Error Compensated Distributed SGD can be Accelerated

Xun Qian   XUN.QIAN@KAUST.EDU.SA
KAUST

Peter Richtárik   PETER.RICHTARIK@KAUST.EDU.SA
KAUST

Tong Zhang   TONGZHANG@UST.HK
Hong Kong University of Science and Technology

Abstract

Gradient compression is a recent and increasingly polular technique for reducing the communication cost in distributed training of large-scale machine learning models. In this work we focus on developing efficient distributed methods that can work for any compressor satisfying a certain contraction property, which includes both unbiased (after appropriate scaling) and biased compressors such as RandK and TopK. Applied naively, gradient compression introduces errors that either slow down convergence or lead to divergence. A popular technique designed to tackle this issue is error compensation/error feedback. Due to the difficulties associated with analyzing biased compressors, it is not known whether gradient compression with error compensation can be combined with Nesterov's acceleration. In this work, we show for the first time that error compensated gradient compression methods can be accelerated. In particular, we propose and study the error compensated loopless Katyusha method, and establish an accelerated linear convergence rate under standard assumptions. We show through numerical experiments that the proposed method converges with substantially fewer communication rounds than previous error compensated algorithms.

1. Introduction

In this work we consider the composite finite-sum optimization problem

$$\min_{x \in \mathbb{R}^d} \left[P(x) := \tfrac{1}{n} \sum_{\tau=1}^{n} f^{(\tau)}(x) + \psi(x) \right], \tag{1}$$

where $f(x) := \frac{1}{n} \sum_\tau f^{(\tau)}(x)$ is an average of n smooth[1] convex functions $f^{(\tau)} : \mathbb{R}^d \to \mathbb{R}$ distributed over n nodes (devices, computers), and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function representing a possibly nonsmooth regularizer. On each node, $f^{(\tau)}(x)$ is an average of m smooth convex functions

$$f^{(\tau)}(x) = \tfrac{1}{m} \sum_{i=1}^{m} f_i^{(\tau)}(x), \tag{2}$$

representing the average loss over the training data stored on node τ. While we specifically focus on the case when $m = 1$, our results are also new in the $m = 1$ case, and hence this regime is relevant as well. We assume throughout that problem (1) has at least one optimal solution x^*. We denote the

---

1. We say that a function $\phi : \mathbb{R}^d \to \mathbb{R}$ is smooth if it is differentiable, and has L_ϕ Lipschitz gradient: $\|\nabla\phi(x) - \phi(y)\| \leq L_\phi \|x - y\|$ for all $x, y \in \mathbb{R}^d$. We say that L_ϕ is the *smoothness constant* of ϕ.

smoothness constants of functions f, $f^{(\tau)}$ and $f_i^{(\tau)}$ using symbols L_f, $\bar{L}$ and L, respectively. These constants are in general related as follows:

$$L_f \leq \bar{L} \leq n L_f, \qquad \bar{L} \leq L \leq m\bar{L}. \tag{3}$$

When training very large scale supervised machine learning problems, such as those arising in the context of federated learning [13, 14, 20] (see also recent surveys [10, 16]), distributed algorithms need to be used. In such settings, communication is generally much slower than (local) computation, which makes communication the key bottleneck in the design of efficient distributed systems. There are several ways to tackle this issue, including reliance on large mini-batches [7, 33], asynchronous learning [1, 18, 23], local updates [8, 12, 19, 27, 31] and communication compression (e.g., quantization and sparsification) [2, 4, 21, 24, 30]. In this work we focus on the last of these techniques: communication compression.

1.1. Communication compression

Contraction and unbiased compressors. We say that a randomized map $Q : \mathbb{R}^d \to \mathbb{R}^d$ is a *contraction compressor* if there exists a constant $0 < \delta \leq 1$ such that

$$\mathbb{E}\left[\|x - Q(x)\|^2\right] \leq (1 - \delta)\|x\|^2, \quad \forall x \in \mathbb{R}^d. \tag{4}$$

Further, we say that a randomized map $\tilde{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is an *unbiased compressor* if there exists a constant $\omega \geq 0$ such that

$$\mathbb{E}[\tilde{Q}(x)] = x \quad \text{and} \quad \mathbb{E}\|\tilde{Q}(x)\|^2 \leq (\omega + 1)\|x\|^2, \qquad \forall x \in \mathbb{R}^d. \tag{5}$$

It is well known that (see, e.g., [5]) after appropriate scaling, any unbiased compressor satisfying (5) becomes a contraction compressor. Indeed, for any $\tilde{Q}$ satisfying (5), $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor satisfying (4) with $\delta = \frac{1}{\omega+1}$, as shown here:

$$\mathbb{E}\left[\left\|\tfrac{1}{\omega+1}\tilde{Q}(x) - x\right\|^2\right] = \tfrac{1}{(\omega+1)^2}\mathbb{E}\left[\|\tilde{Q}(x)\|^2\right] + \|x\|^2 - \tfrac{2}{\omega+1}\mathbb{E}\left[\langle\tilde{Q}(x), x\rangle\right]$$

$$\leq \tfrac{1}{\omega+1}\|x\|^2 + \|x\|^2 - \tfrac{2}{\omega+1}\|x\|^2 = \left(1 - \tfrac{1}{\omega+1}\right)\|x\|^2.$$

Since compressors are typically applied in a scaled fashion, using a scaling stepsize, this means that for all practical purposes, the class of unbiased compressors is included in the class of contraction compressors. For examples of contraction and unbiased compressors, we refer the reader to [5].

1.2. Error compensation

While compression reduces the communicated bits in each communication round, it introduces errors, which generally leads to an increase in the number of communication rounds needed to find a solution of any predefined accuracy. Still, compression has been found useful in practice, as the trade-off often seems to prefer compression to no compression. In order to deal with the errors introduced by compression, some form of error compensation/error feedback is needed.

If we assume that the accumulated error is bounded, and in the case of unbiased compressors, the convergence rate of error compensated SGD was shown to be the same as that of vanilla SGD

[28]. However, if we only assume bounded second moment of the stochastic gradients, in order to guarantee the boundedness of the accumulated quantization error, some decaying factor needs to be involved in general, and error compensated SGD is proved to have some advantage over QSGD in some perspective for convex quadratic problem [32]. On the other hand, for contraction compressors (for example, the TopK compressor [3]), error compensated SGD actually has the same convergence rate as vanilla SGD [25, 26, 29]. Since SGD only has a sublinear convergence rate, the current error compensated methods could not get linear convergence rate. If f is non-smooth and $\psi = 0$, error compensated SGD was studied in [11] in the single node case, and the convergence rate is of order $O\left(1/\sqrt{\delta k}\right)$.

For variance-reduced methods, QSVRG [2] handles the smooth case ($\psi \equiv 0$) and VR-DIANA [9] handles the composite case (general ψ). However, the compressors of both algorithms need to be unbiased. Error compensation in VR-DIANA does not need to be used since this method successfully employs variance reduction (of the variance introduced by the compressor) instead. In this paper, we study error compensation in conjunction with the acceleration mechanism employed in loopless Katyusha (L-Katyusha) [15], for any contraction compressor.

1.3. Contributions

We now summarize the main contributions of our work.

Acceleration for error compensation. We develop a new communication efficient algorithm for solving the distributed optimization problem (1)–(2) which we call *Error Compensated Loopless Katyusha* (ECLK); see Algorithm 1. ECLK is the *first accelerated* error compensated SGD method, and can be seen as an EC variant of the Loopless Katyusha method developed in [15].

Iteration complexity. We obtain the *first accelerated linear convergence rate* for error compensated methods using contraction operators. The iteration complexity of ECLK is

$$O\left(\left(\tfrac{1}{\delta} + \tfrac{1}{p} + \sqrt{\tfrac{L_f}{\mu}} + \sqrt{\tfrac{L}{\mu pn}} + \tfrac{1}{\delta}\sqrt{\tfrac{(1-\delta)\bar{L}}{\mu p}} + \sqrt{\tfrac{(1-\delta)L}{\mu p \delta}}\right) \log \tfrac{1}{\epsilon}\right),$$

where $p \in (0, 1]$ is a parameter of the method described later. This is an improvement over the previous best known result for error compensated SGD by Beznosikov et al. [5], who obtain *nonaccelerated* linear rate. Moreover, they only consider the special case when $\psi \equiv 0$, and for their linear rate, they need to assume that $\nabla f^{(\tau)}(x^*) = 0$ for all τ, and that full gradients are computed by all nodes.

If we invoke additional assumptions (Assumption 2.1 or Assumption 2.2) on the contraction compressor, the iteration complexity is improved to

$$O\left(\left(\tfrac{1}{\delta} + \tfrac{1}{p} + \sqrt{\tfrac{L_f}{\mu}} + \sqrt{\tfrac{L}{\mu pn}} + \tfrac{1}{\delta}\sqrt{\tfrac{(1-\delta)L_f}{\mu p}} + \sqrt{\tfrac{(1-\delta)L}{\mu p \delta n}}\right) \log \tfrac{1}{\epsilon}\right).$$

This is indeed an improvement since $L_f \leq \bar{L}$ (see (3)), and because of the extra scaling factor of n in the last term. If $\delta = 1$, i.e., if no compression is used, we recover the iteration complexity of the accelerated method L-Katyusha [22].

Communication complexity. Considering the communication complexity, the optimal choice of p is $O(r(Q))$, where $r(Q)$ is the *compression ratio* for the compressor Q defined in (12). In particular, when $L_f = \bar{L} = L$, by choosing the optimal p, the communication complexity becomes

$$O\left(\Delta_1\left(\frac{r(Q)}{\delta} + \left(r(Q) + \frac{\sqrt{r(Q)}}{\sqrt{n}} + \frac{\sqrt{(1-\delta)r(Q)}}{\delta}\right)\sqrt{\frac{L}{\mu}}\right)\log\frac{1}{\epsilon}\right),$$

where Δ_1 is the communication cost of the uncompressed vector $x \in \mathbb{R}^d$.

2. Gradient Compression Methods

2.1. TopK and RandK

We now give two canonical examples of contraction and unbiased compression operators.

Example 1 (TopK compressor) *For a parameter $1 \leq K \leq d$, the TopK compressor is defined as*

$$(\mathrm{TopK}(x))_{\pi(i)} = \begin{cases} (x)_{\pi(i)} & \text{if } i \leq K, \\ 0 & \text{otherwise,} \end{cases}$$

where π is a permutation of $\{1, 2, ..., d\}$ such that $(|x|)_{\pi(i)} \geq (|x|)_{\pi(i+1)}$ for $i = 1, ..., d-1$, and if $(|x|)_{\pi(i)} = (|x|)_{\pi(i+1)}$, then $\pi(i) \leq \pi(i+1)$.

The definition of TopK compressor is slightly different with that of [26]. In this way, TopK compressor is a deterministic operator (well-defined when there are equal dimensions).

Example 2 (RandK compressor) *Given $1 \leq K \leq d$, the* RandK *compressor is defined as*

$$(\mathrm{RandK}(x))_i = \begin{cases} (x)_i & \text{if } i \in S, \\ 0 & \text{otherwise,} \end{cases}$$

where S is chosen uniformly from the set of all K element subsets of $\{1, 2, ..., d\}$. RandK *can be used to define an unbiased compressor via scaling. Indeed, it is easy to see that $\mathbb{E}\left(\frac{d}{K}\mathrm{RandK}(x)\right) = x$ for all $x \in \mathbb{R}^d$.*

For the TopK and RandK compressors, we have the following property.

Lemma 1 (Lemma A.1 in [26]) *For the* TopK *and* RandK *compressors with $1 \leq K \leq d$, we have*

$$\mathbb{E}\left[\|\mathrm{TopK}(x) - x\|^2\right] \leq \left(1 - \frac{K}{d}\right)\|x\|^2, \qquad \mathbb{E}\left[\|\mathrm{RandK}(x) - x\|^2\right] \leq \left(1 - \frac{K}{d}\right)\|x\|^2.$$

2.2. Further assumptions

We will optionally use the following additional assumptions for the contraction compressor. These assumptions are not necessary, but when used, they will lead to better complexity.

Assumption 2.1 $\mathbb{E}[Q(x)] = \delta x$ *and all $x \in \mathbb{R}^d$.*

It is easy to verify that RandK compressor satisfies Assumption 2.1 with $\delta = \frac{K}{d}$, and $\frac{1}{\omega+1}\tilde{Q}$, where $\tilde{Q}$ is any unbiased compressor, also satisfies Assumption 2.1 with $\delta = \frac{1}{\omega+1}$.

Assumption 2.2 *For $x_\tau = \frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k \in \mathbb{R}^d$, $\tau = 1, ..., n$ and $k \geq 0$ in Algorithm 1, there exist $\delta' > 0$ such that $\mathbb{E}[Q(x_\tau)] = Q(x_\tau)$, and $\left\| \sum_{\tau=1}^n (Q(x_\tau) - x_\tau) \right\|^2 \leq (1 - \delta') \left\| \sum_{\tau=1}^n x_\tau \right\|^2$.*

Since TopK is deterministic, we have $\mathbb{E}[Q(x)] = Q(x)$ for any $x \in \mathbb{R}^d$. If $Q(x_\tau)$ is close to x_τ, then δ' could be larger than $\frac{K}{d}$. Whenever Assumption 2.2 is needed, if $\delta > \delta'$, we could decrease δ such that $\delta = \min\{\delta, \delta'\}$. In this way, we have the uniform parameter δ for the contraction compressor.

3. Error Compensated L-Katyusha

3.1. Description of the method

In this section we describe our method: error compensated L-Katyusha (see Algorithm 1). The search direction in L-Katyusha in the distributed setting ($n \geq 1$) at iteration k is

$$\frac{1}{n} \sum_{\tau=1}^n \left(\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k) + \nabla f^{(\tau)}(w^k) \right), \tag{6}$$

where i_k^τ is sampled uniformly and independently from $[m] := \{1, 2, ..., m\}$ on the τ-th node for $1 \leq \tau \leq n$, x^k is the current iteration, and w^k is the current reference point. Whenever ψ is nonzero in problem (1), $\nabla f(x^*)$ is nonzero in general, and so is $\nabla f^{(\tau)}(x^*)$. Thus, compressing the direction $\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k) + \nabla f^{(\tau)}(w^k)$ directly on each node would cause nonzero noise even if x^k and w^k converged to the optimal solution x^*. On the other hand, since $f_i^{(\tau)}$ is L-smooth, $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k)$ could be small if x^k and w^k are close enough. Thus, we compress the vector $\frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k$ on each node instead. The accumulated error e_τ^{k+1} is equal to the compression error at iteration k for each node. On each node, a scalar u_τ^k is also maintained, and only u_1^k will be updated. The summation of u_τ^k is u^k, and we use u^k to control the update frequency of the reference point w^k. All nodes maintain the same copies of $x^k, w^k, y^k, z^k, \tilde{g}^k$, and u^k. Each node sends their compressed vector $\tilde{g}_\tau^k = Q(\frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k)$ and u_τ^{k+1} to the other nodes. If $u^k = 1$, each node also sends $\nabla f^{(\tau)}(w^k)$ to the other nodes. After the compressed vector $\tilde{g}_\tau^k$ is received, we add $\frac{\eta}{\mathcal{L}_1} \nabla f(w^k)$ to it as the search direction. We also need the following standard proximal operator: $\text{prox}_{\eta\psi}(x) := \arg\min_y \left\{ \frac{1}{2} \|x - y\|^2 + \eta\psi(y) \right\}$. The reference point w^k will be updated if $u^{k+1} = 1$. It is easy to see that w^k will be updated with propobility p at each iteration.

3.2. Convergence analysis: preliminaries

We now introduce some perturbed vectors which will be used in the convergence analysis. In Algorithm 1, let $e^k = \frac{1}{n} \sum_{\tau=1}^n e_\tau^k$, $g^k = \frac{1}{n} \sum_{\tau=1}^n g_\tau^k$, and $\tilde{x}^k = x^k - \frac{1}{1+\eta\sigma_1} e^k$, $\tilde{z}^k = z^k - \frac{1}{1+\eta\sigma_1} e^k$

Algorithm 1: Error Compensated Loopless Katyusha (ECLK)

Parameters: stepsize parameters $\eta = \frac{1}{3\theta_1} > 0$, $\mathcal{L}_1 > 0$, $\sigma_1 = \frac{\mu_f}{2\mathcal{L}_1} \geq 0$, $\theta_1, \theta_2 \in (0,1)$;
probability $p \in (0,1]$

Initialization: $x^0 = y^0 = z^0 = w^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$

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

    for $\tau = 1, \ldots, n$ **do**

        Sample i_k^τ uniformly and independently in $[m]$ on each node

        $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k), \quad \tilde{g}_\tau^k = Q(\frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k), \quad e_\tau^{k+1} = e_\tau^k + \frac{\eta}{\mathcal{L}_1} g_\tau^k - \tilde{g}_\tau^k$

        $u_\tau^{k+1} = 0$ for $\tau = 2, \ldots, n$, $\quad u_1^{k+1} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$

        Send $\tilde{g}_\tau^k$ and u_τ^{k+1} to the other nodes. Send $\nabla f^{(\tau)}(w^k)$ to the other nodes if $u^k = 1$

        Receive $\tilde{g}_\tau^k$ and u_τ^{k+1} from the other nodes. Receive $\nabla f^{(\tau)}(w^k)$ from the other nodes if
        $u^k = 1$

    end

    $\tilde{g}^k = \frac{1}{n}\sum_{\tau=1}^n \tilde{g}_\tau^k, \quad u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1}$

    $z^{k+1} = \text{prox}_{\frac{\eta}{(1+\eta\sigma_1)\mathcal{L}_1}\psi}\left(\frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1 x^k + z^k - \tilde{g}^k - \frac{\eta}{\mathcal{L}_1}\nabla f(w^k)\right)\right)$

    $y^{k+1} = x^k + \theta_1(z^{k+1} - z^k), \quad w^{k+1} = \begin{cases} y^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$

    $x^{k+1} = \theta_1 z^{k+1} + \theta_2 w^{k+1} + (1 - \theta_1 - \theta_2)y^{k+1}$

end

---

for $k \geq 0$. Then $e^{k+1} = \frac{1}{n}\sum_{\tau=1}^n\left(e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k - \tilde{g}_\tau^k\right) = e^k + \frac{\eta}{\mathcal{L}_1}g^k - \tilde{g}^k$, and

$$\begin{aligned}
\tilde{z}^{k+1} &= z^{k+1} - \frac{1}{1+\eta\sigma_1}e^{k+1} \\
&= \frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1 x^k + z^k - \tilde{g}^k - \frac{\eta}{\mathcal{L}_1}\nabla f(w^k)\right) - \frac{\eta\partial\psi(z^{k+1})}{(1+\eta\sigma_1)\mathcal{L}_1} - \frac{e^{k+1}}{1+\eta\sigma_1} \\
&= \frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1 x^k + z^k - e^k - \frac{\eta}{\mathcal{L}_1}g^k - \frac{\eta}{\mathcal{L}_1}\nabla f(w^k)\right) - \frac{\eta\partial\psi(z^{k+1})}{(1+\eta\sigma_1)\mathcal{L}_1} \\
&= \frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1\tilde{x}^k + \tilde{z}^k - \frac{\eta}{\mathcal{L}_1}g^k - \frac{\eta}{\mathcal{L}_1}\nabla f(w^k)\right) - \frac{\eta\partial\psi(z^{k+1})}{(1+\eta\sigma_1)\mathcal{L}_1}. \quad (7)
\end{aligned}$$

The above relation plays an important role in the convergence analysis, and allows us to follow the analysis of original L-Katyusha. We need the following assumption in this section.

Assumption 3.1 $f_i^{(\tau)}$ is L-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth and μ_f-strongly convex, and ψ is μ_ψ-strongly convex. $\mu_f \geq 0$, $\mu_\psi \geq 0$, and $\mu = \mu_f + \mu_\psi > 0$.

We define some notations which will be used to construct the Lyapunov functions in the convergence analysis. Define $\tilde{\mathcal{Z}}^k = \frac{\mathcal{L}_1 + \eta\mu/2}{2\eta}\|\tilde{z}^k - x^*\|^2$, $\mathcal{Y}^k = \frac{1}{\theta_1}(P(y^k) - P^*)$, and $\mathcal{W}^k = \frac{\theta_2}{pq\theta_1}(P(w^k) - P^*)$. From the update rule of w^k in Algorithm 1, it is easy to see that

$$\mathbb{E}_k[\mathcal{W}^{k+1}] = (1-p)\mathcal{W}^k + \frac{\theta_2}{q}\mathcal{Y}^k, \quad (8)$$

for $k \geq 0$. In the next lemma, we describe the evolution of the terms $\tilde{\mathcal{Z}}^k$ and $\mathcal{Y}^k$.

Lemma 2 *If $\mathcal{L}_1 \geq L_f$ and $\theta_1 + \theta_2 \leq 1$, then $\mathbb{E}_k \left[\tilde{\mathcal{Z}}^{k+1} + \mathcal{Y}^{k+1} \right]$ can be upper bounded by*

$$\frac{\mathcal{L}_1 \tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2)\mathcal{Y}^k + pq\mathcal{W}^k + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2} \right) \|e^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2} \right) \mathbb{E}_k \|e^{k+1}\|^2$$

$$- \frac{1}{\theta_1} \left(\theta_2 - \frac{2L}{n\mathcal{L}_1} \right) (f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).$$

Because of the compression, we have the additional error terms $\|e^k\|^2$ and $\|e^{k+1}\|^2$ in the evolution of $\tilde{\mathcal{Z}}^k$ and $\mathcal{Y}^k$ in Lemma 2. However, from the contraction property of the compressor, we can obtain inequalities controlling the evolution of $\frac{1}{n} \sum_{\tau=1}^{n} \|e_\tau^k\|^2$ and $\|e^k\|^2$ in the following two lemmas.

Lemma 3 *Th quantity $\mathbb{E}_k \left[\frac{1}{n} \sum_{\tau=1}^{n} \|e_\tau^{k+1}\|^2 \right]$ is upper bounded by the expression*

$$\left(1 - \frac{\delta}{2} \right) \frac{1}{n} \sum_{\tau=1}^{n} \|e_\tau^k\|^2 + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2} \left(\frac{2\bar{L}}{\delta} + L \right) (f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).$$

Lemma 4 *Under Assumption 2.1 or 2.2, the quantity $\mathbb{E}_k[\|e^{k+1}\|^2]$ is upper bounded by*

$$\left(1 - \frac{\delta}{2} \right) \|e^k\|^2 + \frac{2(1-\delta)\delta}{n^2} \sum_{\tau=1}^{n} \|e_\tau^k\|^2 + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2} \left(\frac{2L_f}{\delta} + \frac{3L}{n} \right) (f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).$$

3.3. Convergence analysis: main results

From the above three lemmas, we can construct suitable Lyapunov functions which enable us to prove linear convergence. First, we construct the Lyapunov function Ψ^k for the general case as follows. Let $\mathcal{L}_2 := \frac{4L}{n} + \frac{112(1-\delta)\bar{L}}{9\delta^2} + \frac{56(1-\delta)L}{9\delta}$, and for $k \geq 0$ define

$$\Phi^k := \tilde{\mathcal{Z}}^k + \mathcal{Y}^k + \mathcal{W}^k + \frac{4\mathcal{L}_1}{\delta\eta} \cdot \frac{1}{n} \sum_{\tau=1}^{n} \|e_\tau^k\|^2.$$

We are now ready to state our main convergence theorems.

Theorem 5 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. If $\mathcal{L}_1 \geq \max\{L_f, 3\mu\eta\}$, $\theta_1 + \theta_2 \leq 1$, and $\theta_2 \geq \frac{\mathcal{L}_2}{2\mathcal{L}_1}$, then we have*

$$\mathbb{E}\left[\Phi^k \right] \leq \left(1 - \min\left(\frac{\mu}{\mu + 6\theta_1\mathcal{L}_1}, \theta_1 + \theta_2 - \frac{\theta_2}{q}, p(1-q), \frac{\delta}{6} \right) \right)^k \Phi^0, \quad \forall k \geq 0.$$

If Assumption 2.1 or Assumption 2.2 holds, we can define the Lyapunov function Ψ^k as follows. Let $\mathcal{L}_3 := \frac{4L}{n} + \frac{784(1-\delta)L_f}{9\delta^2} + \frac{56(1-\delta)L}{\delta n}$, and for $k \geq 0$ define

$$\Psi^k := \tilde{\mathcal{Z}}^k + \mathcal{Y}^k + \mathcal{W}^k + \frac{4\mathcal{L}_1}{\delta\eta} \|e^k\|^2 + \frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n} \cdot \frac{1}{n} \sum_{\tau=1}^{n} \|e_\tau^k\|^2,$$

Theorem 6 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Assume Assumption 2.1 or Assumption 2.2 holds. If $\mathcal{L}_1 \geq \max\{L_f, 3\mu\eta\}$, $\theta_1 + \theta_2 \leq 1$, and $\theta_2 \geq \frac{\mathcal{L}_3}{2\mathcal{L}_1}$, then we have*

$$\mathbb{E}\left[\Psi^k \right] \leq \left(1 - \min\left(\frac{\mu}{\mu + 6\theta_1\mathcal{L}_1}, \theta_1 + \theta_2 - \frac{\theta_2}{q}, p(1-q), \frac{\delta}{6} \right) \right)^k \Psi^0, \quad \forall k \geq 0.$$

In order to cast the above results into a more digestable form, we formulate the following corollary.

Corollary 7 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Let $\mathcal{L}_1 = \max(\mathcal{L}_4, L_f, 3\mu\eta)$, $\theta_2 = \frac{\mathcal{L}_4}{2\max\{L_f, \mathcal{L}_4\}}$ and*

$$
\theta_1 = \begin{cases} \min\left(\sqrt{\frac{\mu}{\mathcal{L}_4 p}}\theta_2, \theta_2\right) & \text{if } L_f \leq \frac{\mathcal{L}_4}{p} \\ \min\left(\sqrt{\frac{\mu}{L_f}}, \frac{p}{2}\right) & \text{otherwise} \end{cases}.
$$

(i) Let $\mathcal{L}_4 = \mathcal{L}_2$. Then with some $q \in [\frac{2}{3}, 1)$, $\mathbb{E}[\Phi^k] \leq \epsilon\Phi^0$ for

$$
k \geq O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \sqrt{\frac{L_f}{\mu}} + \sqrt{\frac{L}{\mu p n}} + \frac{1}{\delta}\sqrt{\frac{(1-\delta)\bar{L}}{\mu p}} + \sqrt{\frac{(1-\delta)L}{\mu p \delta}}\right)\log\frac{1}{\epsilon}\right). \tag{9}
$$

(ii) Let $\mathcal{L}_4 = \mathcal{L}_3$. If Assumption 2.1 or 2.2 holds, then for some $q \in [\frac{2}{3}, 1)$, we have $\mathbb{E}[\Psi^k] \leq \epsilon\Psi^0$ for

$$
k \geq O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \sqrt{\frac{L_f}{\mu}} + \sqrt{\frac{L}{\mu p n}} + \frac{1}{\delta}\sqrt{\frac{(1-\delta)L_f}{\mu p}} + \sqrt{\frac{(1-\delta)L}{\mu p \delta n}}\right)\log\frac{1}{\epsilon}\right). \tag{10}
$$

Noticing that $L_f \leq \bar{L} \leq nL_f$ and $\bar{L} \leq L \leq m\bar{L}$, the iteration complexity in (10) could be better than that in (9). On the other hand, if $L_f = \bar{L} = L$, then both iteration complexities in (9) and (10) become

$$
O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \sqrt{\frac{L}{\mu}} + \sqrt{\frac{L}{\mu p n}} + \frac{1}{\delta}\sqrt{\frac{(1-\delta)L}{\mu p}}\right)\log\frac{1}{\epsilon}\right). \tag{11}
$$

References

[1] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[2] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[3] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

[4] J. Bernstein, Y. X. Wang, K. Azizzadenesheli, and A. Anandkumar. Signsgd: Compressed optimisation for non-convex problems. *The 35th International Conference on Machine Learning*, pages 560–569, 2018.

[5] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On biased compression for distributed learning. *arXiv:2002.12410*, 2020.

[6] 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.

[7] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv: 1706.2677*, 2017.

[8] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv:2002.05516*, 2020.

[9] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv: 1904.05115*, 2019.

[10] Peter et al Kairouz. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

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

[12] 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.

[13] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016.

[14] 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.

[15] D. Kovalev, S.l Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

[16] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

[17] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[18] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

[19] Chenxin Ma, Jakub Konečný, Martin Jaggi, Virginia Smith, Michael I. Jordan, Peter Richtárik, and Martin Takáč. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017.

[20] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[21] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed learning with compressed gradient differences. *arXiv: 1901.09269*, 2019.

[22] Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019.

[23] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[24] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data- parallel distributed training of speech DNNs. *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[25] S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv: 1909.05350*, 2019.

[26] S. U. Stich, J. B. Cordonnier, and M. Jaggi. Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[27] Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2020.

[28] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu. Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, pages 7652–7662, 2018.

[29] H. Tang, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6155–6165, 2019.

[30] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

[31] Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? *arXiv preprint arXiv:2002.07839*, 2020.

[32] J. Wu, W. Huang, J. Huang, and T. Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. *The 35th International Conference on Machine Learning*, pages 5321–5329, 2018.

[33] Y. You, I. Gitman, and B. Ginsburg. Scaling SGD batch size to 32k for imagenet training. *arXiv: 1708.03888*, 2017.

Appendix

Contents

Appendix A. Experiments

In this section, we experimentally study the performance of error compensated L-Katyusha used with several contraction compressors on the logistic regression problem for binary classification:

$$x \mapsto \log\left(1 + \exp(-y_i A_i^T x)\right) + \frac{\lambda}{2}\|x\|^2,$$

where $\{A_i, y_i\}$ is the data point. We use two datasets, namely, $a5a$ and $mushrooms$ from the LIBSVM library [6]. The regularization parameter $\lambda = 10^{-3}$. The number of nodes in our experiments is 20, and the optimal solution is obtained by running the uncompressed L-Katyusha for 10^5 iterations. We use the parameter setting in Corollary 7 (ii). We calculate the theoretical L_f and L as L_f^{th} and L^{th} respectively. Then we choose $L_f = t \cdot L_f^{th}$ and $L = t \cdot L^{th}$, and search the best t for $t \in \{10^{-k}|k = 0, 1, 2...\}$ in each case.

Compressors. In our experiments, we use two contraction compressors: TopK compressor with $K = 1$ and compressor $\frac{1}{\omega+1}\tilde{Q}$, where $\tilde{Q}$ is the unbiased random dithering compressor in [2] with level $s = 2^1$. For TopK compressor, $r(Q) = \frac{K(64+\lceil \log d \rceil)}{64d}$. For random dithering compressor, from Theorem 3.2 in [2], we can get

$$r(Q) = \frac{1}{64d}\left(\left(3 + \left(\frac{3}{2} + o(1)\right)\log\left(\frac{2(s^2+d)}{s(s+\sqrt{d})}\right)\right)s(s + \sqrt{d}) + 64\right).$$

A.1. TopK vs Random dithering vs No compression

In this subsection, we compare the uncompressed L-Katyusha with the error compensated L-Katyusha with two contraction compressors: TopK compressor and random dithering compressor. For simplicity, we choose $p = r(Q)$, and explore the influence of p in the next subsection. Figure 5 and figure 6 show the iteration complexity and communication complexity of them respectively. We can see that compared with the uncompressed L-Katyusha, the error compensated L-Katyusha with TopK and random dithering compressors need more iterations to reach the optimal solution, but need much less communication bits. In particular, the error compensated L-Katyusha with Top1 compressor is more than 10 times faster than the umcompressed L-Katyusha considering the communication complexity.

A.2. Influence of p

In this subsection, we show the influence of the parameter p for the communication complexity of the error compensated L-Katyusha with TopK and random dithering compressors respectively. We choose $p = t \cdot r(Q)$ for $t \in \left\{3, 1, \frac{1}{3}, \frac{1}{9}\right\}$. Figure 1 shows that $p = r(Q)$ or $p = \frac{1}{3}r(Q)$ achieves the best performance, which coincides with our analysis in Section B.

A.3. Comparison to ECSGD and ECGD

In this subsection, we compare error compensated L-Katyusha (ECLK) with error compensated SGD (ECSGD) and error compensated GD (ECGD) for TopK compressor and random dithering compressor. ECGD is actually a special case of ECSGD with $m = 1$, where the full gradient $\nabla f^{(\tau)}(x^k)$ is calculated on each node. Let ECLK-full be the special case of ECLK with $m = 1$, where the full gradient $\nabla f^{(\tau)}(x^k)$ is calculated on each node. For ECLK, we choose $p = r(Q)$. Figure 2 and Figure 3 show that ECSGD and ECGD can only converge to a neighborhood of the optimal solution, while ECLK and ECLK-full converge to the optimal solution, and at a linear rate.

A.4. Comparison to ADIANA

ADIANA [17] is an accelerated method for any unbiased compressor where the full gradient is used on each node. In this subsection, we compare the EC-LKatyusha-full with ADIANA. For the unbiased compressor $\tilde{Q}$ for ADIANA, we use random dithering compressor with $s = 2$ and $s = \sqrt{d}$. For the contraction compressor, we use TopK compressor with $K = 1$ and $\frac{1}{\omega+1}\tilde{Q}$ where $\tilde{Q}$ is the random dithering compressor with $s = 2$ and $s = \sqrt{d}$. Figure 4 shows that for the communication complexity, the EC-LKatyusha-full with Top1 compressor is the best. For the random dithering compressor with $s = 2$ or $s = \sqrt{d}$, the communication complexity of EC-LKatyusha-full is also better than that of ADIANA.
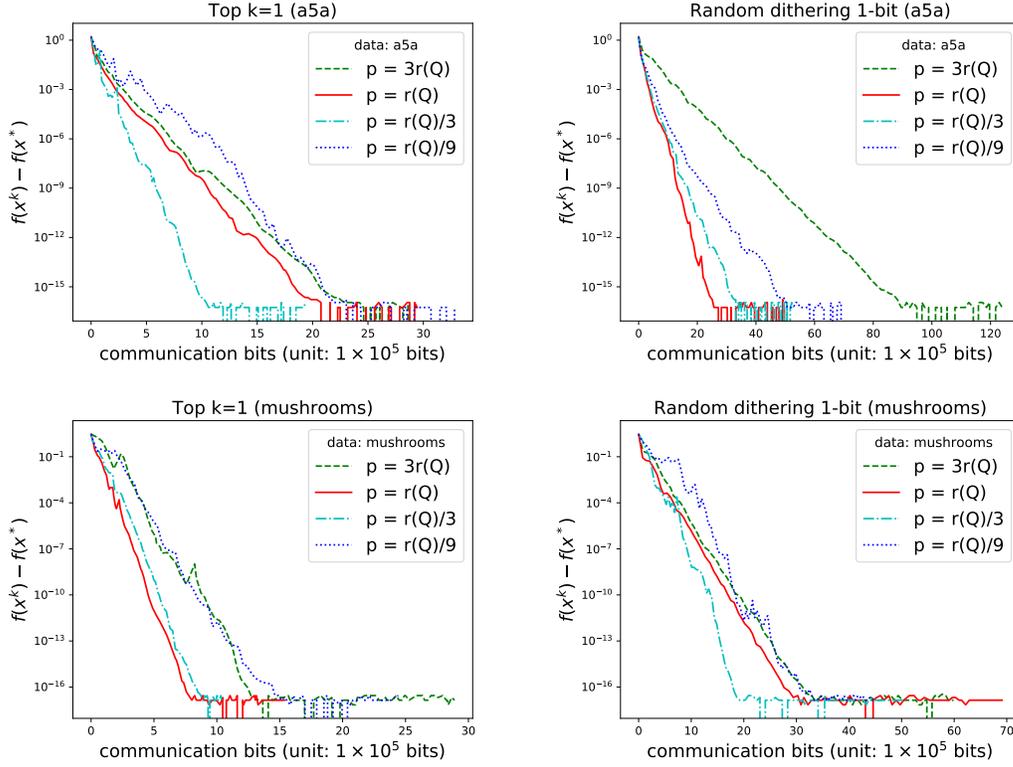
Figure 1: The influence of p for the communication complexity performance of Top k=1 and Random dithering 1-bit for the error compensated L-Katyusha on `a5a` and `mushrooms` datasets.
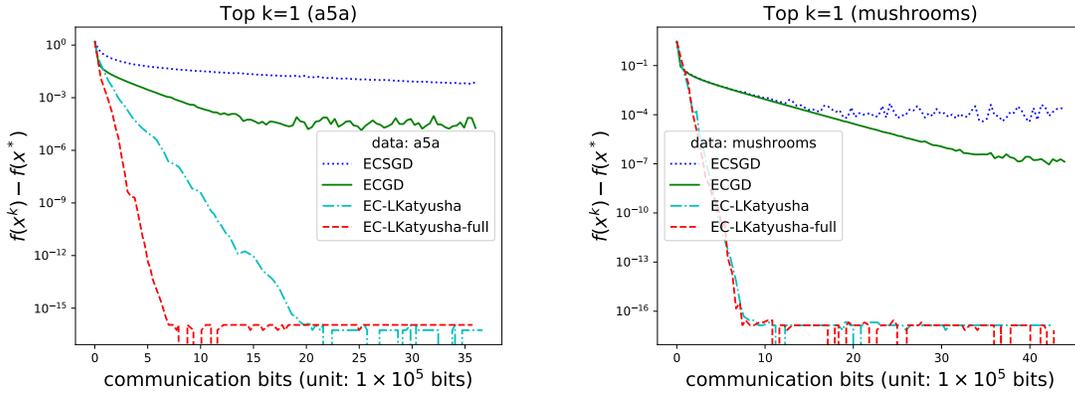
Figure 2: The communication complexity performance of ECSGD vs ECGD vs EC-LKatyusha vs EC-LKatyusha-full for Top k=1 on `a5a` and `mushrooms` datasets.

Appendix B. Communication Cost

Optimal choice of p. In Algorithm 1, when w^k is updated, the uncompressed vector $\nabla f^{(\tau)}(w^k)$ need to be communicated. We denote Δ_1 as the communication cost of the uncompressed vector $x \in \mathbb{R}^d$. Define the compress ratio
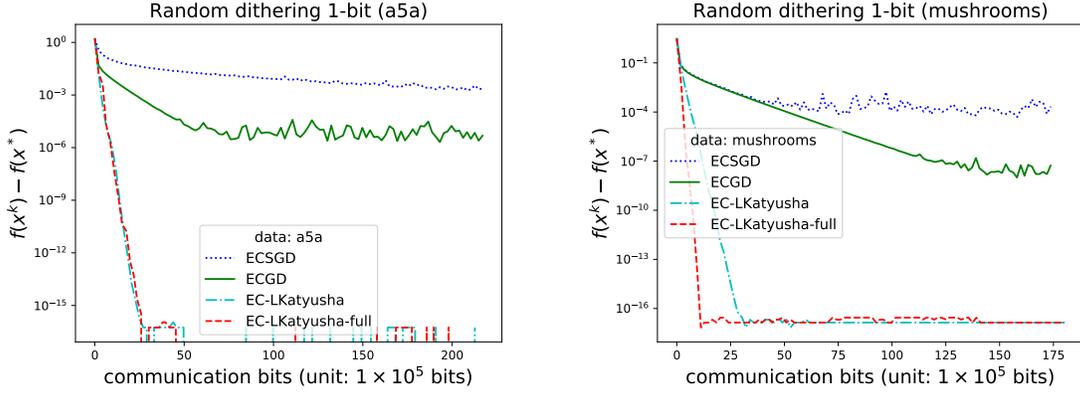
Figure 3: The communication complexity performance of ECSGD vs ECGD vs EC-LKatyusha vs EC-LKatyusha-full for Random dithering 1-bit on `a5a` and `mushrooms` datasets.
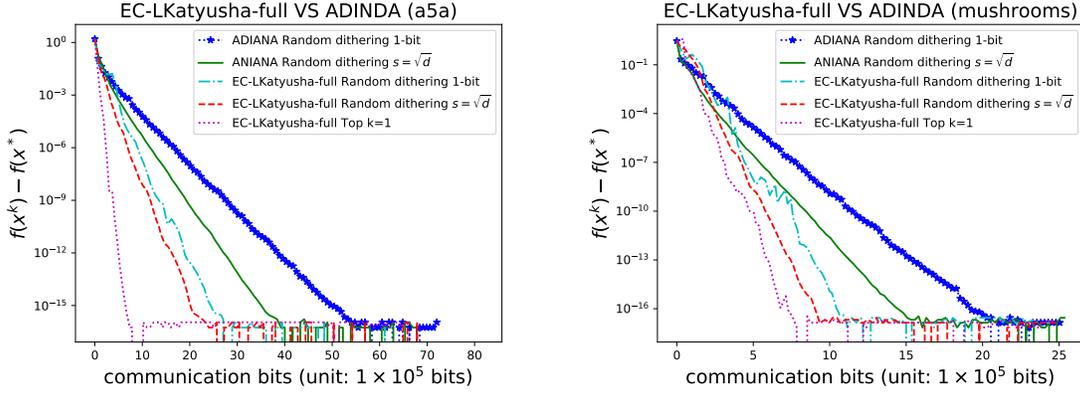
Figure 4: The communication complexity performance of EC-LKatyusha-full vs ADIANA on `a5a` and `mushrooms` datasets.

$r(Q)$ for the contraction compressor Q as

$$r(Q) := \sup_{x \in \mathbb{R}^d} \left\{ \mathbb{E} \left[\frac{\text{communication cost of } Q(x)}{\Delta_1} \right] \right\}. \tag{12}$$

Denote the total expected communication cost for k iterations as $\mathcal{T}_k$. The expected communication cost at iteration $k \geq 1$ is bounded by $\Delta_1 r(Q) + 1 + p\Delta_1$, where 1 bit is needed to communicate u_τ^{k+1}, and the expected communication cost at iteration $k = 0$ is bounded by $\Delta_1 r(Q) + 1 + \Delta_1$. Hence,

$$
\begin{aligned}
\mathcal{T}_k &\leq \Delta_1 r(Q) + 1 + \Delta_1 + (\Delta_1 r(Q) + 1 + p\Delta_1)k \\
&\leq \Delta_1 r(Q) + 1 + \Delta_1 + (\Delta_1 r(Q) + 1) \left(1 + \frac{p}{r(Q)} \right) k. \tag{13}
\end{aligned}
$$

Next, we discuss how to choose p to minimize the total expected communication cost. From Corollary 7 (i) and (13), we have $\mathbb{E}[\Phi^k] \leq \epsilon \Phi^0$ for

$$
\begin{aligned}
\mathcal{T}_k &= O\left((\Delta_1 r(Q) + 1) \left(1 + \frac{p}{r(Q)} \right) \left(a + \frac{1}{p} + \frac{b}{\sqrt{p}} \right) \log \frac{1}{\epsilon} \right) \\
&= O\left((\Delta_1 r(Q) + 1) \left(a + \frac{pa}{r(Q)} + \frac{1}{p} + \frac{1}{r(Q)} + \frac{b}{\sqrt{p}} + \frac{b\sqrt{p}}{r(Q)} \right) \log \frac{1}{\epsilon} \right),
\end{aligned}
$$

14
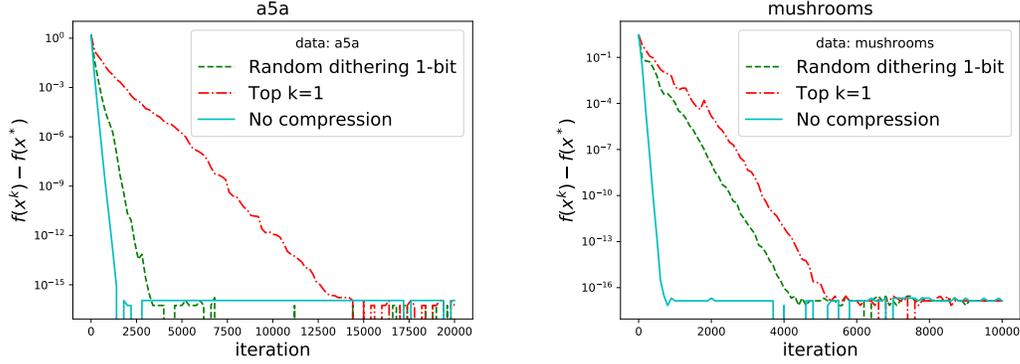
Figure 5: The iteration complexity performance of Top k=1 vs Random dithering 1-bit vs No compression for the error compensated L-Katyusha on `a5a` and `mushrooms` datasets.
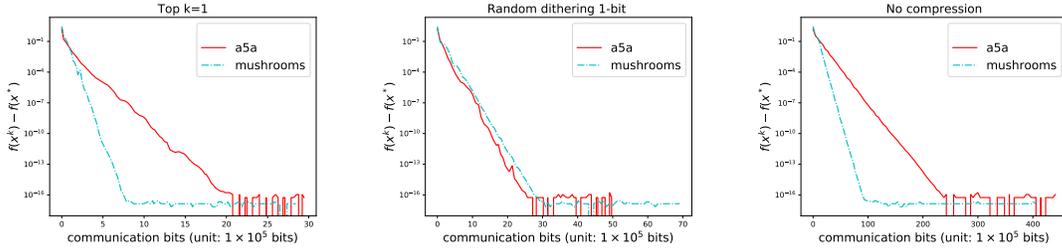
Figure 6: The communication complexity performance of TopK (with $K = 1$) vs Random dithering 1-bit vs No compression for the error compensated L-Katyusha on `a5a` and `mushrooms` datasets.

where we denote $a = \frac{1}{\delta} + \sqrt{\frac{L_f}{\mu}}$ and $b = \sqrt{\frac{L}{\mu n}} + \frac{1}{\delta}\sqrt{\frac{(1-\delta)\bar{L}}{\mu}} + \sqrt{\frac{(1-\delta)L}{\mu\delta}}$. Noticing that $\frac{b}{\sqrt{p}} + \frac{b\sqrt{p}}{r(Q)} \geq \frac{2b}{\sqrt{r(Q)}}$, we have

$$O\left(a + \frac{pa}{r(Q)} + \frac{1}{p} + \frac{1}{r(Q)} + \frac{b}{\sqrt{p}} + \frac{b\sqrt{p}}{r(Q)}\right) \geq O\left(a + \frac{1}{r(Q)} + \frac{b}{\sqrt{r(Q)}}\right),$$

and the above lower bound holds for $p = O(r(Q))$. Hence, in order to minimize the total expected communication cost, the optimal choice of p is $O(r(Q))$.

Under Assumption 2.1 or 2.2, from Corollary 7 (ii), by the same analysis, in order to minimize the total expected communication cost for $\mathbb{E}[\Psi^k] \leq \epsilon\Psi^0$, the optimal choice of p is also $O(r(Q))$.

Comparison to the uncompressed L-Katyusha. For simplicity, we assume $L_f = \bar{L} = L$ and $\Delta_1 r(Q) \geq O(1)$. From (11) and (13), by choosing $p = O(r(Q))$, we have

$$\mathcal{T}_k = O\left(\Delta_1\left(\frac{r(Q)}{\delta} + \left(r(Q) + \frac{\sqrt{r(Q)}}{\sqrt{n}} + \frac{\sqrt{(1-\delta)r(Q)}}{\delta}\right)\sqrt{\frac{L}{\mu}}\right)\log\frac{1}{\epsilon}\right). \tag{14}$$

For uncompressed L-Katyusha, by choosing $p = 1$, we have

$$\mathcal{T}_k = O\left(\Delta_1\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right). \tag{15}$$

If $\frac{\sqrt{r(Q)}}{\delta} < 1$, then the communication cost in (14) is less than that in (15). For TopK compressor, $r(Q) = \frac{K(64+\lceil\log d\rceil)}{64d}$, and in practice δ can be much larger than $\frac{K}{d}$, sometimes even in order $O(1)$.

Appendix C. Lemmas

We bound the varaince of g^k in the following lemma.

15

Lemma 8 *We have*

$$\mathbb{E}_k\left[\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2\right] \leq \frac{2L}{n}\left(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\right). \tag{16}$$

Proof

Since $f_i^{(\tau)}$ is L-smooth, we have

$$\|\nabla f_i^{(\tau)}(x) - \nabla f_i^{(\tau)}(y)\|^2 \leq 2L(f_i^{(\tau)}(x) - f_i^{(\tau)}(y) - \langle\nabla f_i^{(\tau)}(y), x - y\rangle),$$

for any $x, y \in \mathbb{R}^d$. Therefore,

$$
\begin{aligned}
\mathbb{E}_k\left[\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2\right] &= \mathbb{E}_k\|g^k\|^2 - \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{n}\sum_{\tau=1}^n g_\tau^k\right\|^2 - \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&= \frac{1}{n^2}\mathbb{E}_k\left\langle\sum_{\tau=1}^n g_\tau^k, \sum_{\tau=1}^n g_\tau^k\right\rangle - \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&= \frac{1}{n^2}\sum_{\tau_1,\tau_2=1}^n \mathbb{E}_k\left\langle g_{\tau_1}^k, g_{\tau_2}^k\right\rangle - \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\|g_\tau^k\|^2 - \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&\quad + \frac{1}{n^2}\sum_{\tau_1\neq\tau_2}\left\langle\nabla f^{(\tau_1)}(x^k) - \nabla f^{(\tau_1)}(w^k), \nabla f^{(\tau_2)}(x^k) - \nabla f^{(\tau_2)}(w^k)\right\rangle \\
&= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\|g_\tau^k\|^2 - \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \\
&\leq \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\|g_\tau^k\|^2 \\
&\leq \frac{2L}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\left(f_{i_k^\tau}^{(\tau)}(w^k) - f_{i_k^\tau}^{(\tau)}(x^k) - \langle\nabla f_{i_k^\tau}^{(\tau)}(x^k), w^k - x^k\rangle\right) \\
&= \frac{2L}{n}\left(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\right).
\end{aligned}
$$

■

Lemma 9 *If $\mathcal{L}_1 \geq L_f$, then we have*

$$\frac{\mathcal{L}_1}{4\eta}\|z^{k+1} - z^k\|^2 + \langle g^k + \nabla f(w^k), z^{k+1} - z^k\rangle \geq \frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) - \frac{1}{\mathcal{L}_1\theta_1}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2. \tag{17}$$

Proof

16

Since $z^{k+1} - z^k = \frac{1}{\theta_1}(y^{k+1} - x^k)$, we have

$$
\begin{aligned}
\frac{\mathcal{L}_1}{4\eta}\|z^{k+1} - z^k\|^2 + \langle g^k + \nabla f(w^k), z^{k+1} - z^k\rangle &= \frac{\mathcal{L}_1}{4\eta\theta_1^2}\|y^{k+1} - x^k\|^2 + \frac{1}{\theta_1}\langle g^k + \nabla f(w^k), y^{k+1} - x^k\rangle \\
&= \frac{1}{\theta_1}\langle\nabla f(x^k), y^{k+1} - x^k\rangle + \frac{3\mathcal{L}_1}{4\theta_1}\|y^{k+1} - x^k\|^2 \\
&\quad + \frac{1}{\theta_1}\langle g^k + \nabla f(w^k) - \nabla f(x^k), y^{k+1} - x^k\rangle \\
&\geq \frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) + \left(\frac{3\mathcal{L}_1}{4\theta_1} - \frac{L_f}{2\theta_1}\right)\|y^{k+1} - x^k\|^2 \\
&\quad + \frac{1}{\theta_1}\langle g^k + \nabla f(w^k) - \nabla f(x^k), y^{k+1} - x^k\rangle \\
&\geq \frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) + \frac{\mathcal{L}_1}{4\theta_1}\|y^{k+1} - x^k\|^2 \\
&\quad + \frac{1}{\theta_1}\langle g^k + \nabla f(w^k) - \nabla f(x^k), y^{k+1} - x^k\rangle \\
&\geq \frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) - \frac{1}{\mathcal{L}_1\theta_1}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2,
\end{aligned}
$$

where the first inequality comes from L_f-smoothness of f, and the last inequality comes from Young's inequality. ∎

Lemma 10 *We have*

$$
\begin{aligned}
\langle g^k + \nabla f(w^k), x^* - z^{k+1}\rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 &\geq \frac{\mathcal{L}_1}{4\eta}\|z^k - z^{k+1}\|^2 + \tilde{\mathcal{Z}}^{k+1} - \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 \\
&\quad - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\|e^{k+1}\|^2 + \psi(z^{k+1}) - \psi(x^*). \qquad (18)
\end{aligned}
$$

Proof

First, from (7) and $\sigma_1 = \frac{\mu_f}{2\mathcal{L}_1}$, we have

$$
\begin{aligned}
g^k + \nabla f(w^k) &= \frac{\mathcal{L}_1}{\eta}(\tilde{z}^k - \tilde{z}^{k+1}) + \mathcal{L}_1\sigma_1(\tilde{x}^k - \tilde{z}^{k+1}) - \partial\psi(z^{k+1}) \\
&= \frac{\mathcal{L}_1}{\eta}(\tilde{z}^k - \tilde{z}^{k+1}) + \frac{\mu_f}{2}(\tilde{x}^k - \tilde{z}^{k+1}) - \partial\psi(z^{k+1}),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\langle g^k + \nabla f(w^k), z^{k+1} - x^* \rangle &= \frac{\mu_f}{2}\langle z^{k+1} - x^*, \tilde{x}^k - \tilde{z}^{k+1}\rangle + \frac{\mathcal{L}_1}{\eta}\langle z^{k+1} - x^*, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&\quad - \langle z^{k+1} - x^*, \partial\psi(z^{k+1})\rangle \\
&\leq \frac{\mu_f}{2}\langle z^{k+1} - x^*, \tilde{x}^k - \tilde{z}^{k+1}\rangle + \frac{\mathcal{L}_1}{\eta}\langle z^{k+1} - x^*, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&\quad + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1} - x^*\|^2 \\
&= \frac{\mu_f}{2}\langle \tilde{z}^{k+1} - x^*, \tilde{x}^k - \tilde{z}^{k+1}\rangle + \frac{\mathcal{L}_1}{\eta}\langle \tilde{z}^{k+1} - x^*, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&\quad + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1} - x^*\|^2 \\
&\quad + \frac{\mu_f}{2}\langle z^{k+1} - \tilde{z}^{k+1}, \tilde{x}^k - \tilde{z}^{k+1}\rangle + \frac{\mathcal{L}_1}{\eta}\langle z^{k+1} - \tilde{z}^{k+1}, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&= \frac{\mu_f}{4}\left(\|\tilde{x}^k - x^*\|^2 - \|\tilde{z}^{k+1} - x^*\|^2 - \|\tilde{x}^k - \tilde{z}^{k+1}\|^2\right) \\
&\quad + \frac{\mathcal{L}_1}{2\eta}\left(\|\tilde{z}^k - x^*\|^2 - \|\tilde{z}^{k+1} - x^*\|^2 - \|\tilde{z}^k - \tilde{z}^{k+1}\|^2\right) \\
&\quad + \frac{\mu_f}{4}\left(\|z^{k+1} - \tilde{z}^{k+1}\|^2 + \|\tilde{x}^k - \tilde{z}^{k+1}\|^2 - \|\tilde{x}^k - z^{k+1}\|^2\right) \\
&\quad + \frac{\mathcal{L}_1}{2\eta}\left(\|z^{k+1} - \tilde{z}^{k+1}\|^2 + \|\tilde{z}^k - \tilde{z}^{k+1}\|^2 - \|\tilde{z}^k - z^{k+1}\|^2\right) \\
&\quad + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1} - x^*\|^2 \\
&\leq -\left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{4}\right)\|\tilde{z}^{k+1} - x^*\|^2 + \frac{\mathcal{L}_1}{2\eta}\|\tilde{z}^k - x^*\|^2 + \frac{\mu_f}{4}\|\tilde{x}^k - x^*\|^2 \\
&\quad + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{4}\right)\|z^{k+1} - \tilde{z}^{k+1}\|^2 - \frac{\mathcal{L}_1}{2\eta}\|\tilde{z}^k - z^{k+1}\|^2 \\
&\quad + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1} - x^*\|^2.
\end{aligned}
$$

For $\|\tilde{x}^k - x^*\|^2$, $\|\tilde{z}^k - z^{k+1}\|^2$, and $\|z^{k+1} - x^*\|^2$, from Young's inequality, we have

$$
\|\tilde{x}^k - x^*\|^2 \leq 2\|\tilde{x}^k - x^k\|^2 + 2\|x^k - x^*\|^2, \qquad \|\tilde{z}^k - z^{k+1}\|^2 \geq \frac{1}{2}\|z^k - z^{k+1}\|^2 - \|z^k - \tilde{z}^k\|^2,
$$

and

$$
\|z^{k+1} - x^*\|^2 \geq \frac{1}{2}\|\tilde{z}^{k+1} - x^*\|^2 - \|z^{k+1} - \tilde{z}^{k+1}\|^2.
$$

Hence, we arrive at

$$
\begin{aligned}
\langle g^k + \nabla f(w^k), z^{k+1} - x^* \rangle &\leq -\left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{4} + \frac{\mu_\psi}{4}\right)\|\tilde{z}^{k+1} - x^*\|^2 + \frac{\mathcal{L}_1}{2\eta}\|\tilde{z}^k - x^*\|^2 + \frac{\mu_f}{2}\|x^k - x^*\|^2 \\
&\quad + \frac{\mu_f}{2}\|\tilde{x}^k - x^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{4} + \frac{\mu_\psi}{2}\right)\|z^{k+1} - \tilde{z}^{k+1}\|^2 \\
&\quad - \frac{\mathcal{L}_1}{4\eta}\|z^k - z^{k+1}\|^2 + \frac{\mathcal{L}_1}{2\eta}\|z^k - \tilde{z}^k\|^2 + \psi(x^*) - \psi(z^{k+1}) \\
&= -\left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{4}\right)\|\tilde{z}^{k+1} - x^*\|^2 + \frac{\mathcal{L}_1}{2\eta}\|\tilde{z}^k - x^*\|^2 + \frac{\mu_f}{2}\|x^k - x^*\|^2 \\
&\quad + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\|e^{k+1}\|^2 \\
&\quad - \frac{\mathcal{L}_1}{4\eta}\|z^k - z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}) \\
&= -\tilde{\mathcal{Z}}^{k+1} + \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 \\
&\quad + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\|e^{k+1}\|^2 - \frac{\mathcal{L}_1}{4\eta}\|z^k - z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}).
\end{aligned}
$$

$\blacksquare$

Appendix D. Proofs of Lemmas 2, 3, and 4

D.1. Proof of Lemma 2

Since $\theta_1 + \theta_2 \leq 1$, and f is μ_f-strong convex, we have

$$
\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 \\
&= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k + z^k - x^k \rangle \\
&= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k \rangle + \frac{\theta_2}{\theta_1}\langle \nabla f(x^k), x^k - w^k \rangle + \frac{1 - \theta_1 - \theta_2}{\theta_1}\langle \nabla f(x^k), x^k - y^k \rangle \\
&\geq f(x^k) + \frac{\theta_2}{\theta_1}\langle \nabla f(x^k), x^k - w^k \rangle + \frac{1 - \theta_1 - \theta_2}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathbb{E}_k\left[\frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle g^k + \nabla f(w^k), x^* - z^{k+1} \rangle + \langle g^k + \nabla f(w^k), z^{k+1} - z^k \rangle\right],
\end{aligned}
$$

where the last inequality follows from the convexity of f and $\mathbb{E}_k[g^k + \nabla f(w^k)] = \nabla f(x^k)$. For the last term in the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}_k\left[\frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle g^k + \nabla f(w^k), x^* - z^{k+1} \rangle + \langle g^k + \nabla f(w^k), z^{k+1} - z^k \rangle - \psi(z^{k+1}) + \psi(x^*) - \tilde{\mathcal{Z}}^{k+1}\right] \\
&\overset{(18)}{\geq} -\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + \mathbb{E}_k\left[\langle g^k + \nabla f(w^k), z^{k+1} - z^k \rangle + \frac{\mathcal{L}_1}{4\eta}\|z^k - z^{k+1}\|^2\right] \\
&\quad - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2 \\
&\overset{(17)}{\geq} -\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2 \\
&\quad + \mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1}) - f(x^k)) - \frac{1}{\mathcal{L}_1\theta_1}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2\right] \\
&\overset{(16)}{\geq} -\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 - \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2 \\
&\quad + \mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1}) - f(x^k)) - \frac{2L}{n\mathcal{L}_1\theta_1}(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle)\right].
\end{aligned}
$$

Therefore,

$$\mathbb{E}_k\left[f(x^*)-\psi(z^{k+1})+\psi(x^*)-\tilde{\mathcal{Z}}^{k+1}\right]+\left(\frac{\mathcal{L}_1}{2\eta}+\frac{\mu_f}{2}\right)\|e^k\|^2+\left(\frac{\mathcal{L}_1}{2\eta}+\frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2$$

$$\geq \quad -\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1+\eta\mu/2}-\frac{1-\theta_1-\theta_2}{\theta_1}f(y^k)+\frac{1}{\theta_1}\mathbb{E}_k[f(y^{k+1})]-\frac{\theta_2}{\theta_1}\left(f(x^k)+\langle\nabla f(x^k),w^k-x^k\rangle\right)$$

$$-\frac{2L}{n\mathcal{L}_1\theta_1}(f(w^k)-f(x^k)-\langle\nabla f(x^k),w^k-x^k\rangle)$$

$$= \quad -\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1+\eta\mu/2}-\frac{1-\theta_1-\theta_2}{\theta_1}f(y^k)+\frac{1}{\theta_1}\mathbb{E}_k[f(y^{k+1})]-\frac{\theta_2}{\theta_1}f(w^k)$$

$$+\frac{1}{\theta_1}\left(\theta_2-\frac{2L}{n\mathcal{L}_1}\right)(f(w^k)-f(x^k)-\langle\nabla f(x^k),w^k-x^k\rangle).$$

From the convexity of ψ, and

$$y^{k+1}=x^k+\theta_1(z^{k+1}-z^k)=\theta_1 z^{k+1}+\theta_2 w^k+(1-\theta_1-\theta_2)y^k,$$

we have

$$\psi(z^{k+1})\geq\frac{1}{\theta_1}\psi(y^{k+1})-\frac{\theta_2}{\theta_1}\psi(w^k)-\frac{1-\theta_1-\theta_2}{\theta_1}\psi(y^k).$$

Hence, we can obtain

$$P(x^*)+\left(\frac{\mathcal{L}_1}{2\eta}+\frac{\mu_f}{2}\right)\|e^k\|^2+\left(\frac{\mathcal{L}_1}{2\eta}+\frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2 \quad \geq \quad \mathbb{E}_k[\tilde{\mathcal{Z}}^{k+1}]-\frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1+\eta\mu/2}-\frac{1-\theta_1-\theta_2}{\theta_1}P(y^k)$$

$$+\frac{1}{\theta_1}\mathbb{E}_k[P(y^{k+1})]-\frac{\theta_2}{\theta_1}P(w^k)$$

$$+\frac{1}{\theta_1}\left(\theta_2-\frac{2L}{n\mathcal{L}_1}\right)(f(w^k)-f(x^k)-\langle\nabla f(x^k),w^k-x^k\rangle).$$

After rearranging we can get the result.

D.2. Proof of Lemma 3

First, we have

$$\mathbb{E}_k[\|e_\tau^{k+1}\|^2] \quad = \quad \mathbb{E}_k\left\|e_\tau^k+\frac{\eta}{\mathcal{L}_1}g_\tau^k-\tilde{g}_\tau^k\right\|^2$$

$$\leq \quad (1-\delta)\mathbb{E}_k\left\|e_\tau^k+\frac{\eta}{\mathcal{L}_1}g_\tau^k\right\|^2$$

$$= \quad (1-\delta)\left\|e_\tau^k+\frac{\eta}{\mathcal{L}_1}(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\right\|^2+\frac{\eta^2}{\mathcal{L}_1^2}(1-\delta)\mathbb{E}_k\|g_\tau^k-(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2$$

$$\leq \quad (1-\delta)(1+\beta)\|e_\tau^k\|^2+(1-\delta)\left(1+\frac{1}{\beta}\right)\frac{\eta^2}{\mathcal{L}_1^2}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2$$

$$+\frac{\eta^2}{\mathcal{L}_1^2}(1-\delta)\mathbb{E}_k\|g_\tau^k-(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2$$

$$\leq \quad \left(1-\frac{\delta}{2}\right)\|e_\tau^k\|^2+\frac{2(1-\delta)\eta^2}{\delta\mathcal{L}_1^2}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2$$

$$+\frac{\eta^2}{\mathcal{L}_1^2}(1-\delta)\mathbb{E}_k\|g_\tau^k-(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2,$$

where we choose $\beta=\frac{\delta}{2(1-\delta)}$ when $\delta<1$. When $\delta=1$, it is easy to see the above inequality also holds. Since

$$\mathbb{E}_k\|g_\tau^k-(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2 \quad \leq \quad \mathbb{E}_k\|g_\tau^k\|^2$$

$$\leq \quad 2L\mathbb{E}_k(f_{i_k^\tau}^{(\tau)}(w^k)-f_{i_k^\tau}^{(\tau)}(x^k)-\langle\nabla f_{i_k^\tau}^{(\tau)}(x^k),w^k-x^k\rangle)$$

$$= \quad 2L(f^{(\tau)}(w^k)-f^{(\tau)}(x^k)-\langle\nabla f^{(\tau)}(x^k),w^k-x^k\rangle),$$

and
$$\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \le 2\bar{L}(f^{(\tau)}(w^k) - f^{(\tau)}(x^k) - \langle \nabla f^{(\tau)}(x^k), w^k - x^k \rangle),$$

we arrive at

$$\mathbb{E}_k[\|e_\tau^{k+1}\|^2] \le \left(1 - \frac{\delta}{2}\right)\|e_\tau^k\|^2 + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2\bar{L}}{\delta} + L\right)(f^{(\tau)}(w^k) - f^{(\tau)}(x^k) - \langle \nabla f^{(\tau)}(x^k), w^k - x^k \rangle).$$

Therefore,

$$
\begin{aligned}
\mathbb{E}_k\left[\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^{k+1}\|^2\right] &\le \left(1 - \frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2 \\
&\quad + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2\bar{L}}{\delta} + L\right)\frac{1}{n}\sum_{\tau=1}^n (f^{(\tau)}(w^k) - f^{(\tau)}(x^k) - \langle \nabla f^{(\tau)}(x^k), w^k - x^k \rangle) \\
&= \left(1 - \frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2 + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2\bar{L}}{\delta} + L\right)(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).
\end{aligned}
$$

D.3. Proof of Lemma 4

Under Assumption 2.1, we have $\mathbb{E}[Q(x)] = \delta x$, and

$$
\begin{aligned}
\mathbb{E}_k\|e^{k+1}\|^2 &= \mathbb{E}_k\left\|\frac{1}{n}\sum_{\tau=1}^n e_\tau^{k+1}\right\|^2 \\
&= \frac{1}{n^2}\sum_{i,j}\mathbb{E}_k\langle e_i^{k+1}, e_j^{k+1}\rangle \\
&= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\|e_\tau^{k+1}\|^2 + \frac{1}{n^2}\sum_{i\ne j}\mathbb{E}_k\langle e_i^{k+1}, e_j^{k+1}\rangle \\
&\le \frac{1-\delta}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\left\|e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k\right\|^2 + \frac{(1-\delta)^2}{n^2}\sum_{i\ne j}\mathbb{E}_k\left\langle e_i^k + \frac{\eta}{\mathcal{L}_1}g_i^k, e_j^k + \frac{\eta}{\mathcal{L}_1}g_j^k\right\rangle \\
&= \frac{(1-\delta)^2}{n^2}\mathbb{E}_k\left\|\sum_{\tau=1}^n (e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k)\right\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\left\|e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k\right\|^2 \\
&\le (1-\delta)\mathbb{E}_k\left\|e^k + \frac{\eta}{\mathcal{L}_1}g^k\right\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\left\|e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k\right\|^2.
\end{aligned}
$$

Under Assumption 2.2, we have

$$
\begin{aligned}
\mathbb{E}_k\|e^{k+1}\|^2 &= \mathbb{E}_k\left\|\frac{1}{n}\sum_{\tau=1}^n e_\tau^{k+1}\right\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{n}\sum_{\tau=1}^n \left(e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k - Q\left(\frac{\eta}{\mathcal{L}_1}g_\tau^k + e_\tau^k\right)\right)\right\|^2 \\
&\overset{Assumption 2.2}{\le} (1-\delta')\mathbb{E}_k\left\|e^k + \frac{\eta}{\mathcal{L}_1}g^k\right\|^2 \\
&\le (1-\delta)\mathbb{E}_k\left\|e^k + \frac{\eta}{\mathcal{L}_1}g^k\right\|^2.
\end{aligned}
$$

Overall, under Assumption 2.1 or Assumption 2.2, we have

$$
\begin{aligned}
\mathbb{E}_k\|e^{k+1}\|^2 &\le (1-\delta)\mathbb{E}_k\left\|e^k + \frac{\eta}{\mathcal{L}_1}g^k\right\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \mathbb{E}_k\left\|e_\tau^k + \frac{\eta}{\mathcal{L}_1}g_\tau^k\right\|^2 \\
&\le (1-\delta)\mathbb{E}_k\left\|e^k + \frac{\eta}{\mathcal{L}_1}g^k\right\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n \|e_\tau^k\|^2 + \frac{2(1-\delta)\delta\eta^2}{n^2\mathcal{L}_1^2}\sum_{\tau=1}^n \mathbb{E}_k\|g_\tau^k\|^2.
\end{aligned}
$$

The first term on the right hand side above can be bounded as

$$
\begin{aligned}
(1-\delta)\mathbb{E}_k \left\| e^k + \frac{\eta}{\mathcal{L}_1} g^k \right\|^2
&= (1-\delta)\mathbb{E}_k \left\| e^k + \frac{\eta}{\mathcal{L}_1}(\nabla f(x^k) - \nabla f(w^k)) + \frac{\eta}{\mathcal{L}_1} g^k - \frac{\eta}{\mathcal{L}_1}(\nabla f(x^k) - \nabla f(w^k)) \right\|^2 \\
&= (1-\delta)\mathbb{E}_k \left\| e^k + \frac{\eta}{\mathcal{L}_1}(\nabla f(x^k) - \nabla f(w^k)) \right\|^2 + (1-\delta)\frac{\eta^2}{\mathcal{L}_1^2}\mathbb{E}_k \| g^k - (\nabla f(x^k) - \nabla f(w^k)) \|^2 \\
&\leq \left(1 - \frac{\delta}{2}\right) \|e^k\|^2 + \frac{2(1-\delta)\eta^2}{\delta \mathcal{L}_1^2} \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&\quad + (1-\delta)\frac{\eta^2}{\mathcal{L}_1^2}\mathbb{E}_k \| g^k - (\nabla f(x^k) - \nabla f(w^k)) \|^2 \\
&\overset{(16)}{\leq} \left(1 - \frac{\delta}{2}\right) \|e^k\|^2 + \frac{2(1-\delta)\eta^2}{\delta \mathcal{L}_1^2} \|\nabla f(x^k) - \nabla f(w^k)\|^2 \\
&\quad + (1-\delta)\frac{2L\eta^2}{n\mathcal{L}_1^2}(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle) \\
&\leq \left(1 - \frac{\delta}{2}\right) \|e^k\|^2 + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2L_f}{\delta} + \frac{L}{n}\right)(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}_k \|g_\tau^k\|^2 &\leq \frac{2L}{n}\sum_{\tau=1}^{n}\mathbb{E}_k(f_{i_k^\tau}^{(\tau)}(w^k) - f_{i_k^\tau}^{(\tau)}(x^k) - \langle \nabla f_{i_k^\tau}^{(\tau)}(x^k), w^k - x^k \rangle) \\
&= 2L(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathbb{E}_k \|e^{k+1}\|^2 &\leq \left(1 - \frac{\delta}{2}\right)\|e^k\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 \\
&\quad + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2L_f}{\delta} + \frac{L}{n} + \frac{2L\delta}{n}\right)(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle) \\
&\leq \left(1 - \frac{\delta}{2}\right)\|e^k\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 \\
&\quad + \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2L_f}{\delta} + \frac{3L}{n}\right)(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k \rangle).
\end{aligned}
$$

Appendix E. Proof of Theorem 5

From $\|e^k\|^2 \leq \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$, Equation (8), and Lemma 2, we can obtain

$$
\begin{aligned}
\mathbb{E}_k\left[\tilde{\mathcal{Z}}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1}\right] \quad \leq \quad & \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k \\
& + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\mathbb{E}_k\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^{k+1}\|^2 \\
& - \frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1}\right)(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle) \\
\overset{\text{Lemma 3}}{\leq} \quad & \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k \\
& + \left(\frac{\mathcal{L}_1}{\eta} + \mu\right)\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 \\
& - \left(\frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1}\right) - \frac{2(1-\delta)\eta^2}{\mathcal{L}_1^2}\left(\frac{2\bar{L}}{\delta} + L\right)\left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\right) \\
& \cdot(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle) \\
\leq \quad & \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k \\
& + \frac{4\mathcal{L}_1}{3\eta}\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 - \frac{1}{\theta_1}\left(\theta_2 - \frac{1}{\mathcal{L}_1}\left(\frac{2L}{n} + \frac{8(1-\delta)\bar{L}}{9\delta} + \frac{4(1-\delta)L}{9}\right)\right) \\
& \cdot(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle).
\end{aligned}
$$

Therefore, from Lemma 3, we have

$$
\begin{aligned}
& \mathbb{E}_k\left[\tilde{\mathcal{Z}}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1} + \frac{4\mathcal{L}_1}{\delta\eta}\cdot\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^{k+1}\|^2\right] \\
\leq \quad & \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k \\
& + \left(1 - \frac{\delta}{6}\right)\frac{4\mathcal{L}_1}{\delta\eta}\cdot\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 - \frac{1}{\theta_1}\left(\theta_2 - \frac{1}{\mathcal{L}_1}\left(\frac{2L}{n} + \frac{56(1-\delta)\bar{L}}{9\delta^2} + \frac{28(1-\delta)L}{9\delta}\right)\right) \\
& \cdot(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle) \\
= \quad & \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k + \left(1 - \frac{\delta}{6}\right)\frac{4\mathcal{L}_1}{\delta\eta}\cdot\frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2 \\
& - \frac{1}{\theta_1}\left(\theta_2 - \frac{\mathcal{L}_2}{2\mathcal{L}_1}\right)(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle).
\end{aligned}
$$

When $\theta_2 \geq \frac{\mathcal{L}_2}{2\mathcal{L}_1}$ we can get the result.

Appendix F. Proof of Theorem 6

From Lemma 2 and (8), we have

$$\mathbb{E}_k\left[\tilde{\mathcal{Z}}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1}\right]$$

$$\leq \quad \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + \left(1 - \theta_1 - \theta_2 + \frac{\theta_2}{q}\right)\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k$$

$$+ \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu_f}{2}\right)\|e^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\mathbb{E}_k\|e^{k+1}\|^2$$

$$- \frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1}\right)\left(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\right)$$

$$\overset{\text{Lemma } 4}{\leq} \quad \frac{\mathcal{L}_1\tilde{\mathcal{Z}}^k}{\mathcal{L}_1 + \eta\mu/2} + \left(1 - \theta_1 - \theta_2 + \frac{\theta_2}{q}\right)\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k$$

$$+ \left(\frac{\mathcal{L}_1}{\eta} + \mu\right)\|e^k\|^2 + \left(\frac{\mathcal{L}_1}{2\eta} + \frac{\mu}{2}\right)\frac{2(1-\delta)\delta}{n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$$

$$- \left(\frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1}\right) - \frac{4(1-\delta)\eta}{3\mathcal{L}_1}\left(\frac{2L_f}{\delta} + \frac{3L}{n}\right)\right)\left(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\right)$$

$$\leq \quad \frac{6\theta_1\mathcal{L}_1\tilde{\mathcal{Z}}^k}{6\theta_1\mathcal{L}_1 + \mu} + \left(1 - \theta_1 - \theta_2 + \frac{\theta_2}{q}\right)\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k + \frac{4\mathcal{L}_1}{3\eta}\|e^k\|^2$$

$$+ \frac{4\mathcal{L}_1(1-\delta)\delta}{3\eta n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$$

$$- \frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1} - \frac{4(1-\delta)}{9\mathcal{L}_1}\left(\frac{2L_f}{\delta} + \frac{3L}{n}\right)\right)\left(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle\right).$$

Therefore, from Lemma 3 and Lemma 4, we can get

$$\mathbb{E}_k\left[\tilde{\mathcal{Z}}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1} + \frac{4\mathcal{L}_1}{\delta\eta}\|e^{k+1}\|^2 + \frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^{k+1}\|^2\right]$$

$$\leq \quad \frac{6\theta_1\mathcal{L}_1\tilde{\mathcal{Z}}^k}{6\theta_1\mathcal{L}_1 + \mu} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k$$

$$+ \left(1 - \frac{\delta}{6}\right)\frac{4\mathcal{L}_1}{\delta\eta}\|e^k\|^2 + \left(1 - \frac{\delta}{6}\right)\frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$$

$$- \frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1} - \frac{28(1-\delta)}{9\delta\mathcal{L}_1}\left(\frac{2L_f}{\delta} + \frac{3L}{n}\right) - \frac{56(1-\delta)}{3\delta\mathcal{L}_1}\left(\frac{2\bar{L}}{\delta n} + \frac{L}{n}\right)\right)$$

$$\cdot(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle)$$

$$\leq \quad \frac{6\theta_1\mathcal{L}_1\tilde{\mathcal{Z}}^k}{6\theta_1\mathcal{L}_1 + \mu} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k$$

$$+ \left(1 - \frac{\delta}{6}\right)\frac{4\mathcal{L}_1}{\delta\eta}\|e^k\|^2 + \left(1 - \frac{\delta}{6}\right)\frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$$

$$- \frac{1}{\theta_1}\left(\theta_2 - \frac{2L}{n\mathcal{L}_1} - \frac{392(1-\delta)L_f}{9\delta^2\mathcal{L}_1} - \frac{28(1-\delta)L}{\delta\mathcal{L}_1 n}\right)$$

$$\cdot(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle)$$

$$= \quad \frac{6\theta_1\mathcal{L}_1\tilde{\mathcal{Z}}^k}{6\theta_1\mathcal{L}_1 + \mu} + (1 - \theta_1 - \theta_2 + \frac{\theta_2}{q})\mathcal{Y}^k + (1 - p + pq)\mathcal{W}^k$$

$$+ \left(1 - \frac{\delta}{6}\right)\frac{4\mathcal{L}_1}{\delta\eta}\|e^k\|^2 + \left(1 - \frac{\delta}{6}\right)\frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n} \cdot \frac{1}{n}\sum_{\tau=1}^{n}\|e_\tau^k\|^2$$

$$- \frac{1}{\theta_1}\left(\theta_2 - \frac{\mathcal{L}_3}{2\mathcal{L}_1}\right)(f(w^k) - f(x^k) - \langle\nabla f(x^k), w^k - x^k\rangle),$$

where we use $\bar{L} \leq nL_f$ in the second inequality. When $\theta_2 \geq \frac{\mathcal{L}_3}{2\mathcal{L}_1}$ we can get the result.

Appendix G. Proof of Corollary 7

(i) First, we have $\frac{1}{2} \geq \theta_2 \geq \frac{\mathcal{L}_2}{2\mathcal{L}_1}$. Form the definition of θ_1, we know $\theta_1 \leq \frac{1}{2}$. Hence $\theta_1 + \theta_2 \leq 1$. Next we discuss two cases:

– **Case 1:** $3\mu\eta < \mathcal{L}_1$**.** In this case, we have $\mathcal{L}_1 = \max\{\mathcal{L}_4, L_f\}$. Then from Theorem 5 and same as the proof of Theorem 3.2 in [22], we have $\mathbb{E}[\Phi^k] \leq \epsilon\Phi^0$ as long as

$$k \geq O\left(\frac{1}{\delta} + \frac{1}{p} + \sqrt{\frac{L_f}{\mu}} + \sqrt{\frac{\mathcal{L}_4}{\mu p}}\right).$$

Since $\mathcal{L}_4 = \mathcal{L}_2$, we can get the result.

– **Case 2:** $3\mu\eta = \mathcal{L}_1$**.** In this case, we have

$$\frac{\mu}{\mu + 6\theta_1\mathcal{L}_1} = \frac{\mu}{\mu + 6\mu} = \frac{1}{7} \geq \frac{p}{7}.$$

Hence, from Theorem 5 and same as the proof of Theorem 3.2 in [22], we also have $\mathbb{E}[\Phi^k] \leq \epsilon\Phi^0$ for

$$k \geq O\left(\frac{1}{\delta} + \frac{1}{p} + \sqrt{\frac{L_f}{\mu}} + \sqrt{\frac{\mathcal{L}_4}{\mu p}}\right).$$

Since $\mathcal{L}_4 = \mathcal{L}_2$, we can get the result.

(ii) By using Theorem 6, same as (i), we can get the result.