{2}\right)M_1^2\|y - z\|^2.$$

Lemma D.3. *(i) If Assumption 4.3 (ii) holds, $\|x^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2H^2}, \frac{M}{d}\}$, $\|z^k - x^*\|^2 \leq \min\{\frac{\mu^2}{4dH^2}, M\}$, and $\mathcal{H}^k \leq \frac{\mu^2}{4dN_{\mathrm{B}}R^2}$ for $k \leq K$ and any $M > 0$, then $\|z^{K+1} - x^*\|^2 \leq \min\{\frac{\mu^2}{4dH^2}, M\}$.*
(ii) If Assumption 4.4 holds, $\mathcal{H}^K \leq \frac{A_{\mathrm{M}}\mu^2}{4N_{\mathrm{B}}R^2B_{\mathrm{M}}}$, $\|z^k - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\}$ for $k \leq K$ and any $M > 0$, then $\|z^{K+1} - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\}$.
(iii) If Assumption 4.5(ii) holds, and $\|z^k - x^*\|^2 \leq \frac{M}{d^2M_2^2}$ for $k \leq K$ and any $M > 0$, then $\mathcal{H}^K \leq M$.
(iv) If Assumption 4.6 holds, $\mathcal{H}^K \leq M$, and $\|z^K - x^*\|^2 \leq \frac{AM}{BM_1^2}$ for any any $M > 0$, then $\mathcal{H}^{K+1} \leq M$.

Proof. (i) If $\xi^k = 1$, from (23), (24), and (26), we have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \frac{1}{\mu^2}\left(\frac{H^2}{2}\|z^k - x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^k\right)\|z^k - x^*\|^2 \\ &\leq \frac{1}{d}\|z^k - x^*\|^2 \\ &\leq \min\left\{\frac{\mu^2}{4d^2H^2}, \frac{M}{d}\right\},\end{aligned}$$

for $0 \leq k \leq K$.
If $\xi^k = 0$, since we also have $\|w^k - x^*\|^2 \leq \min\{\frac{\mu^2}{4dH^2}, M\}$, then from (23), (25), and (26), we can get the above inequality in the same way.
Since $\|x^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2H^2}, \frac{M}{d}\}$, we know $\|x^k - x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2H^2}, \frac{M}{d}\}$ for all $0 \leq k \leq K+1$. Then from Assumption 4.3 (ii), we can get

$$\begin{aligned}\|z^{K+1} - x^*\|^2 &\leq d\max_j |z_j^{K+1} - x_j^*|^2 \\ &\leq d\max_{0\leq t\leq K+1}\|x^t - x^*\|^2 \\ &\leq \min\left\{\frac{\mu^2}{4dH^2}, M\right\}.\end{aligned}$$

(ii) First, from the update rule of w^k, we know $\|w^k - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\}$ for $k \leq K$. If $\xi^K = 1$, from (23), (24), and (26), we have

$$\begin{aligned}\|x^{K+1} - x^*\|^2 &\leq \frac{1}{\mu^2}\left(\frac{H^2}{2}\|z^K - x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^K\right)\|z^K - x^*\|^2 \\ &\leq \left(\frac{A_{\mathrm{M}}}{8B_{\mathrm{M}}} + \frac{A_{\mathrm{M}}}{2B_{\mathrm{M}}}\right)\|z^K - x^*\|^2 \\ &\leq \frac{A_{\mathrm{M}}}{B_{\mathrm{M}}}\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\right\}.\end{aligned}$$

If $\xi^K = 0$, from $\|w^K - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\}$ and (25), we can obtain the above inequality similarly. Then from Lemma D.1 (ii), we arrive at

$$\begin{aligned}\|z^{K+1} - z^*\|^2 &\leq (1 - A_{\mathrm{M}})\|z^K - x^*\|^2 + B_{\mathrm{M}}\|x^{K+1} - x^*\|^2 \\ &\leq (1 - A_{\mathrm{M}})\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\right\} + A_{\mathrm{M}}\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\right\} \\ &= \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, M\right\}.\end{aligned}$$

(iii) From Assumption 4.5(ii), we have

$$\begin{aligned}\mathcal{H}^K &= \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \\ &\leq \frac{1}{n}\sum_{i=1}^n d^2 \max_{jl}\{|(\mathbf{L}_i^K)_{jl} - (\mathbf{L}_i^*)_{jl}|^2\} \\ &\leq d^2 M_2^2 \max_{0\leq t\leq K} \|z^t - x^*\|^2 \\ &\leq M.\end{aligned}$$

(iv) From Assumption 4.6 and Lemma D.2 (ii), we have

$$\begin{aligned}\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 &\leq (1-A)\|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 + BM_1^2\|z^K - x^*\|^2 \\ &\leq (1-A)\|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 + AM,\end{aligned}$$

which implies that

$$\mathcal{H}^{K+1} = \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq (1-A)M + AM \leq M.$$

□

D.3 Proof of Theorem 4.9

First we have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|z^k - x^* - [\mathbf{H}^k]_\mu^{-1} g^k\|^2 \\ &= \left\|[\mathbf{H}^k]_\mu^{-1}\left([\mathbf{H}^k]_\mu(z^k - x^*) - (g^k - \nabla f(x^*))\right)\right\|^2 \\ &\leq \frac{1}{\mu^2}\left\|[\mathbf{H}^k]_\mu(z^k - x^*) - (g^k - \nabla f(x^*))\right\|^2, \end{aligned} \tag{23}$$

where we use $\nabla f(x^*) = 0$ in the second equality, and $\|[\mathbf{H}^k]_\mu^{-1}\| \leq \frac{1}{\mu}$ in the last inequality.

If $\xi^k = 1$, then

$$\begin{aligned}&\left\|[\mathbf{H}^k]_\mu(z^k - x^*) - (g^k - \nabla f(x^*))\right\|^2 \\ &= \left\|\nabla f(z^k) - \nabla f(x^*) - \nabla^2 f(x^*)(z^k - x^*) + (\nabla^2 f(x^*) - [\mathbf{H}^k]_\mu)(z^k - x^*)\right\|^2 \\ &\leq 2\left\|\nabla f(z^k) - \nabla f(x^*) - \nabla^2 f(x^*)(z^k - x^*)\right\|^2 + 2\left\|(\nabla^2 f(x^*) - [\mathbf{H}^k]_\mu)(z^k - x^*)\right\|^2 \\ &\leq \frac{H^2}{2}\|z^k - x^*\|^4 + 2\|[\mathbf{H}^k]_\mu - \nabla^2 f(x^*)\|^2 \cdot \|z^k - x^*\|^2 \\ &\leq \frac{H^2}{2}\|z^k - x^*\|^4 + 2\|\mathbf{H}^k - \nabla^2 f(x^*)\|_{\mathrm{F}}^2\|z^k - x^*\|^2 \\ &= \frac{H^2}{2}\|z^k - x^*\|^4 + 2\left\|\frac{1}{n}\mathbf{H}_i^k - \frac{1}{n}\nabla^2 f_i(x^*)\right\|_{\mathrm{F}}^2\|z^k - x^*\|^2 \\ &\leq \frac{H^2}{2}\|z^k - x^*\|^4 + \frac{2}{n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\|z^k - x^*\|^2, \end{aligned} \tag{24}$$

where in the second inequality, we use the Lipschitz continuity of the Hessian of f, and in the last inequality, we use the convexity of $\|\cdot\|_{\mathrm{F}}^2$.

If $\xi^k = 0$, then

$$\begin{aligned}
&\left\|[\mathbf{H}^k]_\mu(z^k - x^*) - (g^k - \nabla f(x^*))\right\|^2 \\
&= \left\|[\mathbf{H}^k]_\mu(z^k - w^k) + \nabla f(w^k) - \nabla f(x^*) - [\mathbf{H}^k]_\mu(z^k - x^*)\right\|^2 \\
&= \left\|[\mathbf{H}^k]_\mu(x^* - w^k) + \nabla f(w^k) - \nabla f(x^*)\right\|^2 \\
&= \left\|\nabla f(w^k) - \nabla f(x^*) - \nabla^2 f(x^*)(w^k - x^*) + (\nabla^2 f(x^*) - [\mathbf{H}^k]_\mu)(w^k - x^*)\right\|^2 \\
&\leq \frac{H^2}{2}\|w^k - x^*\|^4 + 2\|\mathbf{H}^k - \nabla^2 f(x^*)\|_{\mathrm{F}}^2\|w^k - x^*\|^2 \\
&\leq \frac{H^2}{2}\|w^k - x^*\|^4 + \frac{2}{n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\|w^k - x^*\|^2. \qquad (25)
\end{aligned}$$

From the above three inequalities, we can obtain

$$\begin{aligned}
\mathbb{E}_k\|x^{k+1} - x^*\|^2 \leq{}& \frac{H^2 p}{2\mu^2}\|z^k - x^*\|^4 + \frac{2p}{n\mu^2}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\|z^k - x^*\|^2 \\
&+ \frac{H^2(1-p)}{2\mu^2}\|w^k - x^*\|^4 + \frac{2(1-p)}{n\mu^2}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\|w^k - x^*\|^2.
\end{aligned}$$

From the definition of $(A_{\mathrm{M}}, B_{\mathrm{M}})$ and Lemma D.1, by choosing $z = z^k$, $x = x^{k+1}$, and $y = x^*$ in Lemma D.1, we can obtain

$$\begin{aligned}
\mathbb{E}_k\|z^{k+1} - x^*\|^2 &= \mathbb{E}_k\|z^k + \eta\mathcal{Q}^k(x^{k+1} - z^k) - x^*\|^2 \\
&\leq (1 - A_{\mathrm{M}})\|z^k - x^*\|^2 + B_{\mathrm{M}}\mathbb{E}_k\|x^{k+1} - x^*\|^2.
\end{aligned}$$

Combining the above two inequalities, we arrive at

$$\begin{aligned}
\mathbb{E}_k\|z^{k+1} - x^*\|^2 \leq{}& (1 - A_{\mathrm{M}})\|z^k - x^*\|^2 + \frac{B_{\mathrm{M}}p}{\mu^2}\left(\frac{H^2}{2}\|z^k - x^*\|^2 + \frac{2}{n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\right)\|z^k - x^*\|^2 \\
&+ \frac{B_{\mathrm{M}}(1-p)}{\mu^2}\left(\frac{H^2}{2}\|w^k - x^*\|^2 + \frac{2}{n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\right)\|w^k - x^*\|^2.
\end{aligned}$$

From the update rule of $\mathbf{H}_i^k$, we know $\mathbf{H}_i^k = \sum_{jl}(\mathbf{L}_i^k)_{jl}\mathbf{B}_i^{jl}$. Denote $\mathbf{L}_i^* = h^i(\nabla^2 f_i(x^*))$. Then we have

$$\begin{aligned}
\|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2 &= \left\|\sum_{jl}(\mathbf{L}_i^k - \mathbf{L}_i^*)_{jl}\mathbf{B}_i^{jl}\right\|_{\mathrm{F}}^2 \\
&\leq N_{\mathrm{B}}\sum_{jl}\|(\mathbf{L}_i^k - \mathbf{L}_i^*)_{jl}\mathbf{B}_i^{jl}\|_{\mathrm{F}}^2 \\
&\leq N_{\mathrm{B}}R^2\|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2. \qquad (26)
\end{aligned}$$

Define $\mathcal{H}^k := \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2$. Then we have

$$\begin{aligned}
\mathbb{E}_k\|z^{k+1} - x^*\|^2 \leq{}& (1 - A_{\mathrm{M}})\|z^k - x^*\|^2 + \frac{B_{\mathrm{M}}p}{\mu^2}\left(\frac{H^2}{2}\|z^k - x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^k\right)\|z^k - x^*\|^2 \\
&+ \frac{B_{\mathrm{M}}(1-p)}{\mu^2}\left(\frac{H^2}{2}\|w^k - x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^k\right)\|w^k - x^*\|^2.
\end{aligned}$$

Assume $\|z^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for $k \geq 0$. Then from the update rule of w^k, we know $\|w^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}$ for $k \geq 0$. Thus we have

$$\mathbb{E}_k\|z^{k+1} - x^*\|^2 \leq \left(1 - A_{\mathrm{M}} + \frac{A_{\mathrm{M}}p}{4}\right)\|z^k - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{4}\|w^k - x^*\|^2. \tag{27}$$

From the update rule of w^k, we have

$$\mathbb{E}_k\|w^{k+1} - x^*\|^2 = p\|z^k - x^*\|^2 + (1-p)\|w^k - x^*\|^2. \tag{28}$$

Define $\Phi_1^k := \|z^k - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{2p}\|w^k - x^*\|^2$. Then we can get

$$\begin{aligned}
\mathbb{E}_k[\Phi_1^{k+1}] &= \mathbb{E}_k\|z^{k+1} - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{2p}\mathbb{E}_k\|w^{k+1} - x^*\|^2 \\
&\overset{(27)}{\leq} \left(1 - A_{\mathrm{M}} + \frac{A_{\mathrm{M}}p}{4}\right)\|z^k - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{4}\|w^k - x^*\|^2 + \frac{A_{\mathrm{M}}(1-p)}{2p}\mathbb{E}_k\|w^{k+1} - x^*\|^2 \\
&\overset{(28)}{\leq} \left(1 - \frac{A_{\mathrm{M}}}{2}\right)\|z^k - x^*\|^2 + \left(1 - \frac{p}{2}\right)\frac{A_{\mathrm{M}}(1-p)}{2p}\|w^k - x^*\|^2 \\
&\leq \left(1 - \frac{\min\{A_{\mathrm{M}}, p\}}{2}\right)\Phi_1^k.
\end{aligned}$$

By applying the tower property, we have

$$\mathbb{E}[\Phi_1^{k+1}] \leq \left(1 - \frac{\min\{A_{\mathrm{M}}, p\}}{2}\right)\mathbb{E}[\Phi_1^k].$$

Unrolling the recursion, we can get the result.

D.4 Proof of Theorem 4.10

Since $\xi^k \equiv 1$, $\eta = 1$, and $\mathcal{Q}^k(x) \equiv x$ for any $x \in \mathbb{R}^d$, it is easy to see that $z^k \equiv x^k$ for $k \geq 0$. In this case, we can view $\mathcal{Q}^k$ as an unbiased compressor with $\omega_{\mathrm{M}} = 0$ or a contraction compressor with $\delta_{\mathrm{M}} = 1$. Then from (27), we have

$$\mathbb{E}_k\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{A_{\mathrm{M}}}{2}\right)\|x^k - x^*\|^2.$$

From Lemma D.2, we can obtain

$$\mathbb{E}_k[\mathcal{H}^{k+1}] \leq (1-A)\mathcal{H}^k + BM_1^2\|x^k - x^*\|^2.$$

Thus,

$$\begin{aligned}
\mathbb{E}_k[\Phi_2^{k+1}] &= \mathbb{E}_k[\mathcal{H}^{k+1}] + \frac{4BM_1^2}{A_{\mathrm{M}}}\mathbb{E}_k\|x^{k+1} - x^*\|^2 \\
&\leq (1-A)\mathcal{H}^k + BM_1^2\|x^k - x^*\|^2 + \frac{4BM_1^2}{A_{\mathrm{M}}}\left(1 - \frac{A_{\mathrm{M}}}{2}\right)\|x^k - x^*\|^2 \\
&\leq \left(1 - \frac{\min\{4A, A_{\mathrm{M}}\}}{4}\right)\Phi_2^k.
\end{aligned}$$

By applying the tower property, we have $\mathbb{E}[\Phi_2^{k+1}] \leq \theta_1\mathbb{E}[\Phi_2^k]$. Unrolling the recursion, we have $\mathbb{E}[\Phi_2^k] \leq \theta_1^k\Phi_2^0$. Then we further have $\mathbb{E}[\mathcal{H}^k] \leq \theta_1^k\Phi_2^0$ and $\mathbb{E}\|x^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}}{4BM_1^2}\theta_1^k\Phi_2^0$. From $z^k \equiv x^k$, (23), and (24), we can get

$$\|x^{k+1} - x^*\|^2 \leq \frac{1}{\mu^2}\left(\frac{H^2}{2}\|x^k - x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^k\right)\|x^k - x^*\|^2.$$

Assume $x^k \neq x^*$ for all $k \geq 0$. Then we have

$$\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2} \leq \frac{1}{\mu^2}\left(\frac{H^2}{2}\|x^k-x^*\|^2 + 2N_{\mathrm{B}}R^2\mathcal{H}^k\right),$$

and by taking expectation, we arrive at

$$\begin{aligned}\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] &\leq \frac{H^2}{2\mu^2}\mathbb{E}\|x^k-x^*\|^2 + \frac{2N_{\mathrm{B}}R^2}{\mu^2}\mathbb{E}[\mathcal{H}^k] \\ &\leq \theta_1^k\left(\frac{A_{\mathrm{M}}H^2}{8BM_1^2\mu^2} + \frac{2N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_2^0.\end{aligned}$$

D.5 Proof of Theorem 4.11

(i) Noticed that under Assumption 4.3, we have $A_{\mathrm{M}} = B_{\mathrm{M}} = \eta$. We prove this by mathematical induction. First, since $z^0 = x^0$, we know $\|z^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{4dH^2}, \frac{\mu^2}{16d^3N_{\mathrm{B}}R^2M_2^2}\}$. Then from Lemma D.3 (iii), we have $\mathcal{H}^0 \leq \frac{\mu^2}{16dN_{\mathrm{B}}R^2}$. Next, assume

$$\|z^k - x^*\|^2 \leq \min\left\{\frac{\mu^2}{4dH^2}, \frac{\mu^2}{16d^3N_{\mathrm{B}}R^2M_2^2}\right\} \quad \text{and} \quad \mathcal{H}^k \leq \frac{\mu^2}{16dN_{\mathrm{B}}R^2},$$

for $k \leq K$. By choosing $M = \frac{\mu^2}{16d^3N_{\mathrm{B}}R^2M_2^2}$ in Lemma D.3 (i), we have $\|z^{K+1} - x^*\|^2 \leq \min\left\{\frac{\mu^2}{4dH^2}, \frac{\mu^2}{16d^3N_{\mathrm{B}}R^2M_2^2}\right\}$. By further using Lemma D.3 (iii), we can get $\mathcal{H}^{K+1} \leq \frac{\mu^2}{16dN_{\mathrm{B}}R^2}$.

(ii) We prove the result by induction. Assume $\|z^k - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, \frac{AA_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}BM_1^2}\right\}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for $k \leq K$. Then by Lemma D.3 (iv), we have $\mathcal{H}^{K+1} \leq \frac{A_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}}$. Moreover, by Lemma D.3 (ii), we have $\|z^{K+1} - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}, \frac{AA_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}BM_1^2}\right\}$.

E PROOFS OF BL2

We denote $\mathbb{E}_k[\cdot]$ as the conditonal expectation on z_i^k, w_i^k, l_i^k, $\mathbf{L}_i^k$, and $\mathbf{H}_i^k$.

E.1 A Lemma

Lemma E.1. *(i) If Assumption 4.3 (ii) holds,* $\|x^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{d^2(6H^2+24H_1^2)}, \frac{M}{d}\}$, $\|z_i^k - x^*\|^2 \leq \min\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\}$, *and* $\mathcal{H}^k \leq \frac{\mu^2}{24dN_{\mathrm{B}}R^2}$ *for* $k \leq K$, $i \in [n]$, *and any* $M > 0$, *then* $\|z_i^{K+1} - x^*\|^2 \leq \min\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\}$ *for* $i \in [n]$.
(ii) If Assumption 4.4 holds, $\mathcal{H}^K \leq \frac{A_{\mathrm{M}}\mu^2}{24N_{\mathrm{B}}R^2B_{\mathrm{M}}}$, $\|z_i^k - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, M\}$ *for* $k \leq K$, $i \in [n]$, *and any* $M > 0$, *then* $\|z_i^{K+1} - x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, M\}$ *for* $i \in [n]$.
(iii) If Assumption 4.5(ii) holds, and $\|z_i^k - x^*\|^2 \leq \frac{M}{d^2M_2^2}$ *for* $k \leq K$, $i \in [n]$, *and any* $M > 0$, *then* $\mathcal{H}^K \leq M$.
(iv) If Assumption 4.6 holds, $\|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq M$, *and* $\|z_i^K - x^*\|^2 \leq \frac{AM}{BM_1^2}$ *for* $i \in [n]$ *and any* $M > 0$, *then* $\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq M$ *for* $i \in [n]$.

Proof. (i) First, from the update rule of w_i^k, we know $\mathcal{Z}^k \leq \min\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\}$ and $\mathcal{W}^k \leq \min\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\}$ for $k \leq K$. Then from (30), we have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \frac{3H^2}{4\mu^2}(\mathcal{W}^k)^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathcal{H}^k\mathcal{W}^k + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k\mathcal{W}^k \\ &\leq \frac{1}{d}\mathcal{W}^k \\ &\leq \min\left\{\frac{\mu^2}{d^2(6H^2+24H_1^2)}, \frac{M}{d}\right\},\end{aligned}$$

for $0 \leq k \leq K$.
Since $\|x^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{d^2(6H^2+24H_1^2)}, \frac{M}{d}\}$, we know $\|x^k - x^*\|^2 \leq \min\{\frac{\mu^2}{d^2(6H^2+24H_1^2)}, \frac{M}{d}\}$ for all $0 \leq k \leq K+1$. Then for $i \in S^k$, from Assumption 4.3 (ii), we can get

$$\begin{aligned}\|z_i^{K+1} - x^*\|^2 &\leq d\max_j |(z_i^{K+1})_j - x_j^*|^2 \\ &\leq d\max_{0\leq t\leq K+1}\|x^t - x^*\|^2 \\ &\leq \min\left\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\right\}.\end{aligned}$$

For $i \notin S^k$, we have

$$\|z_i^{K+1} - x^*\|^2 = \|z_i^K - x^*\|^2 \leq \min\left\{\frac{\mu^2}{d(6H^2+24H_1^2)}, M\right\}.$$

(ii) First, from the update rule of w^k, we know $\mathcal{Z}^k \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, M\}$ and $\mathcal{W}^k \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, M\}$ for $k \leq K$. Then from (30), we have

$$\begin{aligned}\|x^{K+1} - x^*\|^2 &\leq \frac{3H^2}{4\mu^2}(\mathcal{W}^K)^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathcal{H}^K\mathcal{W}^K + \frac{3H_1^2}{\mu^2}\mathcal{Z}^K\mathcal{W}^K \\ &\leq \frac{A_{\mathrm{M}}}{B_{\mathrm{M}}}\mathcal{W}^K \\ &\leq \frac{A_{\mathrm{M}}}{B_{\mathrm{M}}}\min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2+24H_1^2)}, M\right\}.\end{aligned}$$

Then for $i \in S^k$, from Lemma D.1 (ii), we arrive at

$$\begin{aligned}\|z_i^{K+1} - z^*\|^2 &\leq (1 - A_{\mathrm{M}})\|z_i^K - x^*\|^2 + B_{\mathrm{M}}\|x^{K+1} - x^*\|^2 \\ &\leq (1 - A_{\mathrm{M}}) \min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2 + 24H_1^2)}, M\right\} + A_{\mathrm{M}} \min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2 + 24H_1^2)}, M\right\} \\ &= \min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2 + 24H_1^2)}, M\right\}.\end{aligned}$$

For $i \notin S^k$, we have

$$\|z_i^{K+1} - x^*\|^2 = \|z_i^K - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{B_{\mathrm{M}}(6H^2 + 24H_1^2)}, M\right\}.$$

(iii) From Assumption 4.5(ii), we have

$$\begin{aligned}\mathcal{H}^K &= \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \\ &\leq \frac{1}{n}\sum_{i=1}^n d^2 \max_{jl}\{|(\mathbf{L}_i^K)_{jl} - (\mathbf{L}_i^*)_{jl}|^2\} \\ &\leq d^2 M_2^2 \max_{i\in[n], 0\leq t\leq K} \|z_i^t - x^*\|^2 \\ &\leq M.\end{aligned}$$

(iv) For $i \in S^k$, from Assumption 4.6 and Lemma D.2 (ii), we have

$$\begin{aligned}\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 &\leq (1 - A)\|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 + BM_1^2\|z_i^K - x^*\|^2 \\ &\leq (1 - A)M + AM \\ &= M.\end{aligned}$$

For $i \notin S^k$, we also have

$$\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 = \|\mathbf{L}_i^K - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq M.$$

□

E.2 Proof of Theorem 4.12

First, from $[\mathbf{H}_i^k]_s + l_i^k\mathbf{I} \succeq \nabla^2 f_i(z_i^k) \succeq \mu\mathbf{I}$, we know $[\mathbf{H}^k]_s + l^k\mathbf{I} = \frac{1}{n}\sum_{i=1}^n([\mathbf{H}_i^k]_s + l_i^k\mathbf{I}) \succeq \mu\mathbf{I}$. Then we have

$$\begin{aligned}\|x^{k+1} - x^*\| &= \left\|\left([\mathbf{H}^k]_s + l^k\mathbf{I}\right)^{-1}\left(g^k - \left([\mathbf{H}^k]_s + l^k\mathbf{I}\right)x^* + \nabla f(x^*)\right)\right\| \\ &\leq \frac{1}{\mu}\left\|\frac{1}{n}\sum_{i=1}^n g_i^k - \frac{1}{n}\sum_{i=1}^n([\mathbf{H}_i^k]_s + l_i^k\mathbf{I})x^* + \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)\right\| \\ &\leq \frac{1}{n\mu}\sum_{i=1}^n \|g_i^k - ([\mathbf{H}_i^k]_s + l_i^k)x^* + \nabla f_i(x^*)\| \\ &= \frac{1}{n\mu}\sum_{i=1}^n \|([\mathbf{H}_i^k]_s + l_i^k\mathbf{I})w_i^k - \nabla f_i(w_i^k) - ([\mathbf{H}_i^k]_s + l_i^k\mathbf{I})x^* + \nabla f_i(x^*)\| \\ &\leq \frac{1}{n\mu}\sum_{i=1}^n \|\nabla f_i(w_i^k) - \nabla f_i(x^*) - \nabla^2 f_i(x^*)(w_i^k - x^*)\| \\ &\quad + \frac{1}{n\mu}\sum_{i=1}^n \|([\mathbf{H}_i^k]_s + l_i^k\mathbf{I} - \nabla^2 f_i(x^*))(w_i^k - x^*)\|,\end{aligned}$$

where we use $\nabla f(x^*) = 0$ in the first equality and $g_i^k = ([\mathbf{H}_i^k]_s + l_i^k \mathbf{I})w_i^k - \nabla f_i(w_i^k)$ in the second equality. Since $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq H\|x - y\|$ for any $x, y \in \mathbb{R}^d$, we further have

$$\begin{aligned}\|x^{k+1} - x^*\| &\leq \frac{H}{2n\mu}\sum_{i=1}^n \|w_i^k - x^*\|^2 + \frac{1}{n\mu}\sum_{i=1}^n \|([\mathbf{H}_i^k]_s + l_i^k\mathbf{I} - \nabla^2 f_i(x^*))\| \cdot (w_i^k - x^*)\| \\ &\leq \frac{H}{2\mu}\mathcal{W}^k + \frac{1}{n\mu}\sum_{i=1}^n \left(\|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\|_{\mathrm{F}} + l_i^k\right)\|w_i^k - x^*\|,\end{aligned}$$

where $\mathcal{W}^k := \frac{1}{n}\sum_{i=1}^n \|w_i^k - x^*\|^2$.

Since $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x - y\|$ for any $x, y \in \mathbb{R}^d$, we can get

$$\begin{aligned}l_i^k &= \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(z_i^k)\|_{\mathrm{F}} \\ &\leq \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\|_{\mathrm{F}} + \|\nabla^2 f_i(z_i^k) - \nabla^2 f_i(x^*)\|_{\mathrm{F}} \\ &\leq \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\|_{\mathrm{F}} + H_1\|z_i^k - x^*\|.\end{aligned}$$

Thus,

$$\begin{aligned}\|x^{k+1} - x^*\| &\leq \frac{H}{2\mu}\mathcal{W}^k + \frac{1}{n\mu}\sum_{i=1}^n \left(\|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\| + \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\|_{\mathrm{F}} + H_1\|z_i^k - x^*\|\right)\|w_i^k - x^*\| \\ &\leq \frac{H}{2\mu}\mathcal{W}^k + \frac{2}{n\mu}\sum_{i=1}^n \|[\mathbf{H}_i^k]_s - \nabla^2 f_i(x^*)\|_{\mathrm{F}}\|w_i^k - x^*\| + \frac{H_1}{n\mu}\sum_{i=1}^n \|z_i^k - x^*\|\|w_i^k - x^*\| \\ &\overset{\text{Lemma 3.1}}{\leq} \frac{H}{2\mu}\mathcal{W}^k + \frac{2}{n\mu}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}\|w_i^k - x^*\| + \frac{H_1}{n\mu}\sum_{i=1}^n \|z_i^k - x^*\|\|w_i^k - x^*\| \\ &\leq \frac{H}{2\mu}\mathcal{W}^k + \frac{2}{n\mu}\left(\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2\right)^{\frac{1}{2}}\left(n\mathcal{W}^k\right)^{\frac{1}{2}} + \frac{H_1}{n\mu}\left(n\mathcal{Z}^k\right)^{\frac{1}{2}}\left(n\mathcal{W}^k\right)^{\frac{1}{2}},\end{aligned}$$

where we use the Cauchy-Schwarz inequality in the last inequality and $\mathcal{Z}^k := \frac{1}{n}\sum_{i=1}^n \|z_i^k - x^*\|^2$. Since $\mathbf{H}_i^k = \sum_{jl}(\mathbf{L}_i^k)_{jl}\mathbf{B}_i^{jl}$, same as (26), we have

$$\|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2 \leq N_{\mathrm{B}} R^2 \|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2, \tag{29}$$

where $\mathbf{L}_i^* = h^i(\nabla^2 f_i(x^*))$ and N_{B} is defined in 10. Then from the convexity of $\|\cdot\|^2$, we further bound $\|x^{k+1} - x^*\|^2$ as

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \frac{3H^2}{4\mu^2}(\mathcal{W}^k)^2 + \frac{12\mathcal{W}^k}{n\mu^2}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|_{\mathrm{F}}^2 + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k\mathcal{W}^k \\ &\overset{(29)}{\leq} \frac{3H^2}{4\mu^2}(\mathcal{W}^k)^2 + \frac{12N_{\mathrm{B}}R^2\mathcal{W}^k}{n\mu^2}\sum_{i=1}^n \|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2 + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k\mathcal{W}^k \\ &= \frac{3H^2}{4\mu^2}(\mathcal{W}^k)^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathcal{H}^k\mathcal{W}^k + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k\mathcal{W}^k,\end{aligned} \tag{30}$$

where $\mathcal{H}^k = \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2$.

For $i \in S^k$, we have $z_i^{k+1} = z_i^k + \eta\mathcal{Q}_i^k(x^{k+1} - z_i^k)$. Then from the definition of $(A_{\mathrm{M}}, B_{\mathrm{M}})$ and Lemma D.1, by choosing $z = z_i^k$, $x = x^{k+1}$, and $y = x^*$ in Lemma D.1, we can obtain

$$\begin{aligned}\mathbb{E}_k[\|z_i^{k+1} - x^*\|^2 \mid i \in S^k] &= \mathbb{E}_k[\|z_i^k + \eta\mathcal{Q}^k(x^{k+1} - z_i^k) - x^*\|^2 \mid i \in S^k] \\ &\leq (1 - A_{\mathrm{M}})\|z_i^k - x^*\|^2 + B_{\mathrm{M}}\mathbb{E}_k\|x^{k+1} - x^*\|^2 \\ &= (1 - A_{\mathrm{M}})\|z_i^k - x^*\|^2 + B_{\mathrm{M}}\|x^{k+1} - x^*\|^2.\end{aligned}$$

Noticing that $\mathbb{P}[i \in S^k] = \tau/n$ and $z_i^{k+1} = z_i^k$ for $i \notin S^k$, we further have

$$\begin{aligned}\mathbb{E}_k\|z_i^{k+1} - x^*\|^2 &= \frac{\tau}{n}\mathbb{E}_k[\|z_i^{k+1} - x^*\|^2 \mid i \in S^k] + \left(1 - \frac{\tau}{n}\right)\mathbb{E}_k[\|z_i^{k+1} - x^*\|^2 \mid i \notin S^k] \\ &\leq \frac{\tau}{n}(1 - A_{\mathrm{M}})\|z_i^k - x^*\|^2 + \frac{\tau B_{\mathrm{M}}}{n}\|x^{k+1} - x^*\|^2 + \left(1 - \frac{\tau}{n}\right)\|z_i^k - x^*\|^2 \\ &= \left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\|z_i^k - x^*\|^2 + \frac{\tau B_{\mathrm{M}}}{n}\|x^{k+1} - x^*\|^2,\end{aligned}$$

which implies that

$$\begin{aligned}\mathbb{E}_k[\mathcal{Z}^{k+1}] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\|z_i^{k+1} - x^*\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\|z_i^k - x^*\|^2 + \frac{\tau B_{\mathrm{M}}}{n}\|x^{k+1} - x^*\|^2 \\ &= \left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k + \frac{\tau B_{\mathrm{M}}}{n}\|x^{k+1} - x^*\|^2. \end{aligned} \tag{31}$$

For $i \in S^k$, from the update rule of w_i^{k+1}, we have

$$\mathbb{E}_k[\|w_i^{k+1} - x^*\|^2 \mid i \in S^k] = p\mathbb{E}_k[\|z_i^{k+1} - x^*\|^2] + (1-p)\|w_i^k - x^*\|^2.$$

For $i \notin S^k$, we have $w_i^{k+1} = w_i^k$. Thus,

$$\begin{aligned}\mathbb{E}_k\|w_i^{k+1} - x^*\|^2 &= \frac{\tau}{n}\mathbb{E}_k[\|w_i^{k+1} - x^*\|^2 \mid i \in S^k] + \left(1 - \frac{\tau}{n}\right)\mathbb{E}_k[\|w_i^{k+1} - x^*\|^2 \mid i \notin S^k] \\ &= \left(1 - \frac{\tau p}{n}\right)\|w_i^k - x^*\|^2 + \frac{\tau p}{n}\mathbb{E}_k\|z_i^{k+1} - x^*\|^2,\end{aligned}$$

which yields that

$$\begin{aligned}\mathbb{E}_k[\mathcal{W}^{k+1}] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\|w_i^{k+1} - x^*\|^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\tau p}{n}\right)\|w_i^k - x^*\|^2 + \frac{1}{n}\sum_{i=1}^n \frac{\tau p}{n}\mathbb{E}_k\|z_i^{k+1} - x^*\|^2 \\ &= \left(1 - \frac{\tau p}{n}\right)\mathcal{W}^k + \frac{\tau p}{n}\mathbb{E}_k[\mathcal{Z}^{k+1}] \\ &\overset{(31)}{\leq} \left(1 - \frac{\tau p}{n}\right)\mathcal{W}^k + \frac{\tau p}{n}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k + \frac{\tau^2 B_{\mathrm{M}} p}{n^2}\|x^{k+1} - x^*\|^2. \end{aligned} \tag{32}$$

Let $\Phi_3^k := \mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k$ for $k \geq 0$. Then from the above inequality we have

$$\begin{aligned}\mathbb{E}_k[\Phi_3^{k+1}] &\leq \left(1 - \frac{\tau p}{n}\right)\mathcal{W}^k + \frac{\tau p}{n}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k + \frac{\tau^2 B_{\mathrm{M}} p}{n^2}\|x^{k+1} - x^*\|^2 + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\mathbb{E}_k[\mathcal{Z}^{k+1}] \\ &\overset{(31)}{\leq} \left(1 - \frac{\tau p}{n}\right)\mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\left(1 - \frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k + \frac{2\tau p B_{\mathrm{M}}}{n A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{2n}\right)\|x^{k+1} - x^*\|^2 \\ &\overset{30}{\leq} \left(1 - \frac{\tau p}{n} + \frac{2\tau p B_{\mathrm{M}}}{n A_{\mathrm{M}}}\left(\frac{3H^2}{4\mu^2}\mathcal{W}^k + \frac{12 N_{\mathrm{B}} R^2}{\mu^2}\mathcal{H}^k + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k\right)\right)\mathcal{W}^k \\ &\quad + \frac{2p}{A_{\mathrm{M}}}\left(1 - \frac{\tau A_{\mathrm{M}}}{n}\right)\left(1 - \frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k.\end{aligned}$$

If $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{(6H^2 + 24H_1^2)B_{\mathrm{M}}}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{96 N_{\mathrm{B}} R^2 B_{\mathrm{M}}}$ for all $k \geq 0$, then we have

$$\frac{3H^2}{4\mu^2}\mathcal{W}^k + \frac{12 N_{\mathrm{B}} R^2}{\mu^2}\mathcal{H}^k + \frac{3H_1^2}{\mu^2}\mathcal{Z}^k \leq \frac{A_{\mathrm{M}}}{4B_{\mathrm{M}}},$$

which implies that

$$\begin{aligned}\mathbb{E}_k[\Phi_3^{k+1}] &\leq \left(1-\frac{\tau p}{2n}\right)\mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\left(1-\frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k \\ &\leq \left(1-\frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)\Phi_3^k.\end{aligned}$$

By applying the tower property, we have

$$\mathbb{E}[\Phi_3^{k+1}] \leq \left(1-\frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)\mathbb{E}[\Phi_3^k].$$

Unrolling the recursion, we can obtain the result.

E.3 Proof of Theorem 4.13

Since $\xi^k \equiv 1$, $\eta = 1$, $S^k \equiv [n]$, and $\mathcal{Q}_i^k(x) \equiv x$ for any $x \in \mathbb{R}^d$, it is easy to see that $w_i^k = z_i^k \equiv x^k$ for all $i \in [n]$ and $k \geq 0$. In this case, we can view $\mathcal{Q}_i^k$ as an unbiased compressor with $\omega_{\mathrm{M}} = 0$ or a contraction compressor with $\delta_{\mathrm{M}} = 1$. Then from (31), we have

$$\begin{aligned}\mathbb{E}_k\|x^{k+1}-x^*\|^2 &\leq (1-A_{\mathrm{M}})\|x^k-x^*\|^2 + B_{\mathrm{M}}\|x^{k+1}-x^*\|^2 \\ &\overset{(30)}{\leq} (1-A_{\mathrm{M}})\|x^k-x^*\|^2 + \frac{1}{4}A_{\mathrm{M}}\|x^k-x^*\|^2 \\ &= \left(1-\frac{3A_{\mathrm{M}}}{4}\right)\|x^k-x^*\|^2.\end{aligned}$$

From Lemma D.2, we can obtain

$$\mathbb{E}_k[\mathcal{H}^{k+1}] \leq (1-A)\mathcal{H}^k + BM_1^2\|x^k-x^*\|^2.$$

Thus,

$$\begin{aligned}\mathbb{E}_k[\Phi_4^{k+1}] &= \mathbb{E}_k[\mathcal{H}^{k+1}] + \frac{4BM_1^2}{A_{\mathrm{M}}}\mathbb{E}_k\|x^{k+1}-x^*\|^2 \\ &\leq (1-A)\mathcal{H}^k + BM_1^2\|x^k-x^*\|^2 + \frac{4BM_1^2}{A_{\mathrm{M}}}\left(1-\frac{3A_{\mathrm{M}}}{4}\right)\|x^k-x^*\|^2 \\ &\leq \left(1-\frac{\min\{2A, A_{\mathrm{M}}\}}{2}\right)\Phi_4^k.\end{aligned}$$

By applying the tower property, we have $\mathbb{E}[\Phi_4^{k+1}] \leq \theta_2\mathbb{E}[\Phi_4^k]$. Unrolling the recursion, we have $\mathbb{E}[\Phi_4^k] \leq \theta_2^k\Phi_4^0$. Then we further have $\mathbb{E}[\mathcal{H}^k] \leq \theta_2^k\Phi_4^0$ and $\mathbb{E}\|x^k-x^*\|^2 \leq \frac{A_{\mathrm{M}}}{4BM_1^2}\theta_2^k\Phi_4^0$. From $w_i^k = z_i^k \equiv x^k$ and (30), we can get

$$\|x^{k+1}-x^*\|^2 \leq \left(\frac{3H^2+12H_1^2}{4\mu^2}\|x^k-x^*\|^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathcal{H}^k\right)\|x^k-x^*\|^2.$$

Assume $x^k \neq x^*$ for all $k \geq 0$. Then we have

$$\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2} \leq \frac{3H^2+12H_1^2}{4\mu^2}\|x^k-x^*\|^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathcal{H}^k,$$

and by taking expectation, we arrive at

$$\begin{aligned}\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] &\leq \frac{3H^2+12H_1^2}{4\mu^2}\mathbb{E}\|x^k-x^*\|^2 + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\mathbb{E}[\mathcal{H}^k] \\ &\leq \theta_2^k\left(\frac{A_{\mathrm{M}}(3H^2+12H_1^2)}{16BM_1^2\mu^2} + \frac{12N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_4^0.\end{aligned}$$

E.4 Proof of Theorem 4.14

(i) Noticed that under Assumption 4.3, we have $A_{\rm M} = B_{\rm M} = \eta$. We prove this by mathematical induction. First, since $z_i^0 = x^0$, we know $\|z_i^0 - x^*\|^2 \leq \min\{\frac{\mu^2}{d(6H^2+24H_1^2)}, \frac{\mu^2}{96d^3N_{\rm B}R^2M_2^2}\}$ for $i \in [n]$. Then from Lemma E.1 (iii), we have $\mathcal{H}^0 \leq \frac{\mu^2}{96dN_{\rm B}R^2}$. Next, assume

$$\|z_i^k - x^*\|^2 \leq \min\left\{\frac{\mu^2}{d(6H^2+24H_1^2)}, \frac{\mu^2}{96d^3N_{\rm B}R^2M_2^2}\right\} \text{ for } i \in [n] \quad \text{and} \quad \mathcal{H}^k \leq \frac{\mu^2}{96dN_{\rm B}R^2},$$

for $k \leq K$. By choosing $M = \frac{\mu^2}{96d^3N_{\rm B}R^2M_2^2}$ in Lemma E.1 (i), we have

$$\|z_i^{K+1} - x^*\|^2 \leq \min\left\{\frac{\mu^2}{d(6H^2+24H_1^2)}, \frac{\mu^2}{96d^3N_{\rm B}R^2M_2^2}\right\},$$

for $i \in [n]$. By further using Lemma E.1 (iii), we can get $\mathcal{H}^{K+1} \leq \frac{\mu^2}{96dN_{\rm B}R^2}$.

(ii) We prove the result by induction. Assume $\|z_i^k - x^*\|^2 \leq \min\left\{\frac{A_{\rm M}\mu^2}{B_{\rm M}(6H^2+24H_1^2)}, \frac{AA_{\rm M}\mu^2}{96N_{\rm B}R^2B_{\rm M}BM_1^2}\right\}$ and $\|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\rm F}^2 \leq \frac{A_{\rm M}\mu^2}{96N_{\rm B}R^2B_{\rm M}}$ for all $i \in [n]$ and $k \leq K$. Then by Lemma E.1 (iv), we have $\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\rm F}^2 \leq \frac{A_{\rm M}\mu^2}{96N_{\rm B}R^2B_{\rm M}}$. Moreover, by Lemma E.1 (ii), we have $\|z_i^{K+1} - x^*\|^2 \leq \min\left\{\frac{A_{\rm M}\mu^2}{B_{\rm M}(6H^2+24H_1^2)}, \frac{AA_{\rm M}\mu^2}{96N_{\rm B}R^2B_{\rm M}BM_1^2}\right\}$ for $i \in [n]$.

F PROOFS OF BL3

We denote $\mathbb{E}_k[\cdot]$ as the conditonal expectation on z_i^k, w_i^k, $\mathbf{L}_i^k$, γ_i^k, β_i^k, $\mathbf{A}_i^k$ and $\mathbf{C}_i^k$.

F.1 Proof of Lemma A.2

(i) If $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \leq H_1\|x-y\|$ for any $x, y \in \mathbb{R}^d$, and $i \in [n]$, then from (15) we have

$$\begin{aligned}\|\tilde{h}^i(\nabla^2 f_i(x)) - \tilde{h}^i(\nabla^2 f_i(y))\|_{\mathrm{F}} &\leq \|svec(\tilde{h}^i(\nabla^2 f_i(x))) - svec(\tilde{h}^i(\nabla^2 f_i(y)))\| \\ &\leq \|(\tilde{\mathcal{B}}_i)^{-1}\| \cdot \|svec(\nabla^2 f_i(x)) - svec(\nabla^2 f_i(y))\| \\ &\leq \sqrt{2}\|(\tilde{\mathcal{B}}_i)^{-1}\| \cdot \|\nabla^2 f_i(x) - \nabla^2 f_i(y)\|_{\mathrm{F}} \\ &\leq \sqrt{2}\|(\tilde{\mathcal{B}}_i)^{-1}\| H_1\|x-y\|,\end{aligned}$$

which implies that M_4 in Assumption A.1 satisfies $M_4 \leq \sqrt{2}\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|\}H_1$.

If $|(\nabla^2 f_i(x))_{jl} - (\nabla^2 f_i(y))_{jl}| \leq \nu\|x-y\|$ for any $x, y \in \mathbb{R}^d$, $i \in [n]$, and $j, l \in [d]$, then from (15), every entry of $\tilde{h}^i(\nabla^2 f_i(x)) - \tilde{h}^i(\nabla^2 f_i(y))$ will be bounded by $2\nu\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty\|x-y\|$. Hence M_5 in Assumption A.1 satisfies $M_5 \leq 2\nu\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty\}$.

(ii) If $|(\nabla^2 f_i(x))_{jl}| \leq \gamma$ for any $x \in \mathbb{R}^d$, $i \in [n]$, and $j, l \in [d]$, then from (15), every entry of $\tilde{h}^i(\nabla^2 f_i(x))$ will be bounded by $2\gamma\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty$, i.e., $\max_{jl}\{|\tilde{h}^i(\nabla^2 f_i(x))_{jl}|\} \leq 2\gamma\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty$. In particular, under Assumption 4.5 (ii), $(\mathbf{L}_i^k)_{jl}$ is a convex combination of $\{\tilde{h}^i(\nabla^2 f_i(z_i^t))_{jl}\}_{t\leq k}$, and thus M_3 in Assumption A.1 satisfies $M_3 \leq 2\gamma\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|_\infty\}$.

(iii) First, from $\|\nabla^2 f_i(x)\|_{\mathrm{F}} \leq \tilde{\gamma}$ and (15), we have

$$\begin{aligned}\|\tilde{h}^i(\nabla^2 f_i(x))\|_{\mathrm{F}} &\leq \|svec(\tilde{h}^i(\nabla^2 f_i(x)))\| \\ &\leq \|(\tilde{\mathcal{B}}_i)^{-1}\| \cdot \|svec(\nabla^2 f_i(x))\| \\ &\leq \sqrt{2}\|(\tilde{\mathcal{B}}_i)^{-1}\| \cdot \|\nabla^2 f_i(x)\|_{\mathrm{F}} \\ &\leq \sqrt{2}\|(\tilde{\mathcal{B}}_i)^{-1}\|\tilde{\gamma},\end{aligned}$$

for any $x \in \mathbb{R}^d$ and $i \in [n]$. Assume $\|\mathbf{L}_i^K\|_{\mathrm{F}}^2 \leq \frac{2B}{A}\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2$. Then under Assumption 4.6, same as Lemma F.1 (ii), we have

$$\begin{aligned}\|\mathbf{L}_i^{K+1}\|_{\mathrm{F}}^2 &= \|\mathbf{L}_i^K + \mathcal{C}_i^K\left(\tilde{h}^i(\nabla^2 f_i(z_i^{K+1})) - \mathbf{L}_i^K\right)\|_{\mathrm{F}}^2 \\ &\leq \left(1 - \frac{\delta}{4}\right)\|\mathbf{L}_i^K\|_{\mathrm{F}}^2 + \left(\frac{6}{\delta} - \frac{7}{2}\right)\|\tilde{h}^i(\nabla^2 f_i(z_i^{K+1}))\|_{\mathrm{F}}^2 \\ &= (1-A)\|\mathbf{L}_i^K\|_{\mathrm{F}}^2 + B\|\tilde{h}^i(\nabla^2 f_i(z_i^{K+1}))\|_{\mathrm{F}}^2 \\ &\leq (1-A)\cdot\frac{2B}{A}\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2 + B\cdot 2\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2 \\ &= \frac{2B}{A}\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2.\end{aligned}$$

Since $\|\mathbf{L}_i^0\|_{\mathrm{F}}^2 \leq \frac{2B}{A}\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2$, by mathematical induction, we can get $\|\mathbf{L}_i^k\|_{\mathrm{F}}^2 \leq \frac{2B}{A}\|(\tilde{\mathcal{B}}_i)^{-1}\|^2\tilde{\gamma}^2$ for $k \geq 0$. At last, from $\max_{jl}\{|(\mathbf{L}_i^k)_{jl}|\} \leq \|\mathbf{L}_i^k\|_{\mathrm{F}} \leq \frac{\sqrt{2B}}{\sqrt{A}}\|(\tilde{\mathcal{B}}_i)^{-1}\|\tilde{\gamma}$, we can obtain $M_3 \leq \frac{\sqrt{2B}\tilde{\gamma}}{\sqrt{A}}\max_i\{\|(\tilde{\mathcal{B}}_i)^{-1}\|\}$.

F.2 Lemmas

The proof of Lemma F.1 is the same as that of Lemma B.1 in (Safaryan et al., 2021). Hence we omit it.

Lemma F.1. *Let $\mathcal{C}$ be a compressor and $\alpha > 0$. For any matrix $\mathbf{L} \in \mathbb{R}^{d\times d}$ and $y, z \in \mathbb{R}^d$, we have the following results.*

(i) If $\mathcal{C}$ is an unbiased compressor with parameter ω and $\alpha \leq 1/\omega+1$, then

$$\mathbb{E}\|\mathbf{L}+\alpha\mathcal{C}(\tilde{h}^i(\nabla^2 f_i(y))-\mathbf{L})-\tilde{h}^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 \leq (1-\alpha)\|\mathbf{L}-\tilde{h}^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2+\alpha M_4^2\|y-z\|^2,$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to $\mathcal{C}$.
(ii) If $\mathcal{C}$ is a contraction compressor with parameter δ and $\alpha=1$, then

$$\mathbb{E}\|\mathbf{L}+\alpha\mathcal{C}(\tilde{h}^i(\nabla^2 f_i(y))-\mathbf{L})-\tilde{h}^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2 \leq \left(1-\frac{\delta}{4}\right)\|\mathbf{L}-\tilde{h}^i(\nabla^2 f_i(z))\|_{\mathrm{F}}^2+\left(\frac{6}{\delta}-\frac{7}{2}\right)M_4^2\|y-z\|^2.$$

The constants c_1 and c_2 in the following lemma are defined in (33).

Lemma F.2. *(i) If Assumption 4.3 (ii) holds, $\|x^0-x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2(H^2+4c_1)},\frac{M}{d}\}$, $\|z_i^k-x^*\|^2 \leq \min\{\frac{\mu^2}{4d(H^2+4c_1)},M\}$, and $\mathcal{H}^k \leq \frac{\mu^2}{4dc_2}$ for $k \leq K$, $i \in [n]$, and any $M>0$, then $\|z_i^{K+1}-x^*\|^2 \leq \min\{\frac{\mu^2}{4d(H^2+4c_1)},M\}$ for $i \in [n]$.*
(ii) If Assumption 4.4 holds, $\mathcal{H}^K \leq \frac{A_{\mathrm{M}}\mu^2}{4c_2B_{\mathrm{M}}}$, $\|z_i^k-x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)},M\}$ for $k \leq K$, $i \in [n]$, and any $M>0$, then $\|z_i^{K+1}-x^*\|^2 \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)},M\}$ for $i \in [n]$.
(iii) If Assumption 4.5(ii) holds, and $\|z_i^k-x^*\|^2 \leq \frac{M}{d^2M_5^2}$ for $k \leq K$, $i \in [n]$, and any $M>0$, then $\mathcal{H}^K \leq M$.
(iv) If Assumption 4.6 holds, $\|\mathbf{L}_i^K-\mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq M$, and $\|z_i^K-x^*\|^2 \leq \frac{AM}{BM_4^2}$ for $i \in [n]$ and any $M>0$, then $\|\mathbf{L}_i^{K+1}-\mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq M$ for $i \in [n]$.

Proof. (i) First, from the update rule of w_i^k, we know $\mathcal{Z}^k \leq \min\{\frac{\mu^2}{4d(H^2+4c_1)},M\}$ and $\mathcal{W}^k \leq \min\{\frac{\mu^2}{4d(H^2+4c_1)},M\}$ for $k \leq K$. Then for Option 1, from (34), we have

$$\begin{aligned}\|x^{k+1}-x^*\|^2 &\leq \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2+\frac{2c_1}{\mu^2}\mathcal{Z}^{k-1}\mathcal{W}^k+\frac{2c_2}{\mu^2}\mathcal{H}^k\mathcal{W}^k\\ &\leq \frac{1}{d}\mathcal{W}^k\\ &\leq \min\left\{\frac{\mu^2}{4d^2(H^2+4c_1)},\frac{M}{d}\right\},\end{aligned}$$

for $0 \leq k \leq K$. For Option 2, we can get the same bound for $\|x^{k+1}-x^*\|^2$ as above from (35).
Since $\|x^0-x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2(H^2+4c_1)},\frac{M}{d}\}$, we know $\|x^k-x^*\|^2 \leq \min\{\frac{\mu^2}{4d^2(H^2+4c_1)},\frac{M}{d}\}$ for all $0 \leq k \leq K+1$. Then for $i \in S^k$, from Assumption 4.3 (ii), we can get

$$\begin{aligned}\|z_i^{K+1}-x^*\|^2 &\leq d\max_j|(z_i^{K+1})_j-x_j^*|^2\\ &\leq d\max_{0\leq t\leq K+1}\|x^t-x^*\|^2\\ &\leq \min\left\{\frac{\mu^2}{4d(H^2+4c_1)},M\right\}.\end{aligned}$$

For $i \notin S^k$, we have

$$\|z_i^{K+1}-x^*\|^2=\|z_i^K-x^*\|^2 \leq \min\left\{\frac{\mu^2}{4d(H^2+4c_1)},M\right\}.$$

(ii) First, from the update rule of w^k, we know $\mathcal{Z}^k \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)},M\}$ and $\mathcal{W}^k \leq \min\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)},M\}$ for $k \leq K$. Then for Option 1, from (34), we have

$$\begin{aligned}\|x^{K+1}-x^*\|^2 &\leq \frac{H^2}{2\mu^2}(\mathcal{W}^K)^2+\frac{2c_1}{\mu^2}\mathcal{Z}^{K-1}\mathcal{W}^K+\frac{2c_2}{\mu^2}\mathcal{H}^K\mathcal{W}^K\\ &\leq \frac{A_{\mathrm{M}}}{B_{\mathrm{M}}}\mathcal{W}^K\\ &\leq \frac{A_{\mathrm{M}}}{B_{\mathrm{M}}}\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)},M\right\}.\end{aligned}$$

For Option 2, we can get the same bound for $\|x^{K+1}-x^*\|^2$ as above from (35). Then for $i \in S^k$, from Lemma D.1 (ii), we arrive at

$$\begin{aligned}\|z_i^{K+1}-z^*\|^2 &\le (1-A_{\mathrm{M}})\|z_i^K-x^*\|^2 + B_{\mathrm{M}}\|x^{K+1}-x^*\|^2\\ &\le (1-A_{\mathrm{M}})\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, M\right\} + A_{\mathrm{M}}\min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, M\right\}\\ &= \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, M\right\}.\end{aligned}$$

For $i \notin S^k$, we have

$$\|z_i^{K+1}-x^*\|^2 = \|z_i^K-x^*\|^2 \le \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, M\right\}.$$

(iii) From Assumption 4.5(ii), we have

$$\begin{aligned}\mathcal{H}^K &= \frac{1}{n}\sum_{i=1}^n \|\mathbf{L}_i^K-\mathbf{L}_i^*\|_{\mathrm{F}}^2\\ &\le \frac{1}{n}\sum_{i=1}^n d^2 \max_{jl}\{|(\mathbf{L}_i^K)_{jl}-(\mathbf{L}_i^*)_{jl}|^2\}\\ &\le d^2 M_5^2 \max_{i\in[n],0\le t\le K}\|z_i^t-x^*\|^2\\ &\le M.\end{aligned}$$

(iv) For $i \in S^k$, from Assumption 4.6 and Lemma F.1 (ii), we have

$$\begin{aligned}\|\mathbf{L}_i^{K+1}-\mathbf{L}_i^*\|_{\mathrm{F}}^2 &\le (1-A)\|\mathbf{L}_i^K-\mathbf{L}_i^*\|_{\mathrm{F}}^2 + BM_4^2\|z_i^K-x^*\|^2\\ &\le (1-A)M + AM\\ &= M.\end{aligned}$$

For $i \notin S^k$, we also have

$$\|\mathbf{L}_i^{K+1}-\mathbf{L}_i^*\|_{\mathrm{F}}^2 = \|\mathbf{L}_i^K-\mathbf{L}_i^*\|_{\mathrm{F}}^2 \le M.$$

□

F.3 Proof of Theorem A.3

Define $\mathbf{H}_i^k := \beta^k\mathbf{A}_i^k - \mathbf{C}_i^k$ for $i \in [n]$ and $k \ge 0$. First, it is easy to verify that $\mathbf{A}^k = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i^k$, $\mathbf{C}^k = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k$, $\mathbf{H}^k = \frac{1}{n}\sum_{i=1}^n \mathbf{H}_i^k$, $g_1^k = \frac{1}{n}\sum_{i=1}^n g_{i,1}^k$, and $g_2^k = \frac{1}{n}\sum_{i=1}^n g_{i,2}^k$ for $k \ge 0$. Then we have

$$\begin{aligned}g^k &= \beta^k g_1^k - g_2^k\\ &= \frac{1}{n}\sum_{i=1}^n \left(\beta^k g_{i,1}^k - g_{i,2}^k\right)\\ &= \frac{1}{n}\sum_{i=1}^n \left(\beta^k \mathbf{A}_i^k w_i^k - \mathbf{C}_i^k w_i^k - \nabla f_i(w_i^k)\right)\\ &= \frac{1}{n}\sum_{i=1}^n \left(\mathbf{H}_i^k w_i^k - \nabla f_i(w_i^k)\right).\end{aligned}$$

Thus, from

$$x^{k+1} = \left(\mathbf{H}^k\right)^{-1} g^k = \left(\mathbf{H}^k\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbf{H}_i^k w_i^k - \nabla f_i(w_i^k)\right)\right],$$

and

$$x^* = \left(\mathbf{H}^k\right)^{-1}\left[\mathbf{H}^k x^* - \nabla f(x^*)\right] = \left(\mathbf{H}^k\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbf{H}_i^k x^* - \nabla f_i(x^*)\right)\right],$$

we can obtain

$$x^{k+1} - x^* = \left(\mathbf{H}^k\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbf{H}_i^k(w_i^k - x^*) - (\nabla f_i(w_i^k) - \nabla f_i(x^*))\right)\right].$$

Then from the triangle inequality and the fact that $\mathbf{H}^k \succeq \mu \mathbf{I}$, we have

$$\begin{aligned}
\|x^{k+1} - x^*\| &\leq \frac{1}{\mu n}\sum_{i=1}^n \left\|\nabla f_i(w_i^k) - \nabla f_i(x^*) - \mathbf{H}_i^k(w_i^k - x^*)\right\| \\
&\leq \frac{1}{\mu n}\sum_{i=1}^n \left\|\nabla f_i(w_i^k) - \nabla f_i(x^*) - \nabla^2 f_i(x^*)(w_i^k - x^*)\right\| + \frac{1}{\mu n}\sum_{i=1}^n \left\|(\mathbf{H}_i^k - \nabla^2 f_i(x^*)(w_i^k - x^*))\right\| \\
&\leq \frac{H}{2\mu n}\sum_{i=1}^n \|w_i^k - x^*\|^2 + \frac{1}{\mu n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\| \cdot \|w_i^k - x^*\| \\
&= \frac{H}{2\mu}\mathcal{W}^k + \frac{1}{\mu n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\| \cdot \|w_i^k - x^*\|.
\end{aligned}$$

We further use Young's inequality to bound $\|x^{k+1} - x^*\|^2$ as

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &\leq \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2 + \frac{2}{\mu^2 n^2}\left(\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\| \cdot \|w_i^k - x^*\|\right)^2 \\
&\leq \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2 + \frac{2}{\mu^2}\left(\frac{1}{n}\sum_{i=1}^n \|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|^2\right)\mathcal{W}^k,
\end{aligned}$$

where we use Cauchy-Schwarz inequality in the last inequality. Next we estimate $\|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|^2$. Denote $\mathbf{L}_i^* = \tilde{h}^i(\nabla^2 f_i(x^*))$ and assume $\max_{jl}\{\|\mathbf{B}_i^{jl}\|_{\mathrm{F}}\} \leq R$ for $i \in [n]$. Then

$$\begin{aligned}
\|\mathbf{H}_i^k - \nabla^2 f_i(x^*)\|^2 &= \|\beta^k \mathbf{A}_i^k - \mathbf{C}_i^k - \nabla^2 f_i(x^*)\|^2 \\
&= \left\|\sum_{jl}\left[\beta^k((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - 2\gamma_i^k - (\mathbf{L}_i^*)_{jl}\right]\mathbf{B}^{jl}\right\|^2 \\
&\leq NR^2 \sum_{jl}\left|\beta^k((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - ((\mathbf{L}_i^*)_{jl} + 2\gamma_i^k)\right|^2.
\end{aligned}$$

Assuming $\max_{jl}\{|(\mathbf{L}_i^k)_{jl}|\} \leq M_3$ for $i \in [n]$, we have

$$\begin{aligned}
\left|\beta^k((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) - ((\mathbf{L}_i^*)_{jl} + 2\gamma_i^k)\right|^2 &= \left|(\beta^k - 1)((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k) + (\mathbf{L}_i^k - \mathbf{L}_i^*)_{jl}\right|^2 \\
&\leq 2\left|(\beta^k - 1)((\mathbf{L}_i^k)_{jl} + 2\gamma_i^k)\right|^2 + 2\left|(\mathbf{L}_i^k - \mathbf{L}_i^*)_{jl}\right|^2 \\
&\leq 2(M_3 + 2\max\{c, M_3\})^2\left|\beta^k - 1\right|^2 + 2\left|(\mathbf{L}_i^k - \mathbf{L}_i^*)_{jl}\right|^2.
\end{aligned}$$

For Option 1 in Algorithm 3, we have $\beta_i^k = \max_{jl}\frac{\tilde{h}^i(\nabla^2 f_i(z_i^{k-1}))_{jl} + 2\gamma_i^k}{(\mathbf{L}_i^k)_{jl} + 2\gamma_i^k}$, where we define $z_i^{-1} = z_i^0$. For any $j, l \in [d]$,

we have

$$\begin{aligned}\left|\frac{\tilde{h}^i(\nabla^2 f_i(z_i^{k-1}))_{jl}+2\gamma_i^k}{(\mathbf{L}_i^k)_{jl}+2\gamma_i^k}-1\right|^2 &= \left|\frac{(\tilde{h}^i(\nabla^2 f_i(z_i^{k-1}))-\mathbf{L}_i^k)_{jl}}{(\mathbf{L}_i^k+2\gamma_i^k\mathbf{I})_{jl}}\right|^2\\ &\le \frac{1}{c^2}\left|(\tilde{h}^i(\nabla^2 f_i(z_i^{k-1}))-\mathbf{L}_i^k)_{jl}\right|^2\\ &\le \frac{2}{c^2}\left|(\tilde{h}^i(\nabla^2 f_i(z_i^{k-1}))-\mathbf{L}_i^*)_{jl}\right|^2+\frac{2}{c^2}\left|(\mathbf{L}_i^k-\mathbf{L}_i^*)_{jl}\right|^2\\ &\le \frac{2M_5^2}{c^2}\|z_i^{k-1}-x^*\|^2+\frac{2}{c^2}\left|(\mathbf{L}_i^k-\mathbf{L}_i^*)_{jl}\right|^2,\end{aligned}$$

where we use $(\mathbf{L}_i^k+2\gamma_i^k\mathbf{I})_{jl}\ge(\mathbf{L}_i^k)_{jl}+|(\mathbf{L}_i^k)_{jl}|+c\ge c$ in the first inequality, in the second inequality, we use the Young's inequality, and the last inequality comes from $\max_{jl}\{|(\tilde{h}^i(\nabla^2 f_i(x))-\mathbf{L}_i^*)_{jl}|\}\le M_5\|x-x^*\|$. Then from the definition of β^k, we arrive at

$$\begin{aligned}\left|\beta^k-1\right|^2 &\le \max_{jl}\left\{\frac{2M_5^2}{c^2}\|z_i^{k-1}-x^*\|^2+\frac{2}{c^2}\left|(\mathbf{L}_i^k-\mathbf{L}_i^*)_{jl}\right|^2\right\}\\ &\le \frac{2M_5^2}{c^2}\|z_i^{k-1}-x^*\|^2+\frac{2}{c^2}\|\mathbf{L}_i^k-\mathbf{L}_i^*\|^2.\end{aligned}$$

For Option 2 in Algorithm 3, we can have the following bound in the same way.

$$\left|\beta^k-1\right|^2 \le \frac{2M_5^2}{c^2}\|z_i^k-x^*\|^2+\frac{2}{c^2}\|\mathbf{L}_i^k-\mathbf{L}_i^*\|^2.$$

For Option 1, from the above inequalities, we can get

$$\begin{aligned}\|\mathbf{H}_i^k-\nabla^2 f_i(x^*)\|^2 \le{}& \frac{4N^2R^2M_5^2(M_3+2\max\{c,M_3\})^2}{c^2}\|z_i^{k-1}-x^*\|^2\\ &+2NR^2\left(1+\frac{2N(M_3+2\max\{c,M_3\})^2}{c^2}\right)\|\mathbf{L}_i^k-\mathbf{L}_i^*\|_{\mathrm{F}}^2,\end{aligned}$$

which implies that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\|\mathbf{H}_i^k-\nabla^2 f_i(x^*)\|^2 \le{}& \frac{4N^2R^2M_5^2(M_3+2\max\{c,M_3\})^2}{c^2}\mathcal{Z}^{k-1}\\ &+2NR^2\left(1+\frac{2N(M_3+2\max\{c,M_3\})^2}{c^2}\right)\mathcal{H}^k,\end{aligned}$$

where $\mathcal{H}^k:=\frac{1}{n}\sum_{i=1}^n\|\mathbf{L}_i^k-\mathbf{L}_i^*\|_{\mathrm{F}}^2$. Similarly, for Option 2, we can get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\|\mathbf{H}_i^k-\nabla^2 f_i(x^*)\|^2 \le{}& \frac{4N^2R^2M_5^2(M_3+2\max\{c,M_3\})^2}{c^2}\mathcal{Z}^{k}\\ &+2NR^2\left(1+\frac{2N(M_3+2\max\{c,M_3\})^2}{c^2}\right)\mathcal{H}^k.\end{aligned}$$

Let

$$c_1:=\frac{4N^2R^2M_5^2(M_3+2\max\{c,M_3\})^2}{c^2},\quad c_2:=2NR^2\left(1+\frac{2N(M_3+2\max\{c,M_3\})^2}{c^2}\right)\tag{33}$$

Then for Option 1, we have

$$\begin{aligned}\|x^{k+1}-x^*\|^2 &\le \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2+\frac{2}{\mu^2}\mathcal{W}^k\left(c_1\mathcal{Z}^{k-1}+c_2\mathcal{H}^k\right)\\ &= \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2+\frac{2c_1}{\mu^2}\mathcal{Z}^{k-1}\mathcal{W}^k+\frac{2c_2}{\mu^2}\mathcal{H}^k\mathcal{W}^k,\end{aligned}\tag{34}$$

and for Option 2, we have

$$\|x^{k+1}-x^*\|^2 \leq \frac{H^2}{2\mu^2}(\mathcal{W}^k)^2 + \frac{2c_1}{\mu^2}\mathcal{Z}^k\mathcal{W}^k + \frac{2c_2}{\mu^2}\mathcal{H}^k\mathcal{W}^k. \tag{35}$$

From the update rule of w_i^k and z_i^k, the results in (31) and (32) also hold for Algorithm 3. Then for Option 1, we have

$$\begin{aligned}
\mathbb{E}_k[\Phi_5^{k+1}] &\overset{(32)}{\leq} \left(1-\frac{\tau p}{n}\right)\mathcal{W}^k + \frac{\tau p}{n}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k + \frac{\tau^2 B_{\mathrm{M}} p}{n^2}\|x^{k+1}-x^*\|^2 + \frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\mathbb{E}_k[\mathcal{Z}^{k+1}] \\
&\overset{(31)}{\leq} \left(1-\frac{\tau p}{n}\right)\mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\left(1-\frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k + \frac{2\tau p B_{\mathrm{M}}}{nA_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{2n}\right)\|x^{k+1}-x^*\|^2 \\
&\overset{34}{\leq} \left(1-\frac{\tau p}{n}+\frac{2\tau p B_{\mathrm{M}}}{nA_{\mathrm{M}}}\left(\frac{H^2}{2\mu^2}\mathcal{W}^k + \frac{2c_2}{\mu^2}\mathcal{H}^k + \frac{2c_1}{\mu^2}\mathcal{Z}^{k-1}\right)\right)\mathcal{W}^k \\
&\quad + \frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\left(1-\frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k.
\end{aligned}$$

If $\|z_i^k - x^*\|^2 \leq \frac{A_{\mathrm{M}}\mu^2}{4(H^2+4c_1)B_{\mathrm{M}}}$ and $\mathcal{H}^k \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2 B_{\mathrm{M}}}$ for all $k \geq 0$, then we have

$$\frac{H^2}{2\mu^2}\mathcal{W}^k + \frac{2c_2}{\mu^2}\mathcal{H}^k + \frac{2c_1}{\mu^2}\mathcal{Z}^{k-1} \leq \frac{A_{\mathrm{M}}}{4B_{\mathrm{M}}},$$

which implies that

$$\begin{aligned}
\mathbb{E}_k[\Phi_5^{k+1}] &\leq \left(1-\frac{\tau p}{2n}\right)\mathcal{W}^k + \frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\left(1-\frac{\tau A_{\mathrm{M}}}{2n}\right)\mathcal{Z}^k \\
&\leq \left(1-\frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)\Phi_5^k.
\end{aligned}$$

By applying the tower property, we have

$$\mathbb{E}[\Phi_5^{k+1}] \leq \left(1-\frac{\tau\min\{p, A_{\mathrm{M}}\}}{2n}\right)\mathbb{E}[\Phi_5^k].$$

Unrolling the recursion, we can obtain the result. For Option 2, we can have the same result.

F.4 Proof of Theorem A.4

Since $\xi^k \equiv 1$, $\eta = 1$, $S^k \equiv [n]$, and $\mathcal{Q}_i^k(x) \equiv x$ for any $x \in \mathbb{R}^d$, it is easy to see that $w_i^k = z_i^k \equiv x^k$ for all $i \in [n]$ and $k \geq 0$. In this case, we can view $\mathcal{Q}_i^k$ as an unbiased compressor with $\omega_{\mathrm{M}} = 0$ or a contraction compressor with $\delta_{\mathrm{M}} = 1$. Since (31) also holds for Algorithm 3, for Option 1, we have

$$\begin{aligned}
\mathbb{E}_k\|x^{k+1}-x^*\|^2 &\leq (1-A_{\mathrm{M}})\|x^k-x^*\|^2 + B_{\mathrm{M}}\|x^{k+1}-x^*\|^2 \\
&\overset{(34)}{\leq} (1-A_{\mathrm{M}})\|x^k-x^*\|^2 + \frac{1}{4}A_{\mathrm{M}}\|x^k-x^*\|^2 \\
&= \left(1-\frac{3A_{\mathrm{M}}}{4}\right)\|x^k-x^*\|^2.
\end{aligned}$$

For Option 2, we can get the same bound for $\mathbb{E}_k\|x^{k+1}-x^*\|^2$ as above from (35).

From Lemma F.1, we can obtain

$$\mathbb{E}_k[\mathcal{H}^{k+1}] \leq (1-A)\mathcal{H}^k + BM_4^2\|x^k-x^*\|^2.$$

Thus,

$$\begin{aligned}\mathbb{E}_k[\Phi_6^{k+1}] &= \mathbb{E}_k[\mathcal{H}^{k+1}] + \frac{4BM_4^2}{A_{\mathrm{M}}}\mathbb{E}_k\|x^{k+1}-x^*\|^2 \\ &\le (1-A)\mathcal{H}^k + BM_4^2\|x^k-x^*\|^2 + \frac{4BM_4^2}{A_{\mathrm{M}}}\left(1-\frac{3A_{\mathrm{M}}}{4}\right)\|x^k-x^*\|^2 \\ &\le \left(1-\frac{\min\{2A, A_{\mathrm{M}}\}}{2}\right)\Phi_6^k.\end{aligned}$$

By applying the tower property, we have $\mathbb{E}[\Phi_6^{k+1}] \le \theta_2\mathbb{E}[\Phi_6^k]$. Unrolling the recursion, we have $\mathbb{E}[\Phi_6^k] \le \theta_3^k\Phi_6^0$. Then we further have $\mathbb{E}[\mathcal{H}^k] \le \theta_3^k\Phi_6^0$ and $\mathbb{E}\|x^k-x^*\|^2 \le \frac{A_{\mathrm{M}}}{4BM_4^2}\theta_3^k\Phi_6^0$. For Option 1, from $w_i^k = z_i^k \equiv x^k$ and (34), we can get

$$\|x^{k+1}-x^*\|^2 \le \left(\frac{H^2}{2\mu^2}\|x^k-x^*\|^2 + \frac{2c_1}{\mu^2}\|x^{k-1}-x^*\|^2 + \frac{2c_2}{\mu^2}\mathcal{H}^k\right)\|x^k-x^*\|^2.$$

Assume $x^k \neq x^*$ for all $k \ge 0$. Then we have

$$\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2} \le \frac{H^2}{2\mu^2}\|x^k-x^*\|^2 + \frac{2c_1}{\mu^2}\|x^{k-1}-x^*\|^2 + \frac{2c_2}{\mu^2}\mathcal{H}^k,$$

and by taking expectation, we arrive at

$$\begin{aligned}\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] &\le \frac{H^2}{2\mu^2}\mathbb{E}\|x^k-x^*\|^2 + \frac{2c_1}{\mu^2}\mathbb{E}\|x^{k-1}-x^*\|^2 + \frac{2c_2}{\mu^2}\mathbb{E}[\mathcal{H}^k] \\ &\le \theta_3^k\left(\frac{A_{\mathrm{M}}(H^2\theta_3+4c_1)}{8BM_4^2\mu^2\theta_3} + \frac{2c_2}{\mu^2}\right)\Phi_6^0.\end{aligned}$$

For Option 2, from $w_i^k = z_i^k \equiv x^k$ and (35), we can get

$$\|x^{k+1}-x^*\|^2 \le \left(\frac{H^2+4c_1}{2\mu^2}\|x^k-x^*\|^2 + \frac{2c_2}{\mu^2}\mathcal{H}^k\right)\|x^k-x^*\|^2.$$

Assume $x^k \neq x^*$ for all $k \ge 0$. Then we have

$$\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2} \le \frac{H^2+4c_1}{2\mu^2}\|x^k-x^*\|^2 + \frac{2c_2}{\mu^2}\mathcal{H}^k,$$

and by taking expectation, we arrive at

$$\begin{aligned}\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right] &\le \frac{H^2+4c_1}{2\mu^2}\mathbb{E}\|x^k-x^*\|^2 + \frac{2c_2}{\mu^2}\mathbb{E}[\mathcal{H}^k] \\ &\le \theta_3^k\left(\frac{A_{\mathrm{M}}(H^2+4c_1)}{8BM_4^2\mu^2} + \frac{2c_2}{\mu^2}\right)\Phi_6^0.\end{aligned}$$

F.5 Proof of Theorem A.5

(i) Noticed that under Assumption 4.3, we have $A_{\mathrm{M}} = B_{\mathrm{M}} = \eta$. We prove this by mathematical induction. First, since $z_i^0 = x^0$, we know $\|z_i^0 - x^*\|^2 \le \min\{\frac{\mu^2}{4d(H^2+4c_1)}, \frac{\mu^2}{16d^3c_2M_5^2}\}$ for $i \in [n]$. Then from Lemma F.2 (iii), we have $\mathcal{H}^0 \le \frac{\mu^2}{16dc_2}$. Next, assume

$$\|z_i^k - x^*\|^2 \le \min\left\{\frac{\mu^2}{4d(H^2+4c_1)}, \frac{\mu^2}{16d^3c_2M_5^2}\right\} \text{ for } i \in [n] \quad \text{and} \quad \mathcal{H}^k \le \frac{\mu^2}{16dc_2},$$

for $k \le K$. By choosing $M = \frac{\mu^2}{16d^3c_2M_5^2}$ in Lemma F.2 (i), we have

$$\|z_i^{K+1} - x^*\|^2 \le \min\left\{\frac{\mu^2}{4d(H^2+4c_1)}, \frac{\mu^2}{16d^3c_2M_5^2}\right\},$$

for $i \in [n]$. By further using Lemma F.2 (iii), we can get $\mathcal{H}^{K+1} \leq \frac{\mu^2}{16dc_2}$.

(ii) We prove the result by induction. Assume $\|z_i^k - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, \frac{AA_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}BM_4^2}\right\}$ and $\|\mathbf{L}_i^k - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$ for all $i \in [n]$ and $k \leq K$. Then by Lemma F.2 (iv), we have $\|\mathbf{L}_i^{K+1} - \mathbf{L}_i^*\|_{\mathrm{F}}^2 \leq \frac{A_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}}$. Moreover, by Lemma F.2 (ii), we have $\|z_i^{K+1} - x^*\|^2 \leq \min\left\{\frac{A_{\mathrm{M}}\mu^2}{4B_{\mathrm{M}}(H^2+4c_1)}, \frac{AA_{\mathrm{M}}\mu^2}{16c_2B_{\mathrm{M}}BM_4^2}\right\}$ for $i \in [n]$.