OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Error Compensated Loopless SVRG for Distributed Optimization

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

Hanze Dong HDONGAJ@UST.HK
Hong Kong University of Science and Technology

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

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

Abstract

A key bottleneck in distributed training of large scale machine learning models is the overhead related to communication of gradients. In order to reduce the communicated cost, gradient compression (e.g., sparsification and quantization) and error compensation techniques are often used. In this paper, we propose and study a new efficient method in this space: error compensated loopless SVRG method (L-SVRG). Our method is capable of working with any contraction compressor (e.g., TopK compressor), and we perform analysis for strongly convex optimization problems in the composite case and smooth case. We prove linear convergence rates for both cases and show that in the smooth case the rate has a better dependence on the contraction factor associated with the compressor. Further, we show that in the smooth case, and under some certain conditions, error compensated L-SVRG has the same convergence rate as the vanilla L-SVRG method. Numerical experiments are presented to illustrate the efficiency of our method.

1. Introduction

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

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

where $f(x) := \frac{1}{n}\sum_{\tau} f^{(\tau)}(x)$ is an average of n smooth 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 $f^{(\tau)}(x) = \frac{1}{m}\sum_{i=1}^{m} f_i^{(\tau)}(x)$, representing the average loss over the training data stored on node τ. We assume that problem (1) has at least one optimal solution x^*.

For large scale machine learning problems, distributed training and parallel training are often used. While in such settings, communication is generally much slower than the computation, which make the communication overhead become a key bottleneck. There are several ways to tackle this issue, such as using large mini-batchs [5, 20], asynchronous learning [1, 9, 12], quantization and error compensation [2, 4, 10, 13, 18]. For quantization, there are mainly two types, i.e., contraction compressor and unbiased compressor, which are defined as follows.

Q is a contraction compressor if there is a $0 < \delta \leq 1$ such that

$$\mathbb{E}\|x - Q(x)\|^2 \leq (1-\delta)\|x\|^2, \tag{2}$$

for all $x \in \mathbb{R}^d$. $\tilde{Q}$ is an unbiased compressor if there is $\omega \geq 0$ such that

$$\mathbb{E}[\tilde{Q}(x)] = x \quad \text{and} \quad \mathbb{E}\|\tilde{Q}(x)\|^2 \leq (\omega+1)\|x\|^2, \tag{3}$$

for all $x \in \mathbb{R}^d$.

Quantization can reduce the communicated bits to improve the communication efficiency, but it will also slow down the convergence rate generally. Hence, error feedback or error compensation scheme is often used to improve the performance of quantization algorithms. For unbiased compressor, if we assume the accumulated quantization error is bounded, the convergence rate of error compensated SGD is the same as vanilla SGD [16]. However, if we only assume bounded stochastic gradient, in order to guarantee the boundedness of the accumulated quantization error, some decaying factor need to be involved in general, and the error compensated SGD is proved to have some advantage over QSGD in some perspective for convex quadratic problem [19]. On the other hand, for contraction compressor (for example TopK compressor [3]), the error compensated SGD actually has the same convergence rate as Vanilla SGD [14, 15, 17]. If f is non-smooth and $\psi = 0$, error compensated SGD was studied in [7] in the single node case, and the convergence rate is of order $O\left(1/\sqrt{\delta k}\right)$.

For variance-reduced methods, there is QSVRG [2] for the smooth case where ψ in problem (1) is zero, and there is VR-DIANA [6] for the composite or regularized case where ψ in problem (1) is nonzero. However, the compressor of both algorithms need to be unbiased. In this paper, we study the error compensated methods for loopless SVRG (L-SVRG) [8] for any contraction compressor.

1.1. Contributions

Iteration complexity. Denote the smoothness of f, $f^{(\tau)}$, and $f_i^{(\tau)}$ as L_f, $\bar{L}$, and L, respectively. In the composite case, the iteration complexity of error compensated L-SVRG (EC-LSVRG) is

$$O\left(\left(\tfrac{1}{\delta} + \tfrac{1}{p} + \tfrac{(1-\delta)\bar{L}}{\delta^2\mu} + \tfrac{(1-\delta)L}{\delta\mu} + \tfrac{L_f}{\mu} + \tfrac{L}{n\mu}\right)\ln\tfrac{1}{\epsilon}\right).$$

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

$$O\left(\left(\tfrac{1}{\delta} + \tfrac{1}{p} + \tfrac{(1-\delta)L_f}{\delta^2\mu} + \tfrac{(1-\delta)L}{n\delta\mu} + \tfrac{L_f}{\mu} + \tfrac{L}{n\mu}\right)\ln\tfrac{1}{\epsilon}\right).$$

In the smooth case, the iteration complexity of EC-LSVRG is

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

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

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

In particular, if $\frac{L_f}{\delta} \leq \frac{L}{n}$, then the above iteration complexity becomes

$$O\left(\left(\frac{1}{p}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\ln\frac{1}{\epsilon}\right),$$

which is actually the iteration complexity of the uncompressed L-SVRG [11]. Noticing that $L_f \leq L \leq mnL_f$, this means that in the extreme case: $L = mnL_f$, the error compensated L-SVRG has the same convergence rate as the uncompressed L-SVRG as long as $\frac{1}{\delta} \leq m$.

Communication complexity. Considering the communication complexity, we obtaint the optimal choice of p. In particular, we can choose $p = O(r(Q))$ to get the optimial communication complexity, where $r(Q)$ is the *compression ratio* for the compressor Q defined in (7). When $L_f = \bar{L} = L$, by choosing the optimal p, the communication complexity of EC-LSVRG in the composite case becomes

$$O\left(\Delta_1\left(\frac{r(Q)}{\delta}+1+\left(r(Q)+\frac{(1-\delta)r(Q)}{\delta^2}\right)\frac{L}{\mu}\right)\ln\frac{1}{\epsilon}\right),$$

where Δ_1 is the communication cost of the uncompressed vector $x \in \mathbb{R}^d$, and the communication complexity of EC-LSVRG in the smooth case becomes

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

2. Gradient Compression Methods

We now give a few examples of contraction compressors:

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 [15]. In this way, TopK compressor is a deterministic operator (well-defined when there are equal dimensions).

RandK compressor. For a parameter $1 \leq K \leq d$, 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 $x \in \mathbb{R}^d$.

In general, given an arbitrary unbiased compressor, we can obtain a contraction compressor via scaling as follows. For any unbiased compressor $\tilde{Q}$ satisfying (3), $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor satisfying (2) with $\delta = \frac{1}{\omega+1}$. Indeed,

$$\begin{aligned} \mathbb{E}\|\tfrac{1}{\omega+1}\tilde{Q}(x) - x\|^2 &= \tfrac{1}{(\omega+1)^2}\mathbb{E}\|\tilde{Q}(x)\|^2 + \|x\|^2 - \tfrac{2}{\omega+1}\mathbb{E}\langle\tilde{Q}(x), x\rangle \\ &\leq \tfrac{1}{\omega+1}\|x\|^2 + \|x\|^2 - \tfrac{2}{\omega+1}\|x\|^2 = \left(1-\tfrac{1}{\omega+1}\right)\|x\|^2. \end{aligned}$$

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

Lemma 1 (Lemma A.1 in [15]) *For the* TopK *and* RandK *compressors with* $1 \leq K \leq d$*, we have*

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

We may use the following assumptions for the contraction compressor in some cases.

Assumption 2.1 $\mathbb{E}[Q(x)] = \delta x$.

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

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

For TopK, 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-SVRG

The following is the error compensated L-SVRG algorithm. The search direction in L-SVRG is

$$\frac{1}{n}\sum_{\tau=1}^n \left(\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k) + \nabla f^{(\tau)}(w^k)\right), \tag{4}$$

where i_k^τ is sampled uniformly and independently from $[m] := \{1, 2, ..., m\}$ on τ-th node for $1 \leq \tau \leq n$, x^k is the current iteration, and w^k is the reference point. Since when 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 when x^k and w^k goes 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. Thus, we compress $\eta g_\tau^k + e_\tau^k$ 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, and u^k. Each node sends their compressed vector y_τ^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 y_τ^k is received, we add $\eta\nabla f(w^k)$ to it as the search direction. The proximal step is taken on each node, where we use the standard proximal operator: $\mathrm{prox}_{\eta\psi}(x) := \arg\min_y\{\frac{1}{2}\|x - y\|^2 + \eta\psi(y)\}$. The reference point

Algorithm 1: Error compensated Loopless SVRG (EC-LSVRG)

Parameters: stepsize $\eta > 0$; probability $p \in (0, 1]$
Initialization: $x^0 = w^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$
for $k = 0, 1, 2, ...$ **do**
  for $\tau = 1, ..., n$ **do**
    Sample i_k^τ uniformly and independently in $[m]$ on each node
    $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k), \quad y_\tau^k = Q(\eta g_\tau^k + e_\tau^k), \quad e_\tau^{k+1} = e_\tau^k + \eta g_\tau^k - y_\tau^k$
    $u_\tau^{k+1} = 0$ for $\tau = 2, ..., n$, $\quad u_1^{k+1} = \begin{cases} 1 & \text{with propobility } p \\ 0 & \text{with probability } 1-p \end{cases}$
    Send y_τ^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 y_τ^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
  $y^k = \frac{1}{n}\sum_{\tau=1}^n y_\tau^k$, $\quad u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1}$, $\quad x^{k+0.5} = x^k - (y^k + \eta \nabla f(w^k))$
  $x^{k+1} = \text{prox}_{\eta\psi}(x^{k+0.5}), \quad w^{k+1} = \begin{cases} x^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$
end

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.

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 $\tilde{x}^k = x^k - e^k$ for $k \geq 0$. Then

$$e^{k+1} = \frac{1}{n}\sum_{\tau=1}^n \left(e_\tau^k + \eta g_\tau^k - y_\tau^k\right) = e^k + \eta g^k - y^k,$$

and

$$\begin{aligned} \tilde{x}^{k+1} &= x^{k+1} - e^{k+1} \\ &= x^{k+0.5} - \eta \partial\psi(x^{k+1}) - e^{k+1} \\ &= x^k - y^k - \eta \nabla f(w^k) - \eta \partial\psi(x^{k+1}) - e^k - \eta g^k + y^k \\ &= \tilde{x}^k - \eta(g^k + \nabla f(w^k) + \partial\psi(x^{k+1})). \end{aligned}$$

3.1. Composite case

We need the following assumption in this subsection.

Assumption 3.1 *$f_i^{(\tau)}$ is L-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth, and ψ is μ-strongly convex. $L_f \geq \mu$.*

The follwoings are the main results. We use two Lyapulov functions for two cases: with or without Assumption 2.1 or Assumption 2.2 in the following two theorems.

Theorem 2 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Define*

$$\begin{aligned}\Phi_1^k \quad := \quad & \left(1+\tfrac{\eta\mu}{2}\right)\|\tilde{x}^k - x^*\|^2 + \tfrac{9}{\delta n}\sum_{\tau=1}^n \|e_\tau^k\|^2 \\ & + \tfrac{2\eta^2}{p}\left(\tfrac{41(1-\delta)}{\delta}\left(\tfrac{\bar{L}}{\delta}+L\right)+16L_f+\tfrac{16L}{n}\right)[P(w^k)-P(x^*)].\end{aligned}$$

If $\eta \leq \frac{1}{4L_f}$, then we have

$$\begin{aligned}\mathbb{E}[\Phi_1^{k+1}] \quad \leq \quad & \left(1-\min\left\{\tfrac{\mu\eta}{3},\tfrac{\delta}{4},\tfrac{p}{2}\right\}\right)\mathbb{E}[\Phi_1^k] + 2\eta\mathbb{E}[P(x^*)-P(x^{k+1})] \\ & + \left(\tfrac{123(1-\delta)}{\delta}\left(\tfrac{\bar{L}}{\delta}+L\right)+48L_f+\tfrac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)].\end{aligned}$$

Theorem 3 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Define*

$$\begin{aligned}\Phi_2^k \quad := \quad & \left(1+\tfrac{\eta\mu}{2}\right)\|\tilde{x}^k - x^*\|^2 + \tfrac{9}{\delta}\|e^k\|^2 + \tfrac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^n \|e_\tau^k\|^2 \\ & + \tfrac{2\eta^2}{p}\left(\tfrac{(1-\delta)}{\delta}\left(\tfrac{82L_f}{\delta}+\tfrac{336\bar{L}}{\delta n}+\tfrac{459L}{n}\right)+16L_f+\tfrac{16L}{n}\right)[P(w^k)-P(x^*)].\end{aligned}$$

Under Assumption 2.1 or Assumption 2.2, if $\eta \leq \frac{1}{4L_f}$, then we have

$$\begin{aligned}\mathbb{E}[\Phi_2^{k+1}] \quad \leq \quad & \left(1-\min\left\{\tfrac{\mu\eta}{3},\tfrac{\delta}{4},\tfrac{p}{2}\right\}\right)\mathbb{E}[\Phi_2^k] + 2\eta\mathbb{E}[P(x^*)-P(x^{k+1})] \\ & + \left(\tfrac{(1-\delta)}{\delta}\left(\tfrac{246L_f}{\delta}+\tfrac{1008\bar{L}}{\delta n}+\tfrac{1377L}{n}\right)+48L_f+\tfrac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)].\end{aligned}$$

From the above two theorems, we can get the iteration complexity.

Theorem 4 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Let $w_k = \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If $\eta \leq \frac{\delta^2}{135(1-\delta)(\bar{L}+L\delta)+53L_f\delta^2+53L\delta^2/n}$, then we have*

$$\mathbb{E}[P(\bar{x}^k)-P(x^*)] \leq \frac{\frac{\mu}{2}\|x^0-x^*\|^2+\frac{1}{2}(P(x^0)-P(x^*))}{1-(1-\min\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1-\min\left\{\tfrac{\mu\eta}{3},\tfrac{\delta}{4},\tfrac{p}{2}\right\}\right)^k.$$

In particular, if we choose $\eta = \frac{\delta^2}{135(1-\delta)(\bar{L}+L\delta)+53L_f\delta^2+53L\delta^2/n}$, then $\mathbb{E}[P(\bar{x}^k)-P(x^*)] \leq \epsilon$, with $\epsilon \leq \frac{\mu}{2}\|x^0-x^*\|^2+\frac{1}{2}(P(x^0)-P(x^*))$, as long as

$$k \geq O\left(\left(\tfrac{1}{\delta}+\tfrac{1}{p}+\tfrac{(1-\delta)\bar{L}}{\delta^2\mu}+\tfrac{(1-\delta)L}{\delta\mu}+\tfrac{L_f}{\mu}+\tfrac{L}{n\mu}\right)\ln\left(\tfrac{\mu\|x^0-x^*\|^2+P(x^0)-P(x^*)}{\epsilon}\right)\right).$$

Theorem 5 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.1 holds. Assume the compressor Q also satisfies Assumption 2.1 or Assumption 2.2. Let $w_k = \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If*

$$\eta \leq \frac{\delta^2}{(1-\delta)(269L_f+1100\bar{L}/n+1503L\delta/n)+53L_f\delta^2+53L\delta^2/n},$$

then we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \frac{\frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))}{1-(1-\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}} \left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

In particular, if we choose $\eta = \frac{\delta^2}{(1-\delta)(269L_f + 1100\bar{L}/n + 1503L\delta/n) + 53L_f\delta^2 + 53L\delta^2/n}$, *then* $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$, *with* $\epsilon \leq \frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))$, *as long as*

$$k \geq O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{(1-\delta)L_f}{\delta^2\mu} + \frac{(1-\delta)L}{n\delta\mu} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right) \ln\left(\frac{\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)}{\epsilon}\right)\right).$$

Noticing that $L_f \leq \bar{L} \leq nL_f$ and $\bar{L} \leq L \leq m\bar{L}$, the iteration complexity in Theorem 5 could be better than that in Theorem 4. On the other hand, if $L_f = \bar{L} = L$, then both iteration complexities in Theorem 4 and Theorem 5 become

$$O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{L}{\mu} + \frac{(1-\delta)L}{\delta^2\mu}\right) \ln\left(\frac{\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)}{\epsilon}\right)\right). \tag{5}$$

3.2. Smooth case: $\psi = 0$

In this subsection, we study the Algorithm 1 for problem (1) with $\psi = 0$. We need the following assumption in this subsection.

Assumption 3.2 *$f_i^{(\tau)}$ is L-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth and f is μ-strongly convex.*

We also use two Lyapulov functions for two cases: with or without Assumption 2.1 or Assumption 2.2 in the following two theorems.

Theorem 6 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.2 holds. Define*

$$\Phi_3^k := \|\tilde{x}^k - x^*\|^2 + \frac{12L_f\eta}{n\delta}\sum_{\tau=1}^n \|e_\tau^k\|^2 + \frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) + \frac{4L\eta^2}{n}\right)[f(w^k) - f(x^*)].$$

If $\eta \leq \frac{1}{4L_f + 8L/n}$, *then*

$$\begin{aligned}\mathbb{E}[\Phi_3^{k+1}] \quad \leq \quad & \left(1 - \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)\mathbb{E}[\Phi_3^k] \\ & -\frac{\eta}{2}\left(1 - \frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) - \frac{16L\eta}{n}\right)\mathbb{E}[f(x^k) - f(x^*)].\end{aligned}$$

Theorem 7 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.2 holds. Define*

$$\begin{aligned}\Phi_4^{k+1} \quad = \quad & \|\tilde{x}^{k+1} - x^*\|^2 + \frac{12L_f\eta}{\delta}\|e^{k+1}\|^2 + \frac{96(1-\delta)L_f\eta}{n^2\delta}\sum_{\tau=1}^n \|e_\tau^{k+1}\|^2 \\ & + \frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta} + \frac{11L}{n} + \frac{8\bar{L}}{n\delta}\right) + \frac{4L\eta^2}{n}\right)[f(w^{k+1}) - f(x^*)].\end{aligned}$$

Under Assumption 2.1 or Assumption 2.2, if $\eta \leq \frac{1}{4L_f + 8L/n}$, *then*

$$\begin{aligned}\mathbb{E}[\Phi_4^{k+1}] \quad \leq \quad & \left(1 - \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)\mathbb{E}[\Phi_4^k] \\ & -\frac{\eta}{2}\left(1 - \frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{2L_f}{\delta} + \frac{11L}{n} + \frac{8\bar{L}}{n\delta}\right) - \frac{16L\eta}{n}\right)\mathbb{E}[f(x^k) - f(x^*)].\end{aligned}$$

From the above two theorems, we can get the iteration complexity.

Theorem 8 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.2 holds. Let $w_k = \left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If $\eta \leq \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_f L}}\right\}$, then we have*

$$\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \frac{9\mu\|x^0-x^*\|^2+9(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)^k.$$

In particular, if we choose $\eta = \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_f L}}\right\}$, then $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$, with $\epsilon \leq 9\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)$, as long as

$$k \geq O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{\sqrt{(1-\delta)L_f\bar{L}}}{\mu\delta}+\frac{\sqrt{(1-\delta)L_f L}}{\mu\sqrt{\delta}}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\ln\left(\frac{18\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right).$$

Theorem 9 *Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3.2 holds. Assume the compressor Q also satisfies Assumption 2.1 or Assumption 2.2. Let $w_k = \left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If*

$$\eta \leq \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_f L}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f\bar{L}}}\right\},$$

then we have

$$\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \frac{12\mu\|x^0-x^*\|^2+12(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)^k.$$

In particular, if we choose $\eta = \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_f L}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f\bar{L}}}\right\}$, then $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$ with $\epsilon \leq 12\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)$ as long as

$$k \geq O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{\sqrt{(1-\delta)}L_f}{\mu\delta}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\ln\left(\frac{24\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right).$$

Noticing that $L_f \leq \bar{L} \leq nL_f$ and $\bar{L} \leq L \leq m\bar{L}$, the iteration complexity in Theorem 9 could be better than that in Theorem 8. On the other hand, if $L_f = \bar{L} = L$, then both iteration complexities in Theorem 8 and Theorem 9 become

$$O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{L}{\mu}+\frac{\sqrt{(1-\delta)}L}{\delta\mu}\right)\ln\left(\frac{24\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right). \tag{6}$$

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] 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.

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

[7] 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.

[8] D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. *arXiv: 1901.08689*, 2019.

[9] 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.

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

[11] Xun Qian, Zheng Qu, and Peter Richtárik. L-svrg and l-katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

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

Appendix

Appendix A. 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 $r(Q)$ for the contraction compressor Q as

$$r(Q) := \sup_{x\in\mathbb{R}^d} \left\{ \mathbb{E}\left[\frac{\text{communication cost of } Q(x)}{\Delta_1} \right] \right\}. \tag{7}$$

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,

$$\begin{aligned} \mathcal{T}_k &\leq \Delta_1 r(Q) + 1 + \Delta_1 + (\Delta_1 r(Q) + 1 + p\Delta_1)k \\ &\leq \Delta_1 r(Q) + 1 + \Delta_1 + (\Delta_1 r(Q) + 1)\left(1 + \tfrac{p}{r(Q)}\right)k. \end{aligned} \tag{8}$$

From Theorem 4 and Theorem 5 in the composite case and Theorem 8 and Theorem 9 in the smooth case, we have $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \leq \epsilon$ as long as $k \geq O\left(\left(\frac{1}{p} + a\right)\ln\frac{1}{\epsilon}\right)$, where a is independent of p. Hence, from (8), we have $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \leq \epsilon$ for

$$\begin{aligned} \mathcal{T}_k &= O\left((\Delta_1 r(Q) + 1)\left(1 + \tfrac{p}{r(Q)}\right)\left(a + \tfrac{1}{p}\right)\ln\tfrac{1}{\epsilon}\right) \\ &= O\left((\Delta_1 r(Q) + 1)\left(a + \tfrac{pa}{r(Q)} + \tfrac{1}{p} + \tfrac{1}{r(Q)}\right)\ln\tfrac{1}{\epsilon}\right). \end{aligned}$$

Noticing that $\frac{pa}{r(Q)} + \frac{1}{p} \leq a + \frac{1}{r(Q)}$ for $\min\left\{r(Q), \frac{1}{a}\right\} \leq p \leq \max\left\{r(Q), \frac{1}{a}\right\}$, we have

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

and the above lower bound holds for $O\left(\min\left\{r(Q), \frac{1}{a}\right\}\right) \leq p \leq O\left(\max\left\{r(Q), \frac{1}{a}\right\}\right)$. Hence, in order to minimize the total expected communication cost, the optimal choice of p is $O\left(\min\left\{r(Q), \frac{1}{a}\right\}\right) \leq p \leq O\left(\max\left\{r(Q), \frac{1}{a}\right\}\right)$.

Comparison to the uncompressed L-SVRG. For simplicity, we assume $L_f = \bar{L} = L$ and $\Delta_1 r(Q) \geq O(1)$. In the composite case, from (5) and (8), by choosing $p = O(r(Q))$, we have

$$\mathcal{T}_k = O\left(\Delta_1\left(\tfrac{r(Q)}{\delta} + 1 + \left(r(Q) + \tfrac{(1-\delta)r(Q)}{\delta^2}\right)\tfrac{L}{\mu}\right)\ln\tfrac{1}{\epsilon}\right). \tag{9}$$

In the smooth case, from (6) and (8), by choosing $p = O(r(Q))$, we have

$$\mathcal{T}_k = O\left(\Delta_1\left(\tfrac{r(Q)}{\delta} + 1 + \left(r(Q) + \tfrac{\sqrt{(1-\delta)}r(Q)}{\delta}\right)\tfrac{L}{\mu}\right)\ln\tfrac{1}{\epsilon}\right). \tag{10}$$

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

$$\mathcal{T}_k = O\left(\Delta_1 \tfrac{L}{\mu}\ln\tfrac{1}{\epsilon}\right). \tag{11}$$

Thus, in the composite case, If $\frac{r(Q)}{\delta^2} < 1$, then the communication cost in (9) is less than that in (11). In the smooth case, If $\frac{r(Q)}{\delta} < 1$, then the communication cost in (10) is less than that in (11). 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 B. Experiments

In this part, we run experiments with EC-LSVRG to demonstrate the empirical effectiveness. In particular, we should highlight the linear convergence rate of our algorithm with biased compressor in strongly convex case. Also, the convergence speed is competitive to other compressed algorithms.

Settings: We implement (1) Linear regression for GD; (2) Logistic regression with L_1-L_2 regularization (detailed information can be found in the Appendix). We use Python 3.7 to perform experiments on a server with 2 processors (Intel Xeon Gold 5120 @ 2.20GHz), 28 cores in total. Library include numpy, sklearn. In particular, for multi-nodes tasks, we use mpi4py to simulate distributed environment. Step size is searched from 10^t, where $t \in \{-4, \cdots, 0, 1\}$. For DIANA, we use the optimal $\alpha = (w+1)^{-1}$.

Datasets: For GD, in order to construct two synthesis data, which are sampled from normal distribution with $\sigma = 0.1$. In particular, we use 2048 and 4096 dimension case for the problem. To increase the difficulty in distributed setting, we only use square matrix, which enlarge the local solution space on each node. Besides, we have real datasets: *Gisette*, *RCV1*, *a5a*, *mushrooms* (details are in the Appendix).

B.1. Biased Compressor GD

Since there are not any algorithm achieves linear convergence in stochastic optimization setting. We firstly compare our algorithm with the error compensated GD (EC-GD) to demonstrate to the gain of variance reduced term. We should mention that EC-GD only converges linearly when $\nabla f^{(\tau)}(x^*) = 0$ for any $\tau = 1, \cdots, n$, which means the global solution is also a local solution. Thus we construct a linear regression problem $\|Ax - Y\|_2^2$ such that $\|Ax^* - Y\|_2^2 = 0$. By Figure 1, we use 2048-dim and 4096-dim to conduct the experiment. We can notice that, though the gap between EC-GD and GD is small, EC-LSVRG bridge this gap with the variance reduced term. Moreover, when the optimal point $\nabla f^{(\tau)}(x^*) \neq 0$, the EC-GD algorithm is unable to acheive the linear convergence rate. As is shown, for *a5a* and *mushrooms* dataset, we conduct logistic regression for each method. We verified the linear convergence of our algorithm emprically and showed that the EC-GD can only converge to neighbourhood of the optimality.

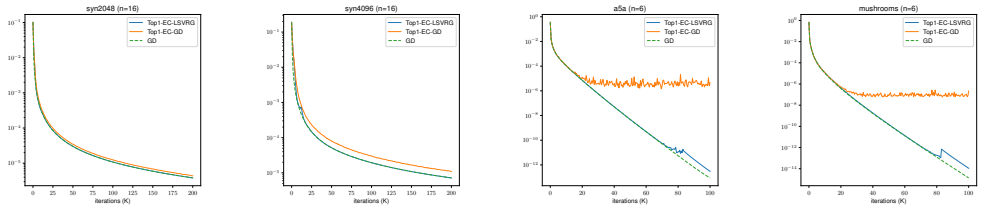

Figure 1: Compare to Error Compensated gradient descent ($p = 10^{-3}$)

B.2. Compressed Stochastic Algorithms with Linear Convergence

For simplicity, we compare the 1-node case to illustrate the effectiveness of our method. By Figure 2, we can notice that with the same compressor, like RandK or quantization (k-bit denotes $s = 2^k$), our EC-LSVRG outperforms VRDIANA, which is a state-of-the-art approach that support proximal operators with linear convergence. More importantly, since the biased compressors are often more effective than the unbiased ones, when we compare these method w.r.t. the communication cost, the TopK compressor is much more efficient than quantization.

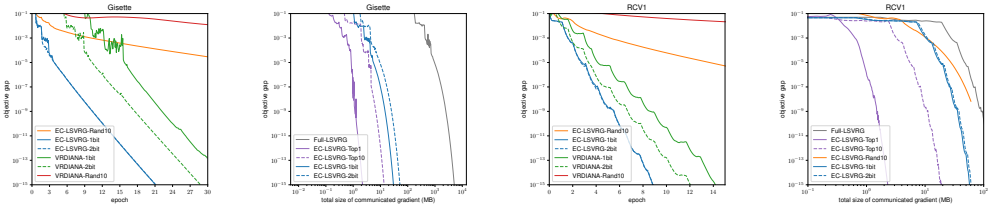

Figure 2: Compare to compressed algorithm with linear convergence ($p = \frac{1}{mn}$)

B.3. Distributed Experiment with Compressed Stochastic Algorithms

In order to reflect the influence of the number of nodes, we compared the convergence speed of different nodes. In Figure 3, we can notice that as the number of nodes increases, the convergence speed can be improved significantly for any compressor. In particular, the TopK compressor is still very competitive, epecially for the communication cost.

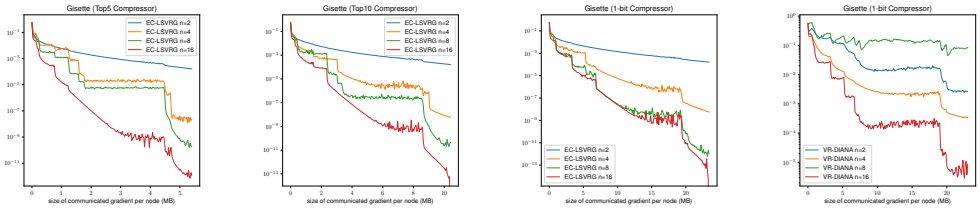

Figure 3: Distributed Experiment with compressed algorithm ($p = 10^{-4}$)

Appendix C. Proofs for the composite case

C.1. Lemmas

The following lemma shows the progress at iteration k for the auxiliary points $\tilde{x}^k$ and $\tilde{x}^{k+1}$.

Lemma 10 *If* $\eta \leq \frac{1}{4L_f}$, *then*

$$\begin{aligned}\left(1+\tfrac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2 \quad &\leq \quad \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\mathbb{E}\|e^k\|^2\\ &\quad +(1+\eta\mu)\mathbb{E}\|e^{k+1}\|^2+4\eta^2\mathbb{E}\|g^k\|^2.\end{aligned}$$

Proof

Since $\tilde{x}^{k+1}=\tilde{x}^k-\eta(g^k+\nabla f(w^k)+\partial\psi(x^{k+1}))$, we have

$$\begin{aligned}&\langle\eta(g^k+\nabla f(w^k)),x^*-x^{k+1}\rangle\\ =\;&\langle\tilde{x}^k-\tilde{x}^{k+1}-\eta\partial\psi(x^{k+1}),x^*-x^{k+1}\rangle\\ =\;&\langle\tilde{x}^k-x^{k+1},x^*-x^{k+1}\rangle+\langle x^{k+1}-\tilde{x}^{k+1},x^*-x^{k+1}\rangle-\eta\langle\partial\psi(x^{k+1}),x^*-x^{k+1}\rangle\\ \geq\;&\frac{1}{2}\left(-\|\tilde{x}^k-x^*\|^2+\|\tilde{x}^k-x^{k+1}\|^2+\|x^{k+1}-x^*\|^2\right)+\frac{1}{2}\left(\|\tilde{x}^{k+1}-x^*\|^2\right.\\ &\left.-\|x^{k+1}-\tilde{x}^{k+1}\|^2-\|x^{k+1}-x^*\|^2\right)+\eta\left(\psi(x^{k+1})-\psi(x^*)+\frac{\mu}{2}\|x^{k+1}-x^*\|^2\right)\\ =\;&\frac{1}{2}\|\tilde{x}^{k+1}-x^*\|^2-\frac{1}{2}\|\tilde{x}^k-x^*\|^2+\frac{1}{2}\|\tilde{x}^k-x^{k+1}\|^2-\frac{1}{2}\|\tilde{x}^{k+1}-x^{k+1}\|^2\\ &+\eta\left(\psi(x^{k+1})-\psi(x^*)+\frac{\mu}{2}\|x^{k+1}-x^*\|^2\right).\end{aligned}$$

From $\|\tilde{x}^k-x^{k+1}\|^2\geq\frac{1}{2}\|x^{k+1}-x^k\|^2-\|\tilde{x}^k-x^k\|^2$, and $\|x^{k+1}-x^*\|^2\geq\frac{1}{2}\|\tilde{x}^{k+1}-x^*\|^2-\|\tilde{x}^{k+1}-x^{k+1}\|^2$, we arrive at

$$\begin{aligned}&\langle\eta(g^k+\nabla f(w^k)),x^*-x^{k+1}\rangle\\ \geq\;&\frac{1+\eta\mu/2}{2}\|\tilde{x}^{k+1}-x^*\|^2-\frac{1}{2}\|\tilde{x}^k-x^*\|^2+\frac{1}{4}\|x^{k+1}-x^k\|^2-\frac{1}{2}\|\tilde{x}^k-x^k\|^2\\ &-\frac{1+\eta\mu}{2}\|\tilde{x}^{k+1}-x^{k+1}\|^2+\eta(\psi(x^{k+1})-\psi(x^*)).\end{aligned}\tag{12}$$

Since f is convex and $\mathbb{E}_k[g^k + \nabla f(w^k)] = \nabla f(x^k)$, we have

$$\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \\
&= f(x^k) + \mathbb{E}_k[\langle g^k + \nabla f(w^k), x^* - x^{k+1} + x^{k+1} - x^k \rangle] \\
&= f(x^k) + \mathbb{E}_k[\langle g^k + \nabla f(w^k), x^* - x^{k+1} \rangle] + \mathbb{E}_k[\langle g^k + \nabla f(w^k) - \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\quad + \mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\geq \mathbb{E}_k[f(x^{k+1})] - \frac{L_f}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[\langle g^k + \nabla f(w^k), x^* - x^{k+1} \rangle] \\
&\quad + \mathbb{E}_k[\langle g^k + \nabla f(w^k) - \nabla f(x^k), x^{k+1} - x^k \rangle] \\
&\geq \mathbb{E}_k[f(x^{k+1})] - \frac{L_f}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[\langle g^k + \nabla f(w^k), x^* - x^{k+1} \rangle] \\
&\quad - \frac{1}{2\beta}\mathbb{E}_k[\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2] - \frac{\beta}{2}\mathbb{E}_k[x^{k+1} - x^k\|^2],
\end{aligned}$$

where the second inequality comes from that f is L_f-smooth and the last inequality comes from Young's inequality.

Since $\mathbb{E}_k[\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2] \leq \mathbb{E}_k\|g^k\|^2$, by choosing $\beta = \frac{1}{4\eta}$, we have

$$\begin{aligned}
& f(x^*) \\
\geq\ & \mathbb{E}_k[f(x^{k+1})] - \left(\frac{L_f}{2} + \frac{1}{8\eta}\right)\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[\langle g^k + \nabla f(w^k), x^* - x^{k+1} \rangle] - 2\eta\mathbb{E}_k\|g^k\|^2 \\
\overset{(12)}{\geq}\ & \mathbb{E}_k[f(x^{k+1})] + \left(\frac{1}{4\eta} - \frac{L_f}{2} - \frac{1}{8\eta}\right)\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{1 + \eta\mu/2}{2\eta}\mathbb{E}_k\|\tilde{x}^{k+1} - x^*\|^2 - \frac{1}{2\eta}\|\tilde{x}^k - x^*\|^2 \\
& - \frac{1}{2\eta}\|\tilde{x}^k - x^k\|^2 - \frac{1 + \eta\mu}{2\eta}\mathbb{E}_k\|\tilde{x}^{k+1} - x^{k+1}\|^2 + \mathbb{E}_k[\psi(x^{k+1})] - \psi(x^*) - 2\eta\mathbb{E}_k\|g^k\|^2.
\end{aligned}$$

Noticing that $\frac{1}{4\eta} - \frac{L_f}{2} - \frac{1}{8\eta} \geq 0$ if $\eta \leq \frac{1}{4L_f}$, we can get the result after rearrangement.

■

Lemma 11 *We have*

$$\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|g_\tau^k\|^2 \quad \leq \quad 4L\mathbb{E}[P(x^k) - P(x^*)] + 4L\mathbb{E}[P(w^k) - P(x^*)], \tag{13}$$

and

$$\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \quad \leq \quad 4\bar{L}\mathbb{E}[P(x^k) - P(x^*)] + 4\bar{L}\mathbb{E}[P(w^k) - P(x^*)], \tag{14}$$

and

$$\mathbb{E}\|g^k\|^2 \quad \leq \quad 4\left(L_f + \tfrac{L}{n}\right)\mathbb{E}[P(x^k) - P(x^*)] + 4\left(L_f + \tfrac{L}{n}\right)\mathbb{E}[P(w^k) - P(x^*)]. \tag{15}$$

and

$$\mathbb{E}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2 \quad \leq \quad \tfrac{4L}{n}\mathbb{E}[P(x^k) - P(x^*) + P(w^k) - P(x^*)]. \tag{16}$$

Proof

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

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

and

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L_f\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)$$

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

$$\begin{aligned}\mathbb{E}\|g_\tau^k\|^2 &= \mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k)\|^2 \\ &\leq 2\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(x^*)\|^2 + 2\mathbb{E}\|\nabla f_{i_k^\tau}^{(\tau)}(w^k) - \nabla f_{i_k^\tau}^{(\tau)}(x^*)\|^2 \\ &\leq 4L\mathbb{E}[f^{(\tau)}(x^k) - f^{(\tau)}(x^*) - \langle \nabla f^{(\tau)}(x^*), x^k - x^*\rangle] \\ &\quad +4L\mathbb{E}[f^{(\tau)}(w^k) - f^{(\tau)}(x^*) - \langle \nabla f^{(\tau)}(x^*), w^k - x^*\rangle],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\|g^k\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{\tau=1}^n g_\tau^k\right\|^2 \\ &= \frac{1}{n^2}\mathbb{E}\left\langle \sum_{\tau=1}^n g_\tau^k, \sum_{\tau=1}^n g_\tau^k\right\rangle \\ &= \frac{1}{n^2}\sum_{\tau_1,\tau_2=1}^n \mathbb{E}\left\langle g_{\tau_1}^k, g_{\tau_2}^k\right\rangle \\ &= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 + \frac{1}{n^2}\sum_{\tau_1\neq\tau_2} \mathbb{E}\left\langle \nabla f^{(\tau_1)}(x^k) - \nabla f^{(\tau_1)}(w^k), \nabla f^{(\tau_2)}(x^k) - \nabla f^{(\tau_2)}(w^k)\right\rangle \\ &= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 + \mathbb{E}\|\nabla f(x^k) - \nabla f(w^k)\|^2 - \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \\ &\leq \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 + 2\mathbb{E}\|\nabla f(x^k) - \nabla f(x^*)\|^2 + 2\mathbb{E}\|\nabla f(w^k) - \nabla f(x^*)\|^2 \\ &\leq \left(\frac{4L}{n} + 4L_f\right)\mathbb{E}[f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle] \\ &\quad + \left(\frac{4L}{n} + 4L_f\right)\mathbb{E}[f(w^k) - f(x^*) - \langle \nabla f(x^*), w^k - x^*\rangle].\end{aligned}$$

Since x^* is an optimal solution, we have $-\nabla f(x^*) \in \partial\psi(x^*)$, which implies that

$$f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle \leq P(x^k) - P(x^*). \tag{17}$$

Thus,

$$\frac{1}{n}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 \leq 4L\mathbb{E}[P(x^k) - P(x^*)] + 4L\mathbb{E}[P(w^k) - P(x^*)],$$

and

$$\mathbb{E}\|g^k\|^2 \leq \left(\frac{4L}{n} + 4L_f\right)\mathbb{E}[P(x^k) - P(x^*)] + \left(\frac{4L}{n} + 4L_f\right)\mathbb{E}[P(w^k) - P(x^*)].$$

For $\mathbb{E}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2$, we have

$$\begin{aligned}\mathbb{E}\|g^k + \nabla f(w^k) - \nabla f(x^k)\|^2 &= \mathbb{E}\|g^k\|^2 - \mathbb{E}\|\nabla f(x^k) - \nabla f(w^k)\|^2 \\ &= \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 - \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|\nabla f^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k)\|^2 \\ &\leq \frac{1}{n^2}\sum_{\tau=1}^n \mathbb{E}\|g_\tau^k\|^2 \\ &\leq \frac{4L}{n}\mathbb{E}[P(x^k) - P(x^*)] + \frac{4L}{n}\mathbb{E}[P(w^k) - P(x^*)].\end{aligned}$$

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

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

Then similarly, we can get

$$\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2 \le 4\bar{L}\mathbb{E}[P(x^k)-P(x^*)]+4\bar{L}\mathbb{E}[P(w^k)-P(x^*)].$$

■

The following two lemmas show the evolution of e_τ^k and e^k, which will be used to construct the Lyapulov functions.

Lemma 12 *We have*

$$\begin{aligned}\tfrac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}[\|e_\tau^{k+1}\|^2] \quad &\le \quad \left(1-\tfrac{\delta}{2}\right)\tfrac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2 \\ &\quad +4(1-\delta)\eta^2\left(\tfrac{\bar{L}}{\delta}+L\right)\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right).\end{aligned}$$

Proof

First, we have

$$\begin{aligned}&\mathbb{E}[\|e_\tau^{k+1}\|^2]\\ \overset{(2)}{\le}\ &(1-\delta)\mathbb{E}\|e_\tau^k+\eta g_\tau^k\|^2\\ =\ &(1-\delta)\mathbb{E}\|e_\tau^k+\eta(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))+\eta g_\tau^k-\eta(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2\\ =\ &(1-\delta)\mathbb{E}\|e_\tau^k+\eta(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2\\ &+(1-\delta)\eta^2\mathbb{E}\|g_\tau^k-(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2\\ \le\ &(1-\delta)\mathbb{E}\|e_\tau^k+\eta(\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k))\|^2+(1-\delta)\eta^2\mathbb{E}\|g_\tau^k\|^2\\ \le\ &(1-\delta)(1+\beta)\mathbb{E}\|e_\tau^k\|^2+(1-\delta)\left(1+\frac{1}{\beta}\right)\eta^2\mathbb{E}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2+(1-\delta)\eta^2\mathbb{E}\|g_\tau^k\|^2\\ \le\ &\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e_\tau^k\|^2+\frac{2(1-\delta)}{\delta}\eta^2\mathbb{E}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2+(1-\delta)\eta^2\mathbb{E}\|g_\tau^k\|^2,\end{aligned}$$

where we use Young's inequality in the third inequality and choose $\beta=\frac{\delta}{2(1-\delta)}$ when $\delta<1$. When $\delta=1$, it is easy to see that the above inequality also holds.

Then we can get

$$\begin{aligned}&\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}[\|e_\tau^{k+1}\|^2]\\ \le\ &\left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{2(1-\delta)}{\delta}\eta^2\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|\nabla f^{(\tau)}(x^k)-\nabla f^{(\tau)}(w^k)\|^2+(1-\delta)\eta^2\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|g_\tau^k\|^2\\ \overset{(14)}{\le}\ &\left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+\frac{8(1-\delta)\eta^2\bar{L}}{\delta}\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right)+(1-\delta)\eta^2\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|g_\tau^k\|^2\\ \overset{(13)}{\le}\ &\left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+4(1-\delta)\eta^2\left(\frac{\bar{L}}{\delta}+L\right)\left(\mathbb{E}[P(x^k)-P(x^*)]+\mathbb{E}[P(w^k)-P(x^*)]\right).\end{aligned}$$

■

Lemma 13 *Under Assumption 2.1 or Assumption 2.2, we have*

$$\begin{aligned}\mathbb{E}\|e^{k+1}\|^2 \quad &\le \quad \left(1-\tfrac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+\tfrac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2\\ &\quad +4(1-\delta)\eta^2\left(\tfrac{2L_f}{\delta}+\tfrac{3L}{n}\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].\end{aligned}$$

Proof

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

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

where we use the definitions of e^k and g^k in the last inequality.

Under Assumption 2.2, we have

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

Overall, under Assumption 2.1 or Assumption 2.2, we have

$$\begin{aligned}
&\mathbb{E}\|e^{k+1}\|^2 \\
\leq\ & (1-\delta)\mathbb{E}\left\|e^k + \eta g^k\right\|^2 + \frac{(1-\delta)\delta}{n^2}\sum_{\tau=1}^n\mathbb{E}\left\|e_\tau^k + \eta g_\tau^k\right\|^2 \\
\leq\ & (1-\delta)\mathbb{E}\left\|e^k + \eta g^k\right\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 + \frac{2(1-\delta)\delta\eta^2}{n^2}\sum_{\tau=1}^n\mathbb{E}\|g_\tau^k\|^2 \\
\overset{(13)}{\leq}\ & (1-\delta)\mathbb{E}\left\|e^k + \eta g^k\right\|^2 + \frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
&+ \frac{8(1-\delta)\delta L\eta^2}{n}\mathbb{E}[P(x^k) - P(x^*) + P(w^k) - P(x^*)]. \qquad (18)
\end{aligned}$$

For $(1-\delta)\mathbb{E}\|e^k+\eta g^k\|^2$, we have

$$\begin{aligned}
&(1-\delta)\mathbb{E}\left\|e^k+\eta g^k\right\|^2\\
=\;&(1-\delta)\mathbb{E}\left\|e^k+\eta(\nabla f(x^k)-\nabla f(w^k))+\eta g^k-\eta(\nabla f(x^k)-\nabla f(w^k))\right\|^2\\
=\;&(1-\delta)\mathbb{E}\left\|e^k+\eta(\nabla f(x^k)-\nabla f(w^k))\right\|^2+(1-\delta)\eta^2\mathbb{E}\|g^k-(\nabla f(x^k)-\nabla f(w^k))\|^2\\
\leq\;&\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+\frac{2(1-\delta)\eta^2}{\delta}\mathbb{E}\|\nabla f(x^k)-\nabla f(w^k)\|^2\\
&+(1-\delta)\eta^2\mathbb{E}\|g^k-(\nabla f(x^k)-\nabla f(w^k))\|^2\\
\overset{(16)}{\leq}\;&\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+\frac{2(1-\delta)\eta^2}{\delta}\mathbb{E}\|\nabla f(x^k)-\nabla f(w^k)\|^2\\
&+(1-\delta)\frac{4L\eta^2}{n}\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}$$

Since f is L_f-smooth, we have

$$\begin{aligned}
\mathbb{E}\|\nabla f(x^k)-\nabla f(w^k)\|^2 &= \mathbb{E}\|\nabla f(x^k)-\nabla f(x^*)+\nabla f(x^*)-\nabla f(w^k)\|^2\\
&\leq 2\mathbb{E}\|\nabla f(x^k)-\nabla f(x^*)\|^2+2\mathbb{E}\|\nabla f(w^k)-\nabla f(x^*)\|^2\\
&\leq 4L_f\mathbb{E}\left[f(x^k)-f(x^*)-\langle\nabla f(x^*),x^k-x^*\rangle\right]\\
&\quad+4L_f\mathbb{E}\left[f(w^k)-f(x^*)-\langle\nabla f(x^*),w^k-x^*\rangle\right]\\
&\overset{(17)}{\leq} 4L_f\mathbb{E}[P(x^k)-P(x^*)]+4L_f\mathbb{E}[P(w^k)-P(x^*)].
\end{aligned}$$

Hence, we arrive at

$$\begin{aligned}
&(1-\delta)\mathbb{E}\left\|e^k+\eta g^k\right\|^2\\
\leq\;&\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+4(1-\delta)\eta^2\left(\frac{2L_f}{\delta}+\frac{L}{n}\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}$$

Combining (18) and the above inequality, we can get

$$\begin{aligned}
&\mathbb{E}\|e^{k+1}\|^2\\
\leq\;&\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+\frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2\\
&+4(1-\delta)\eta^2\left(\frac{2L_f}{\delta}+\frac{L}{n}+\frac{2\delta L}{n}\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)]\\
\leq\;&\left(1-\frac{\delta}{2}\right)\mathbb{E}\|e^k\|^2+\frac{2(1-\delta)\delta}{n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2\\
&+4(1-\delta)\eta^2\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}$$

■

C.2. Proof of Theorem 2

Let $\eta\leq\frac{1}{4L_f}$. From $\|e^k\|^2\leq\frac{1}{n}\sum_{\tau=1}^n\|e_\tau^k\|^2$ and Lemma 10, we have

$$
\begin{aligned}
& \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2 \\
\leq\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\mathbb{E}\|e^k\|^2+(1+\eta\mu)\mathbb{E}\|e^{k+1}\|^2+4\eta^2\mathbb{E}\|g^k\|^2 \\
\leq\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\frac{1}{n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{5}{4n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2+4\eta^2\mathbb{E}\|g^k\|^2 \\
\overset{Lemma12}{\leq}\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\frac{1}{n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{5}{4}\left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +5(1-\delta)\eta^2\left(\frac{\bar{L}}{\delta}+L\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)]+4\eta^2\mathbb{E}\|g^k\|^2 \\
\overset{(15)}{\leq}\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\frac{9}{4}\cdot\frac{1}{n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\left(5(1-\delta)\left(\frac{\bar{L}}{\delta}+L\right)+16L_f+\frac{16L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}
$$

From the definition of w^{k+1}, we have

$$
\mathbb{E}[P(w^{k+1})-P(x^*)]=p\mathbb{E}[P(x^k)-P(x^*)]+(1-p)\mathbb{E}[P(w^k)-P(x^*)]. \tag{19}
$$

From Lemma 12, we have

$$
\begin{aligned}
& \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{9}{\delta n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\frac{9}{\delta n}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\left(\frac{41(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+16L_f+\frac{16L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}
$$

Combining the above inequality and (19), we can get

$$
\begin{aligned}
\mathbb{E}[\Phi_1^{k+1}] =\ & \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{9}{\delta n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
& +\frac{2\eta^2}{p}\left(\frac{41(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+16L_f+\frac{16L}{n}\right)\mathbb{E}[P(w^{k+1})-P(x^*)] \\
\leq\ & \mathbb{E}\|\tilde{x}^{k}-x^*\|^2+\frac{9}{\delta n}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\frac{2\eta^2}{p}\left(\frac{41(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+16L_f+\frac{16L}{n}\right)\left(1-\frac{p}{2}\right)\mathbb{E}[P(w^k)-P(x^*)] \\
& +2\eta\mathbb{E}[P(x^*)-P(x^{k+1})]+\left(\frac{123(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+48L_f+\frac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)] \\
\leq\ & \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)\mathbb{E}[\Phi_1^k]+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})] \\
& +\left(\frac{123(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+48L_f+\frac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)],
\end{aligned}
$$

where we use $\left(1+\frac{\eta\mu}{2}\right)^{-1}\leq 1-\frac{\mu\eta}{3}$ for $\mu\eta<1$.

C.3. Proof of Theorem 3

Let $\eta \leq \frac{1}{4L_f}$. From Lemma 10, we have

$$\begin{aligned}
& \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2 \\
\leq\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\mathbb{E}\|e^k\|^2+(1+\eta\mu)\mathbb{E}\|e^{k+1}\|^2+4\eta^2\mathbb{E}\|g^k\|^2 \\
\leq\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\mathbb{E}\|e^k\|^2+\frac{5}{4}\mathbb{E}\|e^{k+1}\|^2+4\eta^2\mathbb{E}\|g^k\|^2 \\
\overset{Lemma13}{\leq}\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}(P(x^*)-P(x^{k+1}))+\frac{9}{4}\mathbb{E}\|e^k\|^2+\frac{5(1-\delta)\delta}{2n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+4\eta^2\mathbb{E}\|g^k\|^2 \\
& +5(1-\delta)\eta^2\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)] \\
\overset{(15)}{\leq}\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})]+\frac{9}{4}\mathbb{E}\|e^k\|^2+\frac{5(1-\delta)\delta}{2n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\left(5(1-\delta)\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)+16L_f+\frac{16L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}$$

Then from Lemmas 12 and 13, we have

$$\begin{aligned}
& \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{9}{\delta}\mathbb{E}\|e^{k+1}\|^2+\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\overset{lemma13}{\leq}\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})]+\left(1-\frac{\delta}{2}+\frac{\delta}{4}\right)\frac{9}{\delta}\mathbb{E}\|e^k\|^2 \\
& +\left(\frac{18(1-\delta)}{n^2}+\frac{5(1-\delta)\delta}{2n^2}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
& +\left((1-\delta)\left(5+\frac{36}{\delta}\right)\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)+16L_f+\frac{16L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)] \\
\overset{Lemma12}{\leq}\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})]+\left(1-\frac{\delta}{4}\right)\frac{9}{\delta}\mathbb{E}\|e^k\|^2+\left(1-\frac{\delta}{4}\right)\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\left(\frac{41(1-\delta)}{\delta}\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)+16L_f+\frac{16L}{n}+\frac{336(1-\delta)}{\delta n}\left(\frac{\bar{L}}{\delta}+L\right)\right) \\
& \cdot\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)] \\
=\ & \mathbb{E}\|\tilde{x}^k-x^*\|^2+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})]+\left(1-\frac{\delta}{4}\right)\frac{9}{\delta}\mathbb{E}\|e^k\|^2+\left(1-\frac{\delta}{4}\right)\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\left(\frac{(1-\delta)}{\delta}\left(\frac{82L_f}{\delta}+\frac{336\bar{L}}{\delta n}+\frac{459L}{n}\right)+16L_f+\frac{16L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)+P(w^k)-P(x^*)].
\end{aligned}$$

Combining the above inequality and (19), we can obtain

$$
\begin{aligned}
& \mathbb{E}[\Phi_2^{k+1}] \\
= \quad & \left(1+\frac{\eta\mu}{2}\right)\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{9}{\delta}\mathbb{E}\|e^{k+1}\|^2+\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^{k+1}\|^2 \\
& +\frac{2\eta^2}{p}\left(\frac{(1-\delta)}{\delta}\left(\frac{82L_f}{\delta}+\frac{336\bar{L}}{\delta n}+\frac{459L}{n}\right)+16L_f+\frac{16L}{n}\right)\mathbb{E}[P(w^{k+1})-P(x^*)] \\
\leq \quad & \mathbb{E}\|\tilde{x}^k-x^*\|^2+\left(1-\frac{\delta}{4}\right)\frac{9}{\delta}\mathbb{E}\|e^k\|^2+\left(1-\frac{\delta}{4}\right)\frac{84(1-\delta)}{\delta n^2}\sum_{\tau=1}^{n}\mathbb{E}\|e_\tau^k\|^2+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})] \\
& +\frac{2\eta^2}{p}\left(\frac{(1-\delta)}{\delta}\left(\frac{82L_f}{\delta}+\frac{336\bar{L}}{\delta n}+\frac{459L}{n}\right)+16L_f+\frac{16L}{n}\right)\left(1-\frac{p}{2}\right)\mathbb{E}[P(w^k)-P(x^*)] \\
& +\left(\frac{(1-\delta)}{\delta}\left(\frac{246L_f}{\delta}+\frac{1008\bar{L}}{\delta n}+\frac{1377L}{n}\right)+48L_f+\frac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)] \\
\leq \quad & \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)\mathbb{E}[\Phi_2^k]+2\eta\mathbb{E}[P(x^*)-P(x^{k+1})] \\
& +\left(\frac{(1-\delta)}{\delta}\left(\frac{246L_f}{\delta}+\frac{1008\bar{L}}{\delta n}+\frac{1377L}{n}\right)+48L_f+\frac{48L}{n}\right)\eta^2\mathbb{E}[P(x^k)-P(x^*)],
\end{aligned}
$$

where we use $\left(1+\frac{\eta\mu}{2}\right)^{-1}\leq 1-\frac{\mu\eta}{3}$ for $\mu\eta<1$.

C.4. Proof of Theorem 4

Let $\eta\leq\frac{1}{4L_f}$. From Theorem 2, we have

$$
\begin{aligned}
& \mathbb{E}[\Phi_1^k] \\
\leq \quad & \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)\mathbb{E}[\Phi_1^{k-1}]+2\eta\mathbb{E}[P(x^*)-P(x^k)] \\
& +\left(\frac{123(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+48L_f+\frac{48L}{n}\right)\eta^2\mathbb{E}[P(x^{k-1})-P(x^*)] \\
\leq \quad & \left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)^k\Phi_1^0-2\eta\sum_{i=1}^{k}\left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{k-i}\mathbb{E}[P(x^i)-P(x^*)] \\
& +\left(\frac{123(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+48L_f+\frac{48L}{n}\right)\eta^2\sum_{i=0}^{k-1}\left(1-\min\left\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\right\}\right)^{k-1-i}\mathbb{E}[P(x^i)-P(x^*)] \\
= \quad & \frac{1}{w_k}\Phi_1^0-\frac{2\eta}{w_k}\sum_{i=1}^{k}w_i\mathbb{E}[P(x^i)-P(x^*)] \\
& +\left(\frac{123(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+48L_f+\frac{48L}{n}\right)\frac{w_1\eta^2}{w_k}\sum_{i=0}^{k-1}w_i\mathbb{E}[P(x^i)-P(x^*)] \\
\leq \quad & \frac{1}{w_k}\Phi_1^0-\frac{2\eta}{w_k}\sum_{i=1}^{k}w_i\mathbb{E}[P(x^i)-P(x^*)] \\
& +\left(\frac{135(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+53L_f+\frac{53L}{n}\right)\frac{\eta^2}{w_k}\sum_{i=0}^{k}w_i\mathbb{E}[P(x^i)-P(x^*)],
\end{aligned}
$$

where we use $w_1 \leq \frac{12}{11}$ in the last inequality. Rearranging the above inequality, we can get

$$\begin{aligned}
&\frac{2}{w_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \\
\leq\ & \frac{1}{\eta w_k}\Phi_1^0 - \frac{1}{\eta}\mathbb{E}[\Phi_1^k] + \frac{2(P(x^0) - P(x^*))}{w_k} \\
&+ \left(\frac{135(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) + 53L_f + \frac{53L}{n}\right)\frac{\eta}{w_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \\
\leq\ & \frac{1}{\eta w_k}\Phi_1^0 + \frac{2(P(x^0) - P(x^*))}{w_k} + \left(\frac{135(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) + 53L_f + \frac{53L}{n}\right)\frac{\eta}{w_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)].
\end{aligned}$$

Hence, if

$$\eta \leq \frac{1}{\left(\frac{135(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) + 53L_f + \frac{53L}{n}\right)} = \frac{\delta^2}{135(1-\delta)(\bar{L} + L\delta) + 53L_f\delta^2 + 53L\delta^2/n},$$

then

$$\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \leq \frac{1}{\eta}\Phi_1^0 + 2(P(x^0) - P(x^*)).$$

Furthermore, since

$$W_k = \sum_{i=0}^{k} w_i = \frac{1 - \frac{1}{(1-\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}}{1 - \frac{1}{1-\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\}}} = \frac{1 - (1 - \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}{\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\}(1 - \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^k},$$

we can get

$$\begin{aligned}
&\frac{1}{W_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \\
\leq\ & \frac{1}{W_k}\left(\frac{1}{\eta}\Phi_1^0 + 2(P(x^0) - P(x^*))\right) \\
=\ & \frac{\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\}}{1 - (1 - \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(\frac{1}{\eta}\Phi_1^0 + 2(P(x^0) - P(x^*))\right)\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.
\end{aligned}$$

From the definition of Φ_1^k and $e_\tau^0 = 0$, we have

$$\begin{aligned}
&\min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\} \cdot \frac{1}{\eta}\Phi_1^0 \\
=\ & \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\}\left(\frac{1}{\eta}\left(1 + \frac{\mu\eta}{2}\right)\|x^0 - x^*\|^2 + \frac{2\eta}{p}\left(\frac{41(1-\delta)}{\delta}\left(\frac{\bar{L}}{\delta} + L\right) + 16L_f + \frac{16L}{n}\right)(P(x^0) - P(x^*))\right) \\
\leq\ & \frac{\mu}{3}\left(1 + \frac{\mu\eta}{2}\right)\|x^0 - x^*\|^2 + \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\} \cdot \frac{2}{3p}(P(x^0) - P(x^*)) \\
\leq\ & \frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{3}(P(x^0) - P(x^*)).
\end{aligned}$$

Therefore, we arrive at

$$\frac{1}{W_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \leq \frac{\frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))}{1 - (1 - \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

For $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^{k} w_i x^i$, from the convexity of P, we have

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \frac{1}{W_k}\sum_{i=0}^{k} w_i \mathbb{E}[P(x^i) - P(x^*)] \leq \frac{\frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))}{1 - (1 - \min\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

If we choose $\eta = \frac{\delta^2}{135(1-\delta)(\bar{L}+L\delta)+53L_f\delta^2+53L\delta^2/n}$, then in order to guarantee $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$, we first let

$$\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k+1} \leq \frac{1}{2},$$

which implies that

$$\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \left(\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)\right)\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

Hence, when $\epsilon \leq \frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))$, $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$ as long as

$$\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k \leq \frac{\epsilon}{\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)},$$

which is equivalent to

$$k \geq \frac{1}{-\ln(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\})} \ln\left(\frac{\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)}{\epsilon}\right).$$

Since $-\ln(1-x) \geq x$ for $x \in [0, 1)$, we have $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \leq \epsilon$ as long as

$$k \geq O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{(1-\delta)\bar{L}}{\delta^2\mu} + \frac{(1-\delta)L}{\delta\mu} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right)\ln\left(\frac{\mu\|x^0 - x^*\|^2 + P(x^0) - P(x^*)}{\epsilon}\right)\right).$$

C.5. Proof of Theorem 5

From $\frac{\bar{L}}{n} \leq L_f$, the proof is same as that of Theorem 4.

Appendix D. Proofs for the smooth case

D.1. Lemma

Thanks to the following lemma, we can get better results than the composite case. The main difference between Lemma 10 and Lemma 14 is that there is an additional stepsize η before $\mathbb{E}\|e^k\|^2$. The following lemma is similar to Lemma 7 in [14]. However, for completeness, we give the proof.

Lemma 14 *If $\eta \leq \frac{1}{4L_f + 8L/n}$, then*

$$\begin{aligned}\mathbb{E}\|\tilde{x}^{k+1} - x^*\|^2 \quad \leq \quad & \left(1 - \tfrac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k - x^*\|^2 - \tfrac{\eta}{2}\mathbb{E}[f(x^k) - f(x^*)] \\ & + 3L_f\eta\mathbb{E}\|e^k\|^2 + \tfrac{4L}{n}\eta^2\mathbb{E}[f(w^k) - f(x^*)].\end{aligned}$$

Proof

Since $\psi = 0$, we have $\tilde{x}^{k+1} = \tilde{x}^k - \eta(g^k + \nabla f(w^k))$. Hence

$$\begin{aligned}& \mathbb{E}\|\tilde{x}^{k+1} - x^*\|^2 \\ = \quad & \mathbb{E}\|\tilde{x}^k - x^* - \eta(g^k + \nabla f(w^k))\|^2 \\ = \quad & \mathbb{E}\|\tilde{x}^k - x^*\|^2 - 2\eta\mathbb{E}\langle\tilde{x}^k - x^*, \nabla f(x^k)\rangle + \eta^2\mathbb{E}\|g^k + \nabla f(w^k)\|^2 \\ = \quad & \mathbb{E}\|\tilde{x}^k - x^*\|^2 - 2\eta\mathbb{E}\langle x^k - x^*, \nabla f(x^k)\rangle + 2\eta\langle x^k - \tilde{x}^k, \nabla f(x^k)\rangle + \eta^2\mathbb{E}\|g^k + \nabla f(w^k)\|^2 \\ \leq \quad & \mathbb{E}\|\tilde{x}^k - x^*\|^2 - 2\eta\mathbb{E}(f(x^k) - f(x^*)) - \mu\eta\mathbb{E}\|x^k - x^*\|^2 + 2\eta\mathbb{E}\langle e^k, \nabla f(x^k)\rangle + \eta^2\mathbb{E}\|g^k + \nabla f(w^k)\|^2,\end{aligned}$$

where the last inequality comes from the μ-strongly convexity of f.

For $\|x^k - x^*\|^2$, we have

$$\|\tilde{x}^k - x^*\|^2 \leq 2\|x^k - x^*\|^2 + 2\|e^k\|^2.$$

For $2\langle e^k, \nabla f(x^k)\rangle$, we have

$$2\langle e^k, \nabla f(x^k)\rangle \leq \frac{1}{2L_f}\|\nabla f(x^k)\|^2 + 2L_f\|e^k\|^2 \leq f(x^k) - f(x^*) + 2L_f\|e^k\|^2.$$

Thus, we arrive at

$$\begin{aligned}
&\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2 \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\eta\mathbb{E}(f(x^k)-f(x^*))+(2L_f+\mu)\eta\mathbb{E}\|e^k\|^2+\eta^2\mathbb{E}\|g^k+\nabla f(w^k)\|^2 \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\eta\mathbb{E}(f(x^k)-f(x^*))+3L_f\eta\mathbb{E}\|e^k\|^2+\eta^2\mathbb{E}\|g^k+\nabla f(w^k)\|^2.
\end{aligned}$$

Finally, for $\mathbb{E}\|g^k+\nabla f(w^k)\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|g^k+\nabla f(w^k)\|^2 &= \mathbb{E}\|g^k+\nabla f(w^k)-\nabla f(x^k)+\nabla f(x^k)-\nabla f(x^*)\|^2 \\
&= \mathbb{E}\|g^k+\nabla f(w^k)-\nabla f(x^k)\|^2+\mathbb{E}\|\nabla f(x^k)-\nabla f(x^*)\|^2 \\
&\leq \mathbb{E}\|g^k+\nabla f(w^k)-\nabla f(x^k)\|^2+2L_f\mathbb{E}(f(x^k)-f(x^*)) \\
&\overset{(16)}{\leq} \left(2L_f+\frac{4L}{n}\right)\mathbb{E}[f(x^k)-f(x^*)]+\frac{4L}{n}\mathbb{E}[f(w^k)-f(x^*)].
\end{aligned}$$

Thereofore,

$$\begin{aligned}
\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2 \leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\eta\left(1-\left(2L_f+\frac{4L}{n}\right)\eta\right)\mathbb{E}[f(x^k)-f(x^*)] \\
&+3L_f\eta\mathbb{E}\|e^k\|^2+\frac{4L}{n}\eta^2\mathbb{E}[f(w^k)-f(x^*)].
\end{aligned}$$

By choosing $\eta\leq\frac{1}{4L_f+8L/n}$, we can get the reslut. ■

D.2. Proof of Theorem 6

Let $\eta\leq\frac{1}{4L_f+8L/n}$. From Lemma 14, Lemma 12, and $\|e^k\|^2\leq\frac{1}{n}\sum_{\tau=1}^n\|e_\tau^k\|^2$, we can obtain

$$\begin{aligned}
&\mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{12L_f\eta}{n\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\frac{\eta}{2}\mathbb{E}[f(x^k)-f(x^*)]+\frac{3L_f\eta}{n}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{4L}{n}\eta^2\mathbb{E}[f(w^k)-f(x^*)] \\
&+\frac{12L_f\eta}{n\delta}\left(1-\frac{\delta}{2}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)\mathbb{E}[f(x^k)-f(x^*)+f(w^k)-f(x^*)] \\
=\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2+\frac{12L_f\eta}{n\delta}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2-\frac{\eta}{2}\left(1-\frac{96(1-\delta)L_f\eta^2}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)\right) \\
&\cdot\mathbb{E}[f(x^k)-f(x^*)]+\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+\frac{4L\eta^2}{n}\right)\mathbb{E}[f(w^k)-f(x^*)].
\end{aligned}$$

Then from (19), we have

$$\begin{aligned}
&\mathbb{E}[\Phi_3^{k+1}] \\
=\ & \mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{12L_f\eta}{n\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2+\frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+\frac{4L\eta^2}{n}\right)\mathbb{E}[f(w^{k+1})-f(x^*)] \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2+\frac{12L_f\eta}{n\delta}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
&+\frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+\frac{4L\eta^2}{n}\right)\left(1-\frac{p}{2}\right)\mathbb{E}[f(w^k)-f(x^*)] \\
&-\frac{\eta}{2}\left(1-\frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)-\frac{16L\eta}{n}\right)\mathbb{E}[f(x^k)-f(x^*)] \\
\leq\ & \left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)\mathbb{E}[\Phi_3^k]-\frac{\eta}{2}\left(1-\frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)-\frac{16L\eta}{n}\right)\mathbb{E}[f(x^k)-f(x^*)].
\end{aligned}$$

D.3. Proof of Theorem 7

Let $\eta \leq \frac{1}{4L_f+8L/n}$. From Lemma 14, we have

$$
\begin{aligned}
& \mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{12L_f\eta}{\delta}\mathbb{E}\|e^{k+1}\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\frac{\eta}{2}\mathbb{E}[f(x^k)-f(x^*)]+3L_f\eta\mathbb{E}\|e^k\|^2+\frac{4L}{n}\eta^2\mathbb{E}[f(w^k)-f(x^*)] \\
& +\frac{12L_f\eta}{\delta}\mathbb{E}\|e^{k+1}\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
\overset{Lemma\ 13}{\leq}\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\frac{\eta}{2}\mathbb{E}[f(x^k)-f(x^*)]+\frac{12L_f\eta}{\delta}\left(1-\frac{\delta}{2}+\frac{\delta}{4}\right)\mathbb{E}\|e^k\|^2 \\
& +\frac{4L}{n}\eta^2\mathbb{E}[f(w^k)-f(x^*)]+\frac{24(1-\delta)L_f\eta}{n^2}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
& +\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta}+\frac{3L}{n}\right)\mathbb{E}[f(x^k)-f(x^*)+f(w^k)-f(x^*)] \\
\overset{Lemma\ 12}{\leq}\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2-\frac{\eta}{2}\mathbb{E}[f(x^k)-f(x^*)]+\frac{12L_f\eta}{\delta}\left(1-\frac{\delta}{4}\right)\mathbb{E}\|e^k\|^2 \\
& +\frac{4L}{n}\eta^2\mathbb{E}[f(w^k)-f(x^*)]+\frac{96(1-\delta)L_f\eta}{n^2\delta}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta}+\frac{3L}{n}+\frac{8\bar{L}}{n\delta}+\frac{8L}{n}\right)\mathbb{E}[f(x^k)-f(x^*)+f(w^k)-f(x^*)] \\
=\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2+\frac{12L_f\eta}{\delta}\left(1-\frac{\delta}{4}\right)\mathbb{E}\|e^k\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& -\frac{\eta}{2}\left(1-\frac{96(1-\delta)L_f\eta^2}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)\right)\mathbb{E}[f(x^k)-f(x^*)] \\
& +\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)+\frac{4L\eta^2}{n}\right)\mathbb{E}[f(w^k)-f(x^*)].
\end{aligned}
$$

Then from (19), we can get

$$
\begin{aligned}
& \mathbb{E}[\Phi_4^{k+1}] \\
=\ & \mathbb{E}\|\tilde{x}^{k+1}-x^*\|^2+\frac{12L_f\eta}{\delta}\mathbb{E}\|e^{k+1}\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\sum_{\tau=1}^n\mathbb{E}\|e_\tau^{k+1}\|^2 \\
& +\frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)+\frac{4L\eta^2}{n}\right)\mathbb{E}[f(w^{k+1})-f(x^*)] \\
\leq\ & \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\|\tilde{x}^k-x^*\|^2+\frac{12L_f\eta}{\delta}\left(1-\frac{\delta}{4}\right)\mathbb{E}\|e^k\|^2+\frac{96(1-\delta)L_f\eta}{n^2\delta}\left(1-\frac{\delta}{4}\right)\sum_{\tau=1}^n\mathbb{E}\|e_\tau^k\|^2 \\
& +\frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^3}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)+\frac{4L\eta^2}{n}\right)\left(1-\frac{p}{2}\right)\mathbb{E}[f(w^k)-f(x^*)] \\
& -\frac{\eta}{2}\left(1-\frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)-\frac{16L\eta}{n}\right)\mathbb{E}[f(x^k)-f(x^*)] \\
\leq\ & \left(1-\min\left\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\right\}\right)\mathbb{E}[\Phi_4^k]-\frac{\eta}{2}\left(1-\frac{288(1-\delta)L_f\eta^2}{\delta}\left(\frac{2L_f}{\delta}+\frac{11L}{n}+\frac{8\bar{L}}{n\delta}\right)-\frac{16L\eta}{n}\right)\mathbb{E}[f(x^k)-f(x^*)].
\end{aligned}
$$

D.4. Proof of Theorem 8

Let $\eta \leq \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_fL}}\right\}$. Then we have

$$\frac{16L\eta}{n} \leq \frac{2}{3}, \quad \frac{288(1-\delta)L_f\bar{L}\eta^2}{\delta^2} \leq \frac{1}{9}, \quad \text{and} \quad \frac{288(1-\delta)L_fL\eta^2}{\delta} \leq \frac{1}{9}.$$

Hence, from Theorem 6, we have

$$\begin{aligned}
\mathbb{E}[\Phi_3^{k+1}] &\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)\mathbb{E}[\Phi_3^k] - \frac{\eta}{18}\mathbb{E}[f(x^k)-f(x^*)] \\
&\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k+1}\Phi_3^0 - \frac{\eta}{18}\sum_{i=0}^{k}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k-i}\mathbb{E}[f(x^i)-f(x^*)] \\
&\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k}\Phi_3^0 - \frac{\eta}{18}\sum_{i=0}^{k}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k-i}\mathbb{E}[f(x^i)-f(x^*)] \\
&= \frac{1}{w_k}\Phi_3^0 - \frac{\eta}{18w_k}\sum_{i=0}^{k} w_i\mathbb{E}[f(x^i)-f(x^*)],
\end{aligned}$$

which implies that

$$\frac{1}{W_k}\sum_{i=0}^{k} w_i\mathbb{E}[f(x^i)-f(x^*)] \leq \frac{18}{\eta W_k}\Phi_3^0 - \frac{18w_k}{\eta W_k}\mathbb{E}[\Phi_3^{k+1}] \leq \frac{18}{\eta W_k}\Phi_3^0.$$

Then, from

$$W_k = \sum_{i=0}^{k} w_i = \frac{1-\frac{1}{(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}}{1-\frac{1}{1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\}}} = \frac{1-(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}{\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\}(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k}}, \tag{20}$$

we can obtain

$$\frac{1}{W_k}\sum_{i=0}^{k} w_i\mathbb{E}[f(x^i)-f(x^*)] \leq \frac{\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\}}{1-(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}} \cdot \frac{18}{\eta}\Phi_3^0\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k}.$$

From the definition of Φ_3^k and $e_\tau^0 = 0$, we have

$$\begin{aligned}
&\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\cdot\frac{1}{\eta}\Phi_3^0 \\
=\ &\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\left(\frac{1}{\eta}\|x^0-x^*\|^2 + \frac{2}{p}\left(\frac{48(1-\delta)L_f\eta^2}{\delta}\left(\frac{\bar{L}}{\delta}+L\right)+\frac{4L\eta}{n}\right)[f(x^0)-f(x^*)]\right) \\
\leq\ &\frac{\mu}{2}\|x^0-x^*\|^2 + \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\frac{2}{p}\left(\frac{1}{54}+\frac{1}{54}+\frac{1}{6}\right)[f(x^0)-f(x^*)] \\
\leq\ &\frac{\mu}{2}\|x^0-x^*\|^2 + \frac{1}{2}[f(x^0)-f(x^*)].
\end{aligned}$$

Therefore, we can get

$$\frac{1}{W_k}\sum_{i=0}^{k} w_i\mathbb{E}[f(x^i)-f(x^*)] \leq \frac{9\mu\|x^0-x^*\|^2 + 9(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k}.$$

For $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^{k} w_i x^i$, from the convexity of f anf the above inequality, we have

$$\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \frac{9\mu\|x^0-x^*\|^2 + 9(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k}.$$

If we choose $\eta = \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_fL}}\right\}$, then in order to guarantee $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$, we first let

$$\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k+1} \leq \frac{1}{2},$$

which implies that

$$\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq 18\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

Hence, when $\epsilon \leq 9\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)$, $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$ as long as

$$\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k \leq \frac{\epsilon}{18\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)},$$

which is equivalen to

$$k \geq \frac{1}{-\ln(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\})}\ln\left(\frac{18\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right).$$

Since $-\ln(1-x) \geq x$ for $x \in [0,1)$, we have $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$ as long as

$$k \geq O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{\sqrt{(1-\delta)L_f\bar{L}}}{\delta}+\frac{\sqrt{(1-\delta)L_fL}}{\sqrt{\delta}}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\ln\left(\frac{18\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right).$$

D.5. Proof of Theorem 9

Let $\eta \leq \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_fL}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f\bar{L}}}\right\}$. Then we have

$$\frac{288(1-\delta)L_f\eta^2}{\delta}\cdot\frac{2L_f}{\delta} \leq \frac{1}{12}, \quad \frac{288(1-\delta)L_f\eta^2}{\delta}\cdot\frac{11L}{n} \leq \frac{1}{6}, \quad \frac{288(1-\delta)L_f\eta^2}{\delta}\cdot\frac{8\bar{L}}{n\delta} \leq \frac{1}{6}, \text{ and } \frac{16L\eta}{n} \leq \frac{1}{2}.$$

Therefore, from Theorem 7, we have

$$\begin{aligned}
\mathbb{E}[\Phi_4^{k+1}] &\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)\mathbb{E}[\Phi_4^k] - \frac{\eta}{24}\mathbb{E}[f(x^k)-f(x^*)] \\
&\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k+1}\Phi_4^0 - \frac{\eta}{24}\sum_{i=0}^{k}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k-i}\mathbb{E}[f(x^i)-f(x^*)] \\
&\leq \left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k}\Phi_4^0 - \frac{\eta}{24}\sum_{i=0}^{k}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{k-i}\mathbb{E}[f(x^i)-f(x^*)] \\
&= \frac{1}{w_k}\Phi_4^0 - \frac{\eta}{24w_k}\sum_{i=0}^{k}w_i\mathbb{E}[f(x^i)-f(x^*)],
\end{aligned}$$

Then same as the proof of Theorem 8, we have

$$\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \frac{12\mu\|x^0-x^*\|^2+12(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\})^{k+1}}\left(1-\min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k,$$

and if we choose $\eta = \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_fL}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f\bar{L}}}\right\}$, then $\mathbb{E}[f(\bar{x}^k)-f(x^*)] \leq \epsilon$ with $\epsilon \leq 12\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)$ as long as

$$\begin{aligned}
k \geq\ & O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{\sqrt{(1-\delta)L_f}}{\mu\delta}+\frac{\sqrt{(1-\delta)L_f\bar{L}}}{\mu\sqrt{n\delta}}+\frac{\sqrt{(1-\delta)L_fL}}{\mu\sqrt{n\delta}}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\right. \\
&\left.\ln\left(\frac{24\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right),
\end{aligned}$$

which is equivalent to

$$k \geq O\left(\left(\frac{1}{\delta}+\frac{1}{p}+\frac{\sqrt{(1-\delta)}L_f}{\mu\delta}+\frac{L_f}{\mu}+\frac{L}{n\mu}\right)\ln\left(\frac{24\left(\mu\|x^0-x^*\|^2+f(x^0)-f(x^*)\right)}{\epsilon}\right)\right),$$

since $\frac{\bar{L}}{n} \leq L_f$, and

$$2\sqrt{\frac{(1-\delta)L_f L}{n\delta}} \leq \frac{\sqrt{1-\delta}L_f}{\delta}+\frac{\sqrt{1-\delta}L}{n} \leq \frac{\sqrt{1-\delta}L_f}{\delta}+\frac{L}{n}.$$