

Supplementary Material to

Communication-efficient Decentralized

Stochastic Projection-free Learning via

Compressed Gradients

APPENDIX

This document contains proofs of Lemmas 1 to 5 and Theorems 1 and 2 in the paper that establish the convergence rate of CE-DSFW summarized in Algorithm 1.

Algorithm 1 The CE-DSFW Algorithm

- 1: Initialize $\mathbf{x}_0^{(i)} \in \mathcal{C}$, $\mathbf{d}_{-1}^{(i)} = \mathbf{0}$, $\mathbf{y}_{-1}^{(i)} = \mathbf{1}$, $\forall i \in [n]$.
 - 2: **for** each node $i \in [n]$ and $k = 0, 1, \dots$ **do**
 - 3: Sample $\zeta_k^{(i)}$ at random according to $\mathcal{D}^{(i)}$
 - 4: $\mathbf{a}_k^{(i)} = \mathbf{a}_{k-1}^{(i)} + \eta_k \text{Comp} \left(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} \right)$
 - 5: $\mathbf{z}_k^{(i)} = \sum_{j=1}^n W_{ij} \mathbf{a}_k^{(j)}$
 - 6: $\mathbf{y}_k^{(i)} = \sum_{j=1}^n W_{ij} \mathbf{y}_{k-1}^{(j)}$
 - 7: $\mathbf{c}_k^{(i)} = \mathbf{z}_k^{(i)} \odot \mathbf{y}_k^{(i)}$
 - 8: $\mathbf{d}_k^{(i)} = (1 - \eta_{k-1}) \mathbf{d}_{k-1}^{(i)} + \eta_{k-1} \mathbf{c}_k^{(i)}$
 - 9: $\mathbf{s}_k^{(i)} = \underset{\mathbf{s} \in \mathcal{C}}{\text{argmin}} \langle \mathbf{d}_k^{(i)}, \mathbf{s} \rangle$
 - 10: $\mathbf{x}_{k+1}^{(i)} = \mathbf{x}_k^{(i)} + \gamma_k (\mathbf{s}_k^{(i)} - \mathbf{x}_k^{(i)})$
 - 11: $\gamma_k \leftarrow \frac{2}{(k+2)^\beta}$ and η_k .
 - 12: **end for**
-

Before proceeding with the derivations, we present the assumptions and some backgrounds.

Assumption 1. L -smooth local objectives. There exists a constant $L > 0$ such that

$$\|\nabla g(\mathbf{y}) - \nabla g(\mathbf{x})\| \leq L \|\mathbf{y} - \mathbf{x}\| \quad (1)$$

or equivalently

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \forall \mathbf{y}, \mathbf{x} \in \mathcal{R}^d. \quad (2)$$

We assume that functions $f(\cdot)$ and $f(\cdot, \cdot)$ satisfy the smoothness assumption.

Assumption 2. Diameter of \mathcal{C} . The constraint set \mathcal{C} is convex and compact with diameter D , i.e.,

$$\|\mathbf{y} - \mathbf{x}\| \leq D, \quad \forall \mathbf{y}, \mathbf{x} \in \mathcal{C} \subset \mathcal{R}^d.$$

Assumption 3. Bounded local stochastic gradients. Local stochastic gradients $\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)})$ have bounded variance

$$\mathbb{E} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)})\|^2 \right] \leq \nu^2 \quad \forall i \in [n], \quad (3)$$

where $\nabla f(\mathbf{x}_k^{(i)})$ is the full (non-stochastic) gradient evaluated at $\mathbf{x}_k^{(i)}$.

Assumption 4. Unbiasedness and variance of compression operator. For any vector $\mathbf{x} \in \mathcal{R}^d$, the compression operator $\text{Comp}(\cdot)$ satisfies

$$\mathbb{E}[\text{Comp}(\mathbf{x})|\mathbf{x}] = \mathbf{x}, \quad \mathbb{E}\|\mathbf{x} - \text{Comp}(\mathbf{x})\|^2 \leq \delta^2. \quad (4)$$

In the following lemma, we give the convergence guarantees for the push-sum method employed to reach a consensus. Consider applying push-sum to the parameter $vp_{k-1}^{(i)}$ to obtain the $vp_k^{(i)}$, then the parameter is guaranteed to converge to the consensus $\bar{\mathbf{p}} = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_k^{(i)}$ and is formally stated below.

Lemma A1. If the update from the push-sum algorithm $\mathbf{p}_k^{(i)} = \mathbf{a}_t^{(i)} \odot \mathbf{y}_t^{(i)}$ converges to a consensus, then the consensus is the average vector $\bar{\mathbf{p}} = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_k^{(i)}$ [1][Theorem 3.1].

Given the push-sum converges to the consensus we bound the error at each iterate in the following lemma.

Lemma A2. For a directed network characterized by a row stochastic connectivity matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ the distance of the parameter to the consensus can be bounded as

$$\|\mathbf{p}_k^{(i)} - \bar{\mathbf{p}}\|_2 \leq \sqrt{n} \|\mathbf{p}_0^{(i)}\|_\infty u(k) \quad (5)$$

where $u(k)$ is a non-increasing sequence.

Proof. Using [1][Lemma 4.1], we have

$$\|\mathbf{p}_k^{(i)} - \bar{\mathbf{p}}\|_\infty \leq \|\mathbf{p}_0^{(i)}\|_\infty u(k).$$

The bound in Equation 5 follows from the norm inequality $\|\mathbf{p}\|_2 \leq \sqrt{n} \|\mathbf{p}\|_\infty$ for $\mathbf{p} \in \mathbb{R}^n$. \square

Equation 5 bounds the error at each iterate obtained from the push-sum algorithm. Now we want to bound the improvement at each round given the worst-case performance until the previous iterate, which is formally stated below.

Lemma A3. Consider the parameter $\mathbf{p}_k^{(i)}$ at each node $i \in [n]$, then we have

$$\|\mathbf{p}_{k+1}^{(i)} - \bar{\mathbf{p}}\|_2 \leq \rho \|\mathbf{p}_k^{(i)} - \bar{\mathbf{p}}\|_2,$$

where $\rho \leq 1$ is the improvement factor at each gossip step.

Proof. It immediately follows from Lemma A2 that

$$\frac{\|\mathbf{p}_{k+1}^{(i)} - \bar{\mathbf{p}}\|_2}{\|\mathbf{p}_k^{(i)} - \bar{\mathbf{p}}\|_2} \leq \frac{u(k+1)}{u(k)} =: \rho_k$$

where $\rho_k = \frac{u(k+1)}{u(k)} \leq 1$ as $u(\cdot)$ is a non-increasing function and $\rho = \max_k \rho_k$. \square

The constant ρ depends on the network topology and is an indicator for improvement at each iterate. For undirected networks, the constant reduces to the second largest eigenvalue of the adjacency matrix \mathbf{W} , i.e., $\rho = \lambda_2(\mathbf{W})$ [2].

Using Assumption 1 and 3, we introduce a non-standard smoothness assumption for the objective function as in the next lemma.

Lemma A4. There exists a constant $\bar{L} > 0$, such that

$$\mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)})\| \leq \bar{L}D, \quad (6)$$

where $\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)})$, is the stochastic gradient evaluated at $\mathbf{x}_k^{(i)}$ and $\nabla f(\mathbf{x}_k^{(j)})$, the deterministic gradient evaluated at $\mathbf{x}_k^{(j)}$.

Proof. Consider $\nabla f(\mathbf{x}_k^{(i)})$, the deterministic gradient at the i th node in the k th iterate, then we have

$$\begin{aligned} \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)})\| &= \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)}) + \nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)})\| \\ &\stackrel{(a)}{\leq} \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)})\| + \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)})\| \\ &\stackrel{(b)}{\leq} \nu + \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)})\| \\ &\stackrel{(c)}{\leq} \nu + L \mathbb{E}\|\mathbf{x}_k^{(i)} - \mathbf{x}_k^{(j)}\| \\ &\leq \bar{L}D, \end{aligned}$$

where (a) follows from triangle inequality, (b) follows as

$$\mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)})\| = \mathbb{E}\sqrt{\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)})\|^2} \leq \sqrt{\mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(i)})\|^2} \leq \nu,$$

(c) follows from Assumption 1, and final expression follows from Assumption 2. \square

Now using Assumption 3, we can bound the variance of the local stochastic gradient at each node after a gossip step as in the next lemma.

Lemma A5. Uncompressed local stochastic gradients after one gossip step, i.e., $\mathbf{z}_k^{(i)} = \sum_{j \in [n]} w_{ij} \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)})$, has bounded variance

$$\mathbb{E}\left[\|\mathbf{z}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|^2\right] \leq \sigma^2 \quad \forall i \in [n], \quad (7)$$

where σ is related to ν and the network topology.

Proof. The average stochastic gradient (averaged over n nodes) at iteration k is $\frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)})$. Then, we have the following

$$\begin{aligned}
\mathbb{E} \left[\|\mathbf{z}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|^2 \right] &= \mathbb{E} \left[\left\| \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) + \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)}) \right\|^2 \right] \\
&= \mathbb{E} \left[\left\| \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)}) \right\|^2 \right] \\
&\quad + 2 \mathbb{E} \left[\left\langle \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}), \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)}) \right\rangle \right] \\
&\stackrel{(a)}{\leq} 2 \mathbb{E} \left[\left\| \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right] + 2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)}) \right\|^2 \right] \\
&\stackrel{(b)}{\leq} \underbrace{2 \mathbb{E} \left[\left\| \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right]}_{:=T_1} \\
&\quad + \underbrace{2 \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)}) \right\|^2 \right]}_{:=T_2},
\end{aligned}$$

where (a) follows from the Young's inequality $2\langle \mathbf{k}, \mathbf{l} \rangle \leq \eta \|\mathbf{k}\|^2 + \frac{1}{\eta} \|\mathbf{l}\|^2$ for all $\eta > 0$, (we use $\mathbf{k} = \mathbf{z}_k^{(i)} - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)})$, $\mathbf{l} = \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \nabla f(\mathbf{x}_k^{(i)})$, and $\eta = 1$) and (b) follows from the Jensen's inequality $l(\sum_i \alpha_i \mathbf{x}_i) \leq \sum_i \alpha_i l(\mathbf{x}_i)$ with $l(\cdot)$ being a convex function and $\sum_{i=1}^n \alpha_i = 1$.

Using Lemma A4, we have

$$T_2 \leq \bar{L}^2 D^2.$$

Recall that $\mathbf{z}_k^{(i)}$ is the uncompressed stochastic gradient after one gossip step and $\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)})$ is the local stochastic gradient before gossip at node i . From Lemma A3, the first term satisfies

$$\begin{aligned}
T_1 &\leq \rho^2 \mathbb{E} \left[\left\| \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right] \\
&\stackrel{(a)}{=} \rho^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right] \\
&\stackrel{(b)}{\leq} \frac{\rho^2}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) \right\|^2 \right] \\
&\stackrel{(c)}{\leq} \frac{L^2 \rho^2}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{x}_k^{(i)} - \mathbf{x}_k^{(j)} \right\|^2 \right] \\
&\leq \rho^2 L^2 D^2,
\end{aligned}$$

where (a) is obtained by substituting $\nabla f(\mathbf{x}_k^{(i)}) = \frac{1}{n} \sum_{j=1}^n \nabla f(\mathbf{x}_k^{(j)})$, (b) follows from Jensen's inequality, and (c) follows from the Assumption 1. Now combining T_1 and T_2 , we have

$$\mathbb{E} \left[\|\mathbf{z}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|^2 \right] \leq \bar{L}^2 D^2 + \rho^2 L^2 D^2 := \sigma^2.$$

□

A. Proof of Lemma 1: Inexact Iterates due to Decentralization

To prove the Lemma 1 (stated below as well), we use the following result from [3].

Lemma A6. [3][Lemma 17] Suppose a sequence of real numbers ϕ_k satisfies

$$\phi_k = \left(1 - \frac{a_1}{(k + k_0)^{a_3}} \right) \phi_{k-1} + \frac{a_2}{(k + k_0)^{2a_3}}, \quad (8)$$

for some scalars $k_0 \geq 0$, $a_1 > 1$, a_2 , and $a_3 \leq 1$. Then the sequence ϕ_k converges at the rate

$$\phi_k \leq \frac{\max\{k_0^{a_3} \phi_0, a_2/(a_1 - 1)\}}{(k + k_0 + 1)^{a_3}}. \quad (9)$$

Lemma 1. The error associated with the iterate $\mathbf{x}_k^{(i)}$ to the global consensus $\bar{\mathbf{x}}_k$ in expectation satisfies

$$q_k^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|^2 \right] \leq \frac{4D^2}{(k + 2)^\beta}.$$

Proof. Consider the distance of the iterate $\mathbf{x}_{k+1}^{(i)}$ and the global consensus $\bar{\mathbf{x}}_{k+1}$, we have

$$\begin{aligned} q_{k+1}^2 &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{x}_{k+1}^{(i)} - \bar{\mathbf{x}}_{k+1}\|^2 \right] \\ &\stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|(1 - \gamma_k)(\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k) + \gamma_k(\mathbf{s}_k^{(i)} - \bar{\mathbf{s}}_k)\|^2 \right] \\ &\stackrel{(b)}{\leq} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|(1 - \gamma_k)(\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k) + \gamma_k(\mathbf{s} - \bar{\mathbf{s}}_k)\|^2 \right] \end{aligned}$$

where (a) is from the update [cf. Algorithm 1, Line 13] and (b) follows from the definition

$$\mathbf{s} = \operatorname{argmax}_{\mathbf{s} \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|(1 - \gamma_k)(\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k) + \gamma_k(\mathbf{s} - \bar{\mathbf{s}}_k)\|^2 \right].$$

On expanding the above expression we get

$$\begin{aligned} &= (1 - \gamma_k)^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|^2 \right] + \gamma_k^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{s} - \bar{\mathbf{s}}_k\|^2 \right] \\ &\quad + 2(1 - \gamma_k)\gamma_k \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\langle \mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k, \mathbf{s} - \bar{\mathbf{s}}_k \rangle \right] \\ &\stackrel{(a)}{=} (1 - \gamma_k)^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|^2 \right] + \gamma_k^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{s}_k^{(i)} - \bar{\mathbf{s}}_k\|^2 \right] \end{aligned}$$

where (a) is from the definition $\bar{\mathbf{x}}_k = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_k^{(i)}$ due to which the cross term vanishes. From Assumption 2, we have

$$\begin{aligned}
 q_{k+1}^2 &\leq (1 - \gamma_k)q_k^2 + \gamma_k^2 D^2 \\
 &\stackrel{(a)}{=} \left(1 - \frac{2}{(k+2)^\beta}\right) q_k^2 + \frac{4D^2}{(k+2)^{2\beta}} \\
 &\stackrel{(b)}{\leq} \frac{\max\{2^\beta q_0^2, 4D^2\}}{(k+2)^\beta} \\
 &\stackrel{(c)}{\leq} \frac{4D^2}{(k+3)^\beta}
 \end{aligned}$$

where we use the step-size $\gamma_k = \frac{2}{(k+2)^\beta}$ in (a), expression (b) follows from Lemma A6 and (c) follows as $2^\beta q_0^2 = 2^\beta \frac{2D^2}{2^\beta} = 2D^2$, hence $\max\{2^\beta q_0^2, 4D^2\} = 4D^2$. \square

B. Proof of Lemma 2: Inexact Iterates due to Compression

Lemma 2. The error associated with the estimate of the compressed gradient with memory $\mathbf{a}_k^{(i)}$ to the stochastic gradient $\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)})$, $\forall i \in [n]$ satisfies

$$\begin{aligned}
 u_k^2 &= \mathbb{E} \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_k^{(i)}\|^2 \\
 &\leq \frac{C_u}{n(k+1)^\alpha},
 \end{aligned}$$

where $C_u = 2L^2 D^2 + 4n^{3/2} \delta^2$.

Proof. Consider the case $k = 0$,

$$\begin{aligned}
 u_0^2 &= \mathbb{E} \|\nabla f(\mathbf{x}_0^{(i)}, \zeta_0^{(i)}) - \mathbf{a}_0^{(i)}\|^2 \\
 &= \mathbb{E} \|\nabla f(\mathbf{x}_0^{(i)}, \zeta_0^{(i)}) - \text{Comp}(\nabla f(\mathbf{x}_0^{(i)}, \zeta_0^{(i)}))\|^2 \\
 &\stackrel{(a)}{\leq} \delta^2,
 \end{aligned} \tag{10}$$

where (a) follows from Assumption 4. Consider the conditional error (conditioned on all randomness until iterate $k+1$)

$$\begin{aligned}
& \mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_k^{(i)}\|^2 \right] \\
& \stackrel{(a)}{=} \mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \eta_k \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \right] \\
& \stackrel{(b)}{=} \mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - (1 - \eta_k) \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) + (1 - \eta_k) \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)} \right. \\
& \quad \left. - \eta_k \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \right] \\
& = \mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\underbrace{\|(1 - \eta_k) (\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}))\|}_{:=\boldsymbol{\alpha}_i}^2 + \underbrace{\|(1 - \eta_k) (\nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|}_{:=\boldsymbol{\beta}_i}^2 \right. \\
& \quad \left. + \underbrace{\|\eta_k (\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)}))\|}_{:=\boldsymbol{\gamma}_i}^2 \right] \\
& = \mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\boldsymbol{\alpha}_i\|^2 + \|\boldsymbol{\beta}_i\|^2 + \|\boldsymbol{\gamma}_i\|^2 \right. \\
& \quad + 2(1 - \eta_k)^2 \langle \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}), \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)} \rangle \\
& \quad + 2(1 - \eta_k) \eta_k \langle \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}), \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)}) \rangle \\
& \quad \left. + 2(1 - \eta_k) \eta_k \langle \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)}, \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)}) \rangle \right]
\end{aligned}$$

where (a) follows from the update of the quantized gradient vector with memory $\mathbf{a}_k^{(i)} = \mathbf{a}_{k-1}^{(i)} + \eta_k \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})$ and we introduce the term $(1 - \eta_k) \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)})$ to obtain (b). The cross terms $\langle \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i \rangle$ and $\langle \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i \rangle$ vanish due to the unbiasedness of the compression operator $\text{Comp}(\cdot)$, and we have

$$\begin{aligned}
\mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_k^{(i)}\|^2 \right] &= \|\boldsymbol{\alpha}_i\|^2 + \|\boldsymbol{\beta}_i\|^2 + \|\boldsymbol{\gamma}_i\|^2 \\
&\quad + 2(1 - \eta_k)^2 \langle \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}), \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)} \rangle.
\end{aligned}$$

Now using the iterated property of conditional expectation, i.e., $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}[\cdot]]$

$$\begin{aligned}
u_k^2 &= \mathbb{E} \left[\mathbb{E}_{\cdot|\mathcal{F}_{k+1}} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_k^{(i)}\|^2 \right] \right] \\
&= \mathbb{E} \left[\|\boldsymbol{\alpha}_i\|^2 + \|\boldsymbol{\beta}_i\|^2 + \|\boldsymbol{\gamma}_i\|^2 \right. \\
&\quad \left. + 2(1 - \eta_k)^2 \langle \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}), \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)} \rangle \right] \\
&= \mathbb{E} \left[(1 - \eta_k)^2 \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)})\|^2 + (1 - \eta_k)^2 \|\nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)}\|^2 \right. \\
&\quad + \eta_k^2 \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \\
&\quad \left. + 2(1 - \eta_k)^2 \langle \nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}), \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)} \rangle \right] \\
&\stackrel{(a)}{=} \mathbb{E} \left[(1 - \eta_k)^2 \left(1 + \frac{1}{\eta_k}\right) \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)})\|^2 + (1 - \eta_k)^2 (1 + \eta_k) \|\nabla f(\mathbf{x}_{k-1}^{(i)}, \zeta_{k-1}^{(i)}) - \mathbf{a}_{k-1}^{(i)}\|^2 \right. \\
&\quad \left. + \eta_k^2 \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \right],
\end{aligned}$$

where we use the Young's inequality $2\langle \alpha_i, \beta_i \rangle \leq \frac{1}{\eta} \|\alpha_i\|^2 + \eta \|\beta_i\|^2$ with $\eta = \eta_k$, to obtain the expression in (a).

Using the L -lipschitz continuity of gradient and from the definition u_{k-1}

$$\begin{aligned} u_k^2 &= (1 - \eta_k)(1 - \eta_k^2)u_{k-1}^2 + (1 - \eta_k)(-\eta_k + \frac{1}{\eta_k})\frac{1}{n} \sum_{i=1}^n L^2 \mathbb{E} \left[\|\mathbf{x}_k^{(i)} - \mathbf{x}_{k-1}^{(i)}\|^2 \right] \\ &\quad + \eta_k^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \right] \\ &\stackrel{(a)}{\leq} (1 - \eta_k)u_{k-1}^2 + \frac{1}{\eta_k} \gamma_k^2 L^2 D^2 + \eta_k^2 \delta^2, \end{aligned} \quad (11)$$

where we substitute $\mathbf{x}_k^{(i)} - \mathbf{x}_{k-1}^{(i)} = \gamma_k (\mathbf{s}_k^{(i)} - \mathbf{x}_{k-1}^{(i)})$ from Step 9 of Algorithm 1 and then using Assumption 2 to bound the 2nd term, and from the unbiasedness of the compression operator Assumption 4 to obtain (a). Consider the case when $k > T$, with step-size $\gamma_k = \frac{2}{(k+2)^\beta}$ and the weight $\eta_k = \frac{2\sqrt{n}}{(k+2)^\alpha}$

$$\begin{aligned} u_k^2 &\leq \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right) u_{k-1}^2 + \frac{1}{\frac{2\sqrt{n}}{(k+2)^\alpha}} \frac{4}{(k+2)^{2\beta}} L^2 D^2 + \frac{4n}{(k+2)^{2\alpha}} \delta^2 \\ &\stackrel{(b)}{=} \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right) u_{k-1}^2 + \frac{2L^2 D^2}{\sqrt{n}(k+2)^{2\alpha}} + \frac{4n\delta^2}{(k+2)^{2\alpha}} \\ &= \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right) u_{k-1}^2 + \frac{2L^2 D^2 + 4n^{3/2}\delta^2}{\sqrt{n}(k+2)^{2\alpha}} \end{aligned}$$

where in (b) we use the fact that $\alpha \leq \frac{2}{3}\beta$. Finally, from Lemma A6 with $a_1 = 2\sqrt{n}$, $a_2 = \frac{2L^2 D^2 + 4n^{3/2}\delta^2}{\sqrt{n}}$ and $a_3 = \alpha$, we obtain

$$\begin{aligned} u_k^2 &\leq \frac{\max\{2u_0^2, \frac{(2L^2 D^2 + 4n^{3/2}\delta^2)/\sqrt{n}}{2\sqrt{n}-1}\}}{(k+2)^\alpha} \\ &\stackrel{(a)}{\leq} \frac{\max\{2u_0^2, \frac{2L^2 D^2 + 4n^{3/2}\delta^2}{n}\}}{(k+2)^\alpha} \\ &\leq \frac{\max\{2nu_0^2, 2L^2 D^2 + 4n^{3/2}\delta^2\}}{n(k+2)^\alpha} \\ &= \frac{C_u}{n(k+2)^\alpha}, \end{aligned}$$

where, (a) follows from the fact that $2\sqrt{n} - 1 \geq \sqrt{n}$, for $n \geq 2$. Using Equation 10, the expression simplifies to $\max\{2nu_0^2, 2L^2 D^2 + 4n^{3/2}\delta^2\} = 2L^2 D^2 + 4n^{3/2}\delta^2$, that is, $C_u = 2L^2 D^2 + 4n^{3/2}\delta^2$.

Now consider the case $k \leq T$,

$$\begin{aligned} u_k^2 &= \mathbb{E} \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_k^{(i)}\|^2 \\ &= \mathbb{E} \|\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)} - \text{Comp}(\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)}) - \mathbf{a}_{k-1}^{(i)})\|^2 \\ &\stackrel{(a)}{\leq} \delta^2, \end{aligned}$$

where (a) follows from Assumption 4. Consider the term

$$\begin{aligned} \frac{4n^{3/2}\delta^2}{n(T+2)^\alpha} &= \frac{4\sqrt{n}\delta^2}{((2\sqrt{n})^{1/\alpha} - 2 + 2)^\alpha} \\ &= \frac{4\sqrt{n}\delta^2}{2\sqrt{n}} = 2\delta^2. \end{aligned}$$

From the above expression we see that $\delta^2 < \frac{4n^{3/2}\delta^2}{n(k+2)^\alpha}$ and $u_k^2 < \frac{C_u}{n(k+2)^\alpha} \forall k \leq T$, therefore we have $u_k^2 \leq \frac{C_u}{n(k+2)^\alpha} \forall k$. The bound follows for every node $i \in [n]$, as we are using same compression operation and variance reduction technique. So we neglected the depends of nodes on u_k . \square

C. Proof of Lemma 3: Inexact Iterates due to Stochasticity and Compression

Lemma 3. The error associated with the stochastic estimate of the gradient $\mathbf{d}_k^{(i)}$ to the deterministic gradient $\nabla f^{(i)}(\mathbf{x}_k^{(i)})$ satisfies

$$r_k^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|^2 \right] \leq \underbrace{\frac{2L^2D^2 + 4n^{(3/2)}\sigma^2}{n(k+1)^\alpha}}_{\text{stochasticity}} + \underbrace{\frac{4\sqrt{n}C_u}{n(k+1)^\alpha}}_{\text{quantization}} := \frac{C_r^2}{n(k+1)^\alpha},$$

where $C_r^2 = 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u$.

Proof. The conditional error (conditioned on all randomness up to iterate k) is given by

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_k} \left[\|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|^2 \right] \\ & \stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_k} \left[\|\nabla f(\mathbf{x}_k^{(i)}) - (1 - \eta_{k-1})\mathbf{d}_{k-1}^{(i)} + \eta_{k-1}\mathbf{c}_k^{(i)}\|^2 \right] \\ & \stackrel{(b)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_k} \left[\|(1 - \eta_{k-1}) \underbrace{(\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)}))}_{:=\boldsymbol{\alpha}_i} \right. \\ & \quad \left. + (1 - \eta_{k-1}) \underbrace{(\nabla f(\mathbf{x}_{k-1}^{(i)}) - \mathbf{d}_{k-1}^{(i)})}_{:=\boldsymbol{\beta}_i} + \eta_{k-1} \underbrace{(\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{z}_k^{(i)})}_{:=\boldsymbol{\gamma}_i} + \eta_{k-1} \underbrace{(\mathbf{z}_k^{(i)} - \mathbf{c}_k^{(i)})}_{:=\boldsymbol{\omega}_i} \|^2 \right], \end{aligned}$$

where (a) follows from Step 8 and we introduce $(1 - \eta_{k-1})\nabla f(\mathbf{x}_{k-1}^{(i)}), \eta_{k-1}\mathbf{z}_k^{(i)}$ to obtain (b). Here the vector $\mathbf{z}_k^{(i)} = \sum_{j \in [n]} w_{ij} \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)})$, the uncompressed stochastic gradient after one round of gossip. On expanding the conditional error is given as

$$\begin{aligned} & = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_k} \left[(1 - \eta_{k-1})^2 \|\boldsymbol{\alpha}_i\|^2 + (1 - \eta_{k-1})^2 \|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2 \|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2 \|\boldsymbol{\omega}_i\|^2 \right. \\ & \quad \left. + 2(1 - \eta_{k-1})^2 \langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i \rangle \right. \\ & \quad \left. + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\beta}_i, \boldsymbol{\omega}_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\alpha}_i, \boldsymbol{\omega}_i \rangle + 2\eta_{k-1}^2 \langle \boldsymbol{\omega}_i, \boldsymbol{\gamma}_i \rangle \right] \\ & \stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^n \left[(1 - \eta_{k-1})^2 \|\boldsymbol{\alpha}_i\|^2 + (1 - \eta_{k-1})^2 \|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2 \|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2 \mathbb{E}_{\cdot|\mathcal{F}_k} \|\boldsymbol{\omega}_i\|^2 \right. \\ & \quad \left. + 2(1 - \eta_{k-1})^2 \langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i \rangle, \right] \end{aligned}$$

where the three cross terms ($\langle \alpha_i, \omega_i \rangle$, $\langle \beta_i, \omega_i \rangle$, and $\langle \omega_i, \gamma_i \rangle$) vanishes due to the unbiasedness of the compression operator $\text{Comp}(\cdot)$, i.e., $\mathbb{E}_{\cdot|\mathcal{F}_k}[\omega_i] = 0 \ \forall i$. Now considering all the randomness until $k-1$ th iterate

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_{k-1}} \left[\mathbb{E}_{\cdot|\mathcal{F}_k} \left[\|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|^2 \right] \right] \\
& \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_{k-1}} \left[(1 - \eta_{k-1})^2 \|\alpha_i\|^2 + (1 - \eta_{k-1})^2 \|\beta_i\|^2 + \eta_{k-1}^2 \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E}_{\cdot|\mathcal{F}_k} \|\omega_i\|^2 \right. \\
& \quad \left. + 2(1 - \eta_{k-1})^2 \langle \alpha_i, \beta_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \alpha_i, \gamma_i \rangle + 2(1 - \eta_{k-1})\eta_{k-1} \langle \beta_i, \gamma_i \rangle \right] \\
& \stackrel{(a)}{\leq} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\cdot|\mathcal{F}_{k-1}} \left[(1 - \eta_{k-1})^2 \|\alpha_i\|^2 + (1 - \eta_{k-1})^2 \|\beta_i\|^2 + \eta_{k-1}^2 \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E}_{\cdot|\mathcal{F}_k} \|\omega_i\|^2 \right. \\
& \quad \left. + 2(1 - \eta_{k-1})^2 \langle \alpha_i, \beta_i \rangle \right].
\end{aligned}$$

In a deterministic setting, since each node has access to the entire data, the sequence of iterates and the full gradients are deterministic and the same (i.e., $\mathbf{x}_k^{(i)} = \mathbf{x}_k^{(j)}$ and $\nabla f(\mathbf{x}_k^{(i)}) = \nabla f(\mathbf{x}_k^{(j)})$ for $i, j \in [n]$) when each node has the same initial point. Due to unbiasedness of the local stochastic gradient $\nabla f(\mathbf{x}_k^{(i)}, \zeta_k^{(i)})$ and that $\mathbf{c}_k^{(i)}$ is obtained by a convex combination of full (deterministic) gradients under $\mathbb{E}_{\cdot|\mathcal{F}_k}$, we have $\mathbb{E}_{\cdot|\mathcal{F}_{k-1}}[\mathbf{c}_k^{(i)}] = \nabla f(\mathbf{x}_k^{(i)})$. Consequently, the two cross terms ($\langle \alpha_i, \gamma_i \rangle$ and $\langle \beta_i, \gamma_i \rangle$) vanish under $\mathbb{E}_{\cdot|\mathcal{F}_{k-1}}$ in (a).

Due to the fact that $\mathbb{E}[\mathbb{E}_{\cdot|\mathcal{F}_{k-1}}[\mathbb{E}_{\cdot|\mathcal{F}_k}[\cdot]]] = \mathbb{E}[\cdot]$, we have

$$\begin{aligned}
r_k^2 &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E}_{\cdot|\mathcal{F}_{k-1}} \left[\mathbb{E}_{\cdot|\mathcal{F}_k} \left[\|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|^2 \right] \right] \right] \\
&\leq \frac{1}{n} \sum_{i=1}^n (1 - \eta_{k-1})^2 \mathbb{E} \|\alpha_i\|^2 + (1 - \eta_{k-1})^2 \mathbb{E} \|\beta_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\omega_i\|^2 \\
&\quad + 2(1 - \eta_{k-1})^2 \mathbb{E} \langle \alpha_i, \beta_i \rangle \\
&\stackrel{(a)}{\leq} \frac{1}{n} \sum_{i=1}^n (1 - \eta_{k-1})^2 \mathbb{E} \|\alpha_i\|^2 + (1 - \eta_{k-1})^2 \mathbb{E} \|\beta_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\omega_i\|^2 \\
&\quad + (1 - \eta_{k-1})^2 \mathbb{E} \left[\frac{1}{\eta_{k-1}} \|\alpha_i\|^2 + \eta_{k-1} \|\beta_i\|^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n (1 - \eta_{k-1})^2 \left(1 + \frac{1}{\eta_{k-1}} \right) \mathbb{E} \|\alpha_i\|^2 + (1 - \eta_{k-1})^2 (1 + \eta_{k-1}) \mathbb{E} \|\beta_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\omega_i\|^2,
\end{aligned}$$

where (a) obtained using Young's inequality with $\eta = \eta_{k-1}$. For the case $k \geq T$, using the inequalities $(1 - \eta_{k-1})(1 + \eta_{k-1}) = (1 - \eta_{k-1}^2) \leq 1$ and $(1 - \eta_{k-1})^2(1 + \frac{1}{\eta_{k-1}}) \leq \frac{1}{\eta_{k-1}}$, we have

$$r_k^2 \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_{k-1}} \mathbb{E} \|\alpha_i\|^2 + (1 - \eta_{k-1}) \mathbb{E} \|\beta_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\gamma_i\|^2 + \eta_{k-1}^2 \mathbb{E} \|\omega_i\|^2. \quad (12)$$

Consider the term,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\omega}_i\|^2 &= \mathbb{E}\|\mathbf{z}_k^{(i)} - \mathbf{c}_k^{(i)}\|^2 \\
&= \mathbb{E}\left\|\sum_{j \in [n]} w_{ij} \nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \sum_{j \in [n]} w_{ij} \mathbf{a}_k^{(j)}\right\|^2 \\
&\stackrel{(a)}{\leq} \sum_{j \in [n]} w_{ij} \mathbb{E}\left\|\nabla f(\mathbf{x}_k^{(j)}, \zeta_k^{(j)}) - \mathbf{a}_k^{(j)}\right\|^2 \\
&\stackrel{(b)}{\leq} u_k^2,
\end{aligned} \tag{13}$$

where (a), follows from Jensen's inequality for convex functions and (b) follows from the definition of u_k^2 . Substituting $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$, Equation 13, and Lemma A5 in Equation 12, we have the following bound

$$\begin{aligned}
r_k^2 &\leq \frac{1}{\eta_{k-1}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\|\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\mathbf{x}_{k-1}^{(i)})\|^2 + \frac{(1 - \eta_{k-1})}{n} \sum_{i=1}^n \mathbb{E}\|\nabla f(\mathbf{x}_{k-1}^{(i)}) - \mathbf{d}_{k-1}^{(i)}\|^2 + \eta_{k-1}^2 \sigma^2 + \eta_{k-1}^2 u_k^2 \\
&\stackrel{(a)}{\leq} (1 - \eta_{k-1}) r_{k-1}^2 + \frac{1}{\eta_{k-1}} \frac{L^2}{n} \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{x}_k^{(i)} - \mathbf{x}_{k-1}^{(i)}\|^2\right] + \eta_{k-1}^2 \sigma^2 + \eta_{k-1}^2 u_k^2 \\
&\stackrel{(b)}{\leq} (1 - \eta_{k-1}) r_{k-1}^2 + \frac{1}{\eta_{k-1}} \gamma_{k-1}^2 \frac{L^2 D^2}{n} + \eta_{k-1}^2 \sigma^2 + \eta_{k-1}^2 u_k^2,
\end{aligned}$$

where, we substitute $\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\|\nabla f(\mathbf{x}_{k-1}^{(i)}) - \mathbf{d}_{k-1}^{(i)}\|^2\right]$ with r_{k-1}^2 and use Assumption 1 in (a) and (b) follows from the Step 9 and Assumption 2. Using Lemma 2,

$$\begin{aligned}
r_k^2 &\leq (1 - \eta_{k-1}) r_{k-1}^2 + \frac{1}{\eta_{k-1}} \gamma_{k-1}^2 L^2 D^2 + \eta_{k-1}^2 \sigma^2 + \eta_{k-1}^2 \frac{C_u}{n(k+2)^\alpha} \\
&\stackrel{(a)}{=} \left(1 - \frac{2\sqrt{n}}{(k+1)^\alpha}\right) r_{k-1}^2 + \frac{1}{\frac{2\sqrt{n}}{(k+1)^\alpha}} \frac{4}{(k+1)^{2\beta}} L^2 D^2 + \frac{4n}{(k+1)^{2\alpha}} \sigma^2 + \frac{4n}{(k+1)^{2\alpha}} \frac{C_u}{n(k+2)^\alpha} \\
&\stackrel{(b)}{\leq} \left(1 - \frac{2\sqrt{n}}{(k+1)^\alpha}\right) r_{k-1}^2 + \frac{2L^2 D^2 + 4n^{(3/2)} \sigma^2 + 4\sqrt{n} C_u}{\sqrt{n}(k+1)^{2\alpha}},
\end{aligned}$$

where we use the step-size $\gamma_{k-1} = \frac{2}{(k+1)^\beta}$ and the weight $\eta_{k-1} = \frac{2\sqrt{n}}{(k+1)^\alpha}$ in (a) and in (b), we use the fact that $\alpha \leq \frac{2}{3}\beta$. Using Lemma A6

$$\begin{aligned}
r_k^2 &\leq \frac{\max\{r_0^2, \frac{(2L^2 D^2 + 4n^{(3/2)} \sigma^2 + 4\sqrt{n} C_u)/\sqrt{n}}{2\sqrt{n}-1}\}}{(k+2)^\alpha} \\
&\stackrel{(a)}{\leq} \frac{\max\{r_0^2, \frac{2L^2 D^2 + 4n^{(3/2)} \sigma^2 + 4\sqrt{n} C_u}{n}\}}{(k+2)^\alpha} \\
&= \frac{\max\{nr_0^2, 2L^2 D^2 + 4n^{(3/2)} \sigma^2 + 4\sqrt{n} C_u\}}{n(k+2)^\alpha} \\
&= \frac{C_r^2}{n(k+2)^\alpha},
\end{aligned}$$

where, (a) follows from the fact that $2\sqrt{n} - 1 \geq \sqrt{n}$, with $n \geq 2$. Consider the term

$$\begin{aligned} r_0^2 &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f(\mathbf{x}_0^{(i)}) - \mathbf{d}_0^{(i)}\|^2 \right] \\ &\stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f(\mathbf{x}_0^{(i)}) - \mathbf{c}_0^{(i)}\|^2 \right] \\ &\stackrel{(b)}{\leq} \sigma^2, \end{aligned} \tag{14}$$

where, (a) follows from Step 8 with $\eta_0 = 1$ and (b) follows from Lemma A5. Using Equation 14, $\max\{nr_0^2, 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u\} = 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u := C_r^2$.

Now consider the case $k \leq T$,

$$r_k^2 = \mathbb{E} \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|^2 = \mathbb{E} \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{c}_k^{(i)}\|^2 \leq \sigma^2,$$

where (a) follows from Lemma A5. Consider the term

$$\begin{aligned} \frac{4n^{3/2}\sigma^2}{n(T+2)^\alpha} &= \frac{4\sqrt{n}\sigma^2}{((2\sqrt{n})^{1/\alpha} - 2 + 2)^\alpha} \\ &= \frac{4\sqrt{n}\sigma^2}{2\sqrt{n}} = 2\sigma^2. \end{aligned}$$

From the above expression we see that $\sigma^2 < \frac{4n^{3/2}\sigma^2}{n(k+2)^\alpha}$ and $r_k^2 < \frac{C_r^2}{n(k+2)^\alpha} \forall k \leq T$, therefore we bound $r_k^2 \leq \frac{C_r^2}{n(k+2)^\alpha} \forall k$.

□

D. Proof of Lemma 4: Generalized Suboptimality Gap

Lemma 4. For an L -smooth convex function $f : \mathcal{R}^d \rightarrow \mathcal{R}$, convex and compact constraint set \mathcal{C} with diameter D , the suboptimality gap $h_k = \mathbb{E}[f(\bar{\mathbf{x}}_k) - f_{\text{opt}}]$ satisfies

$$h_{k+1} \leq (1 - \gamma_k)h_k + \gamma_k^2 \frac{LD^2}{2} + \underbrace{\gamma_k 2LDq_k}_{\text{decentralization}} + \underbrace{\gamma_k 2Dr_k}_{\text{stochasticity}}.$$

Proof. We derive an upper bound on the suboptimality gap due to inexact updates using L -smoothness of f [cf. Assumption 1] as

$$\begin{aligned} f(\bar{\mathbf{x}}_{k+1}) &\leq f(\bar{\mathbf{x}}_k) + \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k \rangle + \frac{L}{2} \|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k\|_2^2 \\ &\stackrel{(a)}{=} f(\bar{\mathbf{x}}_k) + \gamma_k \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \gamma_k^2 \frac{L}{2} \|\bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k\|_2^2 \\ &\stackrel{(b)}{=} f(\bar{\mathbf{x}}_k) + \gamma_k \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2, \end{aligned} \tag{15}$$

where (a) follows from the FW update Step 9 and (b) follows from Assumption 2.

Let us introduce the average variance-reduced gradient $\bar{\mathbf{d}}_k = \frac{1}{n} \sum_{i=1}^n \mathbf{d}_k^{(i)}$ and the average deterministic gradient $\bar{\nabla} f_k = \frac{1}{n} \sum_{i=1}^n \nabla f(\mathbf{x}_k^{(i)})$ in $\langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle$ as

$$\begin{aligned}
\langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle &= \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \langle \nabla f(\bar{\mathbf{x}}_k) - \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&= \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \langle \nabla f(\bar{\mathbf{x}}_k) - \bar{\nabla} f_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&\quad + \langle \bar{\nabla} f_k - \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&= \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)}), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&\quad + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle &\leq \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)})\|_2 \|\bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k\|_2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2 \|\bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k\|_2 \\
&\stackrel{(a)}{\leq} \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)})\|_2 D \\
&\quad + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2 D \\
&\stackrel{(b)}{\leq} \langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n LD \|\bar{\mathbf{x}}_k - \mathbf{x}_k^{(i)}\|_2 + \frac{1}{n} \sum_{i=1}^n D \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2, \tag{16}
\end{aligned}$$

where (a) follows from Assumption 2 and (b) follows from Assumption 1.

Next we bound $\langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle$ by introducing the terms $\nabla f(\bar{\mathbf{x}}_k)$ and $\overline{\nabla f}_k$ as

$$\begin{aligned}
\langle \bar{\mathbf{d}}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle &= \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \langle \overline{\nabla f}_k - \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&\quad + \langle \bar{\mathbf{d}}_k - \overline{\nabla f}_k, \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&= \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&\quad + \frac{1}{n} \sum_{i=1}^n \langle \mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)}), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \\
&\stackrel{(a)}{\leq} \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\bar{\mathbf{x}}_k)\|_2 \|\bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k\|_2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2 \|\bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k\|_2 \\
&\stackrel{(b)}{\leq} \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \nabla f(\bar{\mathbf{x}}_k)\|_2 D \\
&\quad + \frac{1}{n} \sum_{i=1}^n \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2 D \\
&\stackrel{(c)}{\leq} \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \frac{1}{n} \sum_{i=1}^n LD \|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|_2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n D \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2, \tag{17}
\end{aligned}$$

where we employ the Cauchy-Schwarz inequality in (a), Assumption 1 in (b), and Assumption 2 in (c).

Using Equation 16 and Equation 17 in Equation 15, we get

$$\begin{aligned}
f(\bar{\mathbf{x}}_{k+1}) &\leq f(\bar{\mathbf{x}}_k) + \gamma_k \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k \frac{1}{n} \sum_{i=1}^n 2LD \|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|_2 \\
&\quad + \gamma_k \frac{1}{n} \sum_{i=1}^n 2D \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2 \\
&\stackrel{(a)}{\leq} f(\bar{\mathbf{x}}_k) + \gamma_k \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LD \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|_2^2} \\
&\quad + \gamma_k 2D \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2^2},
\end{aligned}$$

where the above expression follows from the concavity of $\sqrt{(\cdot)}$. Finally, we have

$$\begin{aligned}
\mathbb{E} \left[f(\bar{\mathbf{x}}_{k+1}) - f_{\text{opt}} \right] &\leq \mathbb{E} \left[f(\bar{\mathbf{x}}_k) - f_{\text{opt}} + \gamma_k \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 \right. \\
&\quad + \gamma_k 2LD \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|_2^2} \\
&\quad \left. + \gamma_k 2D \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2^2} \right] \\
&\stackrel{(a)}{\leq} \mathbb{E} \left[f(\bar{\mathbf{x}}_k) - f_{\text{opt}} \right] + \gamma_k \mathbb{E} \left[\langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle \right] + \gamma_k^2 \frac{L}{2} D^2 \\
&\quad + \gamma_k 2LD \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k\|_2^2 \right]} \\
&\quad + \gamma_k 2D \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{d}_k^{(i)} - \nabla f(\mathbf{x}_k^{(i)})\|_2^2 \right]},
\end{aligned}$$

where we use Jensen's inequality for concave functions in (a). Thus we have

$$\begin{aligned}
h_{k+1} &\leq h_k + \gamma_k \mathbb{E}[\langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{s}}_k - \bar{\mathbf{x}}_k \rangle] + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LDq_k + \gamma_k 2Dr_k \\
&\stackrel{(a)}{\leq} h_k - \gamma_k \mathbb{E}[g_k] + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LDq_k + \gamma_k 2Dr_k \\
&\stackrel{(b)}{\leq} (1 - \gamma_k)h_k + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LDq_k + \gamma_k 2Dr_k,
\end{aligned} \tag{18}$$

where (a) follows from the definition of the Frank-Wolfe duality gap

$$g_k = \max_{\mathbf{s} \in \mathcal{C}} \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{x}}_k - \mathbf{s} \rangle$$

and (b) is because of the fact that the duality gap upper bounds the primal suboptimality gap for convex functions. \square

E. Proof of Theorem 1: Convergence Rate for Convex Objectives

Theorem 1. For an L -smooth convex function $f : \mathcal{R}^d \rightarrow \mathcal{R}$, convex and compact constraint set \mathcal{C} with diameter D , the average suboptimality gap at the k th iterate with $\gamma_k = \frac{2}{k+2}$ and $\alpha = 2/3$ satisfies

$$\mathbb{E}[f(\bar{\mathbf{x}}_k) - f_{\text{opt}}] \leq \frac{2LD^2}{(k+1)} (-1 + \log(k+1)) + \frac{16LD^2}{\sqrt{k+1}} + \frac{6C_r D}{\sqrt{n}(k+1)^{1/3}}, \tag{19}$$

where $C_r^2 = 2L^2 D^2 + 4n^{(3/2)} \sigma^2 + 4\sqrt{n} C_u$.

Proof. Consider the bound on the suboptimality gap in Lemma 4

$$h_{k+1} \leq (1 - \gamma_k)h_k + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LDq_k + \gamma_k 2Dr_k.$$

Unrolling the bound up to K iterations

$$\begin{aligned}
h_K &\leq \prod_{j=0}^{K-1} (1 - \gamma_j) h_0 + \frac{LD^2}{2} \sum_{i=0}^{K-1} \gamma_i^2 \prod_{j=i+1}^{K-1} (1 - \gamma_j) + 2LD \sum_{i=0}^{K-1} \gamma_i q_i \prod_{j=i+1}^{K-1} (1 - \gamma_j) \\
&\quad + 2D \sum_{i=0}^{K-1} \gamma_i r_i \prod_{j=i+1}^{K-1} (1 - \gamma_j) \\
&\stackrel{(a)}{=} 0 + \frac{LD^2}{2} \sum_{i=0}^{K-1} \gamma_i^2 \prod_{j=i+1}^{K-1} (1 - \gamma_j) + 2LD \sum_{i=0}^{K-1} \gamma_i q_i \prod_{j=i+1}^{K-1} (1 - \gamma_j) + 2D \sum_{i=0}^{K-1} \gamma_i r_i \prod_{j=i+1}^{K-1} (1 - \gamma_j) \\
&\stackrel{(b)}{\leq} \frac{LD^2}{2} \sum_{i=0}^{K-1} \frac{4}{(i+2)^2} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right) + 2LD \sum_{i=0}^{K-1} \frac{2}{i+2} \frac{2D}{\sqrt{i+1}} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right) \\
&\quad + 2D \sum_{i=0}^{K-1} \frac{2}{i+2} \frac{C_r}{\sqrt{n}(i+2)^{1/3}} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right)
\end{aligned}$$

where we invoke the fact that $\gamma_0 = 1$ and $1 - \gamma_j \leq 1$, $\forall j$ to obtain (a). We substitute for q_k from Lemma 2 and r_k from Lemma 3, with $\alpha = 2/3$ and use the fact that step-size of $\gamma_j = \frac{2}{j+2}$ in (b). Consider the term

$$\begin{aligned}
\sum_{i=0}^{K-1} \frac{4}{(i+2)^2} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right) &= \frac{4}{(K+1)^2} + \frac{4}{K^2} \frac{K-1}{K+1} + \frac{4}{(K-1)^2} \frac{K-2}{K} \frac{K-1}{K+1} + \dots \\
&= \frac{1}{K+1} \left(\frac{4}{K+1} + \frac{4(K-1)}{K^2} + \frac{4(K-2)}{(K-1)K} + \dots \right) \\
&\leq \frac{1}{K+1} \left(\frac{4}{K+1} + \frac{4}{K} + \frac{4}{(K-1)} + \dots \right) \\
&= \frac{1}{K+1} \sum_{i=0}^{K-1} \frac{4}{i+2}.
\end{aligned}$$

Now consider the term

$$\begin{aligned}
\sum_{i=0}^{K-1} \frac{2}{i+2} \frac{2D}{\sqrt{i+1}} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right) &= \frac{4D}{(K+1)\sqrt{K}} + \frac{4D}{K\sqrt{K-1}} \frac{K-1}{K+1} + \frac{4D}{(K-1)\sqrt{K-2}} \frac{K-2}{K} \frac{K-1}{K+1} + \dots \\
&= \frac{1}{K+1} \left(\frac{4D}{\sqrt{K}} + \frac{4D(K-1)}{K\sqrt{K-1}} + \frac{4D(K-2)}{K\sqrt{K-2}} + \dots \right) \\
&\leq \frac{1}{K+1} \left(\frac{4D}{\sqrt{K}} + \frac{4D}{\sqrt{K-1}} + \frac{4D}{\sqrt{K-2}} + \dots \right) \\
&= \frac{1}{K+1} \sum_{i=0}^{K-1} \frac{4D}{\sqrt{i+1}}.
\end{aligned}$$

Similarly, we have

$$\sum_{i=0}^{K-1} \frac{2}{i+2} \frac{C_r}{\sqrt{n}(i+2)^{1/3}} \prod_{j=i+1}^{K-1} \left(1 - \frac{2}{j+2}\right) = \frac{1}{K+1} \sum_{i=0}^{K-1} \frac{2C_r}{\sqrt{n}(i+2)^{1/3}}.$$

Using the above inequalities we have

$$h_K \leq \frac{LD^2}{2(K+1)} \sum_{i=0}^{K-1} \frac{4}{(i+2)} + \frac{8LD^2}{K+1} \sum_{i=0}^{K-1} \frac{1}{\sqrt{i+1}} + \frac{4C_r D}{\sqrt{n}(K+1)} \sum_{i=0}^{K-1} \frac{1}{(i+2)^{1/3}}.$$

Now let us use the following inequalities

$$\begin{aligned}
\sum_{j=0}^{K-1} \frac{1}{j+2} &\leq -1 + \int_{j=1}^{K+1} \frac{1}{j} dj \\
&= -1 + \log(K+1), \\
\sum_{j=0}^{K-1} \frac{1}{\sqrt{j+1}} &\leq \int_{j=0}^K \frac{1}{\sqrt{j}} dj \\
&= 2\sqrt{K},
\end{aligned}$$

and

$$\begin{aligned}
\sum_{j=0}^{K-1} \frac{1}{(j+2)^{1/3}} &\leq \int_{j=0}^{K+1} \frac{1}{j^{1/3}} dj \\
&= \frac{(K+1)^{1-1/3}}{1-1/3} \\
&= \frac{3}{2}(K+1)^{2/3}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
h_K &\leq \frac{2LD^2}{(K+1)} (-1 + \log(K+1)) + \frac{8LD^2}{K+1} 2\sqrt{K} + \frac{4C_r D}{\sqrt{n}(K+1)} \frac{3}{2}(K+1)^{2/3} \\
&\leq \frac{2LD^2}{(K+1)} (-1 + \log(K+1)) + \frac{16LD^2}{\sqrt{K+1}} + \frac{6C_r D}{\sqrt{n}(K+1)^{1/3}}.
\end{aligned}$$

□

F. Proof of Lemma 5: Convergence of Average Duality Gap

Lemma 5. For an L -smooth but non-convex $f : \mathcal{R}^d \rightarrow \mathcal{R}$, convex and compact constraint set \mathcal{C} , the distance between the duality gaps g_k and \hat{g}_k , in expectation satisfies

$$\mathbb{E} [|g_k - \hat{g}_k|] \leq LDq_k + Dr_k. \quad (20)$$

Proof. To obtain a bound on the distance between the true duality gap and the average duality gap at each node, we

use the definition of the duality gap at node i

$$\begin{aligned}
g_k &= \max_{s \in \mathcal{C}} \langle \nabla f(\bar{\mathbf{x}}_k), \bar{\mathbf{x}}_k - s \rangle = \max_{s \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_k), \mathbf{x}_k^{(i)} - s \rangle \\
&\leq \frac{1}{n} \sum_{i=1}^n \max_{s \in \mathcal{C}} \langle \nabla f(\bar{\mathbf{x}}_k), \mathbf{x}_k^{(i)} - s \rangle \\
&\stackrel{(a)}{\leq} \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_k), \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle \\
&\stackrel{(b)}{=} \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_k^{(i)}), \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)}), \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle \\
&\stackrel{(c)}{=} \frac{1}{n} \sum_{i=1}^n \langle \mathbf{d}_k^{(i)}, \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}, \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle \\
&\quad + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)}), \mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)} \rangle,
\end{aligned}$$

where we use the solution to the linear program $\mathbf{s}_k^{(i)}$ in (a). We introduce the full gradient $\nabla f(\mathbf{x}_k^{(i)})$ in (b) and the variance-reduced stochastic gradient direction $\mathbf{d}_k^{(i)}$ in (c). From the definition of \hat{g}_k and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
g_k &\leq \hat{g}_k + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2 \|\mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)}\|_2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)})\|_2 \|\mathbf{x}_k^{(i)} - \mathbf{s}_k^{(i)}\|_2 \\
&\stackrel{(a)}{\leq} \hat{g}_k + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2 D + \frac{1}{n} \sum_{i=1}^n \|\nabla f(\bar{\mathbf{x}}_k) - \nabla f(\mathbf{x}_k^{(i)})\|_2 D \\
&\stackrel{(b)}{\leq} \hat{g}_k + \frac{D}{n} \sum_{i=1}^n \|\nabla f(\mathbf{x}_k^{(i)}) - \mathbf{d}_k^{(i)}\|_2 + \frac{LD}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_k - \mathbf{x}_k^{(i)}\|_2
\end{aligned}$$

$$\mathbb{E}[g_k] \leq \mathbb{E}[\hat{g}_k] + Dr_k + LDq_k$$

where we use Assumption 2 in (a) and Assumption 1 in (b). \square

G. Proof of Theorem 2: Convergence Rate for Non-convex Objectives

Theorem 2. For an L -smooth but non-convex function $f : \mathcal{R}^d \rightarrow \mathcal{R}$, over a convex and compact constraint set \mathcal{C} with diameter D , the average FW duality gap $\mathbb{E}[\bar{g}_K]$, with $\beta = 2/3$ and $\alpha = 4/9$ satisfies

$$\begin{aligned}
\mathbb{E}[\bar{g}_K] &= \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[g_k] \\
&\leq \frac{\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + 6LD^2}{(K+1)^{-1/3}} + \frac{LD^2}{2(K+1)^{-2/3}} + \frac{3C_r D}{\sqrt{n}(K+2)^{-2/9}}.
\end{aligned}$$

Proof. From the bound on suboptimality gap in Equation 18, we have

$$\begin{aligned} h_{k+1} &\leq h_k - \gamma_k \mathbb{E}[g_k] + \gamma_k^2 \frac{LD^2}{2} + \gamma_k 2LDq_k + \gamma_k 2Dr_k \\ \gamma_k \mathbb{E}[g_k] &\leq \mathbb{E}[f(\bar{\mathbf{x}}_k) - f(\bar{\mathbf{x}}_{k+1})] + \gamma_k^2 \frac{LD^2}{2} + \gamma_k 2LDq_k + \gamma_k 2Dr_k. \end{aligned}$$

Summing from $k = 0, \dots, K$, $f(\bar{\mathbf{x}}_k) - f(\bar{\mathbf{x}}_{k+1})$ telescopes to

$$\sum_{k=0}^K \gamma_k \mathbb{E}[g_k] \leq \mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + \sum_{k=0}^K \gamma_k^2 \frac{LD^2}{2} + \sum_{k=0}^K \gamma_k 2LDq_k + \sum_{k=0}^K \gamma_k 2Dr_k.$$

Consider

$$\mathbb{E}[\bar{g}_K] = \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[g_k] \quad (21)$$

$$\stackrel{(a)}{\leq} \frac{1}{(K+1) \cdot (K+2)^{-2/3}} \left[\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + \frac{LD^2}{2} \sum_{k=0}^K (K+2)^{-4/3} + 2LD \sum_{k=0}^K (K+2)^{-2/3} q_k \right. \quad (22)$$

$$\left. + 2D \sum_{k=0}^K (K+2)^{-2/3} r_k \right]$$

$$\stackrel{(b)}{\leq} \frac{1}{(K+1)^{1/3}} \left[\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + \frac{LD^2}{2} \sum_{k=0}^K (K+2)^{-4/3} + 2LD \sum_{k=0}^K (K+2)^{-2/3} \frac{2D}{(k+2)^{1/3}} \right. \quad (23)$$

$$\left. + 2D \sum_{k=0}^K (K+2)^{-2/3} \frac{C_r}{\sqrt{n}(k+2)^{2/9}} \right], \quad (24)$$

where (a) follows by substituting $\gamma_k = (K+2)^{-2/3}$ and (b) follows by substituting for r_k and q_k with $\alpha = 4/9$.

Now consider

$$\begin{aligned} \sum_{j=0}^K \frac{1}{(j+2)^{2/9}} &\leq \int_{j=1}^{K+2} \frac{1}{j^{2/9}} dj \\ &= \frac{(K+2)^{1-2/9}}{1-2/9} \\ &= \frac{9}{7} (K+2)^{7/9}. \end{aligned}$$

Using the above inequality in Equation 24

$$\begin{aligned} \mathbb{E}[\bar{g}_K] &\leq \frac{1}{(K+1)^{1/3}} \left[\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + \frac{LD^2}{2} (K+1) \cdot (K+2)^{-4/3} + 4LD^2 (K+2)^{-2/3} \frac{3}{2} (K+2)^{2/3} \right. \\ &\quad \left. + \frac{2C_r D}{\sqrt{n}} (K+2)^{-2/3} \frac{9}{7} (K+2)^{7/9} \right] \\ &\leq \frac{1}{(K+1)^{1/3}} \left[\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + \frac{LD^2}{2} (K+1)^{-1/3} + 6LD^2 + \frac{2C_r D}{\sqrt{n}} \frac{9}{7} (K+2)^{1/9} \right] \\ &= \frac{\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})]}{(K+1)^{1/3}} + \frac{LD^2}{2(K+1)^{2/3}} + \frac{6LD^2}{(K+1)^{1/3}} + \frac{3C_r D}{\sqrt{n}(K+2)^{2/9}} \\ &= \frac{\mathbb{E}[f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_{K+1})] + 6LD^2}{(K+1)^{1/3}} + \frac{LD^2}{2(K+1)^{2/3}} + \frac{3C_r D}{\sqrt{n}(K+2)^{2/9}}. \end{aligned}$$

□

REFERENCES

- [1] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, “Weighted gossip: Distributed averaging using non-doubly stochastic matrices,” in *Proceedings of the IEEE International Symposium on Information Theory*, 2010, pp. 1753–1757.
- [2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Gossip algorithms: design, analysis and applications,” in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3, 2005, pp. 1653–1664 vol. 3.
- [3] A. Mokhtari, H. Hassani, and A. Karbasi, “Stochastic conditional gradient methods: From convex minimization to submodular maximization,” 2018.