Appendix

In this section, we first provide further numerical experiments as well as presenting the hyper-parameters of all methods used in the experiments. We then continue with presenting proofs of the Theorems 5.1-5.3.

7.1. Neural Network Training on CIFAR-10 dataset

We train a neural network of 20 hidden-units with sigmoid activation functions for binary classification of the CIFAR-10 dataset. We use the topology of $\mathcal{G}_1$ as in Figure 1 with the corresponding weight matrix constructed according to Example 1. The value of step size α is fine-tuned among 15 values of α uniformly selected in $[0.1, 3]$ and up to iteration 300 for both algorithms.

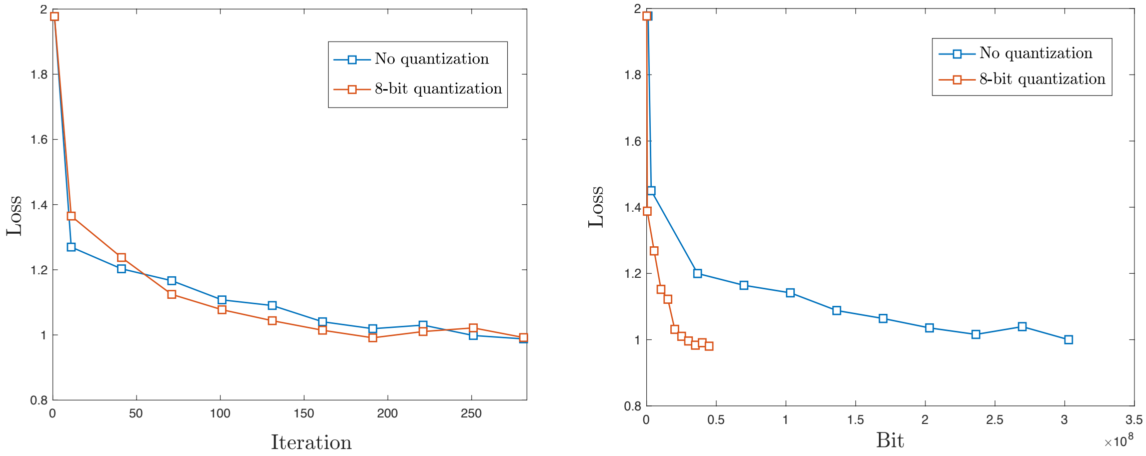

Figure 5. Comparison of the proposed method and exact-communication push-sum method in training a neural network with CIFAR-10 data-set, based on the iteration number (Left) and total number of bits communicated between two neighbor nodes (Right).

Figure 5 illustrates the performance of the proposed algorithm based on iteration number (Left) and number of bits communicated between two neighbor nodes (Right), and compares the vanilla push-sum method and the proposed communication-efficient method with 8 bits for quantization. We use SGD with mini-batch size of 100 samples for each node in both methods. We highlight the close similarity in training loss of the two methods for a fixed number of iteration. More importantly the proposed method uses remarkably smaller number of bits implying faster communication while reaching the same level of training loss.

7.2. Details of the Numerical Experiments

The step-sizes of the algorithms used in Section 6 are fine tuned over the interval $[0.01,\ 3]$, so that the best error achieved by each method is compared. In Table 1 we present the fine-tuned step sizes as well as model size (i.e. dimension of parameteres) and the mini-batch size of the algorithms that are used throughout the numerical experiments.

Table 1. Details on the hyper-parameters of the vanilla push-sum algorithm with no quantization and the proposed quantized push-sum algorithm.

Objective	Iteration	Model size	Mini-batch size	Graph	Step size
Square loss	50	256	1	$\mathcal{G}_1$	1.7
Square loss	50	256	1	$\mathcal{G}_2$	1.1
NN, MNIST	200	7960	10	$\mathcal{G}_1$	2.2
NN, CIFAR-10	300	20542	100	$\mathcal{G}_1$	1.1
Square loss (4-bits)	50	256	1	$\mathcal{G}_1$	1.1
Square loss (16-bits)	50	256	1	$\mathcal{G}_2$	1.1
NN, MNIST (8-bits)	200	7960	10	$\mathcal{G}_1$	1.9
NN, CIFAR-10 (8-bits)	300	20542	100	$\mathcal{G}_1$	0.3

Proofs

Notation

Throughout this section we set the following notation. For variables, stochastic gradients and gradients, we concatenate the row vectors corresponding to each node to form the following matrices:

$$\begin{aligned}
&Z(t) := [\mathbf{z}_1(t)\,; \quad \mathbf{z}_2(t) \quad \cdots \quad \mathbf{z}_n(t)] \in \mathbb{R}^{n\times d},\\
&\partial F\left(Z(t), \zeta_t\right) := \left[\nabla F_1\left(\mathbf{z}_1(t), \zeta_{1,t}\right)\,; \quad \nabla F_2\left(\mathbf{z}_2(t), \zeta_{2,t}\right) \quad \cdots \quad \nabla F_n\left(\mathbf{z}_n(t), \zeta_{n,t}\right)\right] \in \mathbb{R}^{n\times d},\\
&\partial f\left(Z(t)\right) := \left[\nabla f_1\left(\mathbf{z}_1(t)\right)\,; \quad \nabla f_2\left(\mathbf{z}_2(t)\right) \quad \cdots \quad \nabla f_n\left(\mathbf{z}_n(t)\right)\right] \in \mathbb{R}^{n\times d}.
\end{aligned}$$

8. Proof of Theorem 5.1 : Quantized Gossip over Directed Graphs

First we write the iterations of Algorithm 1 in matrix notation to derive the following:

$$\begin{cases}
Q(t) = Q\left(X(t) - \widehat{X}(t)\right)\\
\widehat{X}(t+1) = \widehat{X}(t) + Q(t)\\
X(t+1) = X(t) + (A-I)\widehat{X}(t+1)\\
\mathbf{y}(t+1) = A\mathbf{y}(t)\\
\mathbf{z}_i(t+1) = \frac{\mathbf{x}_i(t+1)}{y_i(t+1)}
\end{cases} \tag{11}$$

Based on this, we can rewrite the update rule for $X(t+1)$ as follows:

$$X(t+1) = AX(t) + (A-I)(\widehat{X}(t+1) - X(t)).$$

By repeating this for $X(t), ..., X(1)$, the update rule for $X(t+1)$ takes the following shape:

$$X(t+1) = A^t X(1) + \sum_{s=0}^{t-1} A(A-I)(X(t-s+1) - X(t-s)). \tag{12}$$

Multiplying both sides by $\mathbf{1}^T$ and recalling that $\mathbf{1}^T A = \mathbf{1}^T$, yields that

$$\mathbf{1}^T X(t+1) = \mathbf{1}^T X(1). \tag{13}$$

With this and (12) we have for all $t \geq 1$:

$$\begin{aligned}
\left\|X(t+1) - \boldsymbol{\phi}\mathbf{1}^T X(1)\right\| &= \left\|A^t X(1) + \sum_{s=0}^{t-1} A^s (A-I)(\widehat{X}(t-s+1) - X(t-s)) - \boldsymbol{\phi}\mathbf{1}^T X(1)\right\|\\
&\leqslant C\lambda^t \left\|X(1)\right\| + 2C \sum_{s=0}^{t-1} \lambda^s \left\|\widehat{X}(t-s+1) - X(t-s)\right\|.
\end{aligned} \tag{14}$$

Furthermore by the iterations of Algorithm 11 as well as the assumption on quantization noise in Assumption 3 we find that

$$\begin{aligned}
\mathbb{E}\left\|X(t+1) - \widehat{X}(t+2)\right\| &= \mathbb{E}\left\|X(t+1) - \widehat{X}(t+1) - Q(t+1)\right\|\\
&\leq \omega\left\|X(t+1) - \widehat{X}(t+1)\right\|\\
&= \omega\left\|X(t) + (A-I)\widehat{X}(t+1) - \widehat{X}(t+1)\right\|\\
&\leq \omega\left\|X(t) - \widehat{X}(t+1)\right\| + \omega\left\|(A-I)\widehat{X}(t+1)\right\|.
\end{aligned}$$

Next, we add and subtract $(A-I)X(t)$ to the RHS and also use the fact that $A\boldsymbol{\phi}=\boldsymbol{\phi}$ ((Zeng & Yin, 2015)) to conclude that

$$\begin{aligned}\mathbb{E}\left\|X(t+1)-\widehat{X}(t+2)\right\| \leq \omega\left\|X(t)-\widehat{X}(t+1)\right\| + \omega\left\|(A-I)(\widehat{X}(t+1)-X(t))\right\| \\ + \omega\left\|(A-I)\left(X(t)-\boldsymbol{\phi}\mathbf{1}^T X(1)\right)\right\|.\end{aligned} \tag{15}$$

Let $\gamma := \|A-I\|$, $\lambda_1 := \omega(1+\gamma)$ and $\lambda_1' := \omega\gamma$, then we conclude that

$$\mathbb{E}\left\|X(t+1)-\widehat{X}(t+2)\right\| = \lambda_1\left\|\widehat{X}(t+1)-X(t)\right\| + \lambda_1'\left\|X(t)-\boldsymbol{\phi}\mathbf{1}^T X(1)\right\|.$$

Denoting by $R(t) := \mathbb{E}\left\|X(t)-\boldsymbol{\phi}\mathbf{1}^T X(1)\right\|$ and $U(t)=\mathbb{E}\left\|\widehat{X}(t+1)-X(t)\right\|$, we derive the next two inequalities based on (14) and (15):

$$\begin{cases} \mathbb{E}\, R(t+1) \leq C\cdot\lambda^t\|X(1)\| + 2\,C\sum_{s=0}^{t-1}\lambda^s U(t-s) \\ \mathbb{E}\, U(t+1) \leq \lambda_1 U(t) + \lambda_1' R(t) \end{cases} \tag{16}$$

Lemma 8.1. *The iterates of $U(t)$ satisfy for all iterations $t \geq 1$*

$$U(t) \leq \xi_1 \lambda^{t/2},$$

where $\xi_1 = \max\{\frac{4C}{\lambda}\lambda_1\|X(1)\|,\ \frac{\lambda_1\|X(1)\|}{\lambda^{1/2}}\}$ and for the values of λ_1 chosen such that $\lambda_1 \leq \frac{1}{2}(\frac{1}{\lambda^{1/2}}+\frac{2C}{\lambda-\lambda^{3/2}})^{-1}$.

Proof. Noting that $\lambda_1' \leq \lambda_1$, we have by (16) for all $t \geq 1$:

$$U(t+1) \leq \lambda_1 U(t) + 2C\cdot\lambda_1\left(\lambda^{t-1}\|X(1)\| + \sum_{s=0}^{t-2}\lambda^s U\left(t-s-1\right)\right). \tag{17}$$

The proof is based on induction on the inequality in (17). Let $U(t) \leq \xi_1\cdot\lambda^{t/2}$, then by (17)

$$\begin{aligned} U(t+1) &\leq \lambda_1\cdot\xi_1\cdot\lambda^{t/2} + 2C\lambda_1\left(\lambda^{t-1}\|X(1)\|\right) + 2C\lambda_1\cdot\xi_1\cdot\frac{\lambda^{\frac{t-1}{2}}}{1-\lambda^{1/2}} \\ &= \frac{\lambda_1\cdot\xi_1}{\lambda^{1/2}}\lambda^{\frac{t+1}{2}} + 2C\lambda_1\left(\|X(1)\|\cdot\lambda^{\frac{t-3}{2}}\right)\lambda^{\frac{t+1}{2}} + \frac{2C\cdot\lambda_1\cdot\xi_1}{\lambda-\lambda^{3/2}}\lambda^{\frac{t+1}{2}} \\ &\leq \lambda_1\xi_1\left(\frac{1}{\lambda^{1/2}}+\frac{2C}{\lambda-\lambda^{3/2}}\right)\lambda^{\frac{t+1}{2}} + \frac{2C\cdot\lambda_1\|X(1)\|}{\lambda}\cdot\lambda^{\frac{t+1}{2}} \\ &\leq \left(\frac{1}{2}\xi_1 + \frac{2C\cdot\lambda_1\|X(1)\|}{\lambda}\right)\cdot\lambda^{\frac{t+1}{2}} \leq \xi_1\lambda^{\frac{t+1}{2}}, \end{aligned}$$

where the last two steps follow by the assumptions of the lemma on ξ_1 and λ_1. Moreover for $U(1)$ we follow the inequalities similar to (15) to find that

$$U(1) \leq \omega\|X(1)\| \leq \xi_1\lambda^{1/2},$$

where we used $\omega \leq \lambda_1$ in the last inequality. This completes the proof of the lemma. □

Lemma 8.1 implies that the quantization error $U(t)$ is decaying with the rate $\lambda^{t/2}$. As we will see shortly, this results in total error, decaying with the same rate. From the update rule in Algorithm 11, we obtain the following for all $t \geq 0$:

$$\mathbf{y}(t+1) = A\mathbf{y}(t) = A^t\mathbf{y}(1).$$

Since by assumption $\mathbf{y}(1) = \mathbf{1}$ it yields that

$$\begin{aligned}\mathbf{y}(t+1) &= A^t\mathbf{1} = \left(A^t - \boldsymbol{\phi}\mathbf{1}^T\right)\mathbf{1} + \boldsymbol{\phi}\mathbf{1}^T\mathbf{1} \\ &= \left(A^t - \boldsymbol{\phi}\mathbf{1}^T\right)\mathbf{1} + \boldsymbol{\phi} n.\end{aligned}$$

Therefore for all $i \in [n]$

$$\mathbf{y}_i(t+1) = \left[\left(A^t - \boldsymbol{\phi}\mathbf{1}^T\right)\right]_i + \boldsymbol{\phi}_i n.$$

Furthermore, based on Algorithm 11 the parameter $\mathbf{z}_i(t+1)$ satisfies:

$$\mathbf{z}_i(t+1) = \frac{\mathbf{x}_i(t+1)}{y_i(t+1)} = \frac{\left[A^t X(1) + \sum_{s=0}^{t-1} A^s (A - I)\left(\widehat{X}\left(t-s+1\right) - X(t-s)\right)\right]_i}{\left[\left(A^t - \boldsymbol{\phi}\mathbf{1}^T\right)\mathbf{1}\right]_i + \boldsymbol{\phi}_i n}. \tag{18}$$

Using this, we find the following expression for the vector representing error of node i :

$$\begin{aligned}&\mathbf{z}_i(t+1) - \frac{\mathbf{1}^T X(1)}{n} \\ &= \frac{\left[A^t X(1) + \sum_{s=0}^{t-1} A^s (A - \mathbf{I})(\widehat{X}(t-s+1) - X(t-s))\right]_i}{n\left(\left[(A^t - \boldsymbol{\phi}\mathbf{1}^T)\mathbf{1}\right]_i + \boldsymbol{\phi}_i n\right)} - \frac{\mathbf{1}^T X(1)\left(\left[\left(A^t - \boldsymbol{\phi}\mathbf{1}^T\right)\mathbf{1}\right]_i + \boldsymbol{\phi}_i n\right)}{n\left(\left[(A^t - \boldsymbol{\phi}\mathbf{1}^T)\mathbf{1}\right]_i + \boldsymbol{\phi}_i n\right)} \\ &= \frac{n\left[A^t - \boldsymbol{\phi}\mathbf{1}^T\right]_i X(1) + n\sum_{s=0}^{t-1}\left[A^s(A - \mathbf{I})\right]_i\left(\widehat{X}(t-s+1) - X(t-s)\right) - \mathbf{1}^T X(1)\left[(A^t - \boldsymbol{\phi}\mathbf{1}^T)\right]_i \mathbf{1}}{n\left(\left[A^t - \boldsymbol{\phi}\mathbf{1}^T\right]_i \mathbf{1} + \boldsymbol{\phi}_i n\right)}.\end{aligned}$$

By Proposition 2.1 it holds that $A^t\mathbf{1} \geq \delta$ and $\left\|[A^t - \boldsymbol{\phi}\mathbf{1}^T]_i\right\| \leq C\lambda^t$ for all $t \geq 1$; thus, we derive the following for the error of parameter of node i :

$$\begin{aligned}\mathbb{E}\left\|\mathbf{z}_i(t+1) - \frac{\mathbf{1}^T X(1)}{n}\right\| &\leq \frac{C}{\delta}\cdot\lambda^t\left\|X(1)\right\| \quad + \frac{C}{\delta}\sum_{s=0}^{t-1}\lambda^s\cdot\left(\xi_1\cdot\lambda^{\frac{t-s}{2}}\right) + \frac{C}{n\delta}\lambda^t\left\|\mathbf{1}^T X(1)\right\| \\ &\leq \frac{C}{\delta}\cdot\lambda^t\left\|X(1)\right\| + \frac{C}{n\delta}\lambda^t\left\|\mathbf{1}^T X(1)\right\|\cdot\sqrt{n} + \frac{C}{\delta}\xi_1\cdot\frac{\lambda^{t/2}}{1-\lambda^{1/2}}.\end{aligned}$$

Note that $\|\mathbf{1}^T X(1)\| \leq \|X(1)\|\cdot\sqrt{n}$. Thus,

$$\mathbb{E}\left\|\mathbf{z}_i(t+1) - \tfrac{\mathbf{1}^T X(1)}{n}\right\| \leq \tfrac{2C}{\delta}\lambda^t\left\|X(1)\right\| + \tfrac{C}{\delta}\cdot\tfrac{\xi_1}{1-\lambda^{1/2}}\lambda^{t/2}.$$

Rewriting the condition on λ_1 in Lemma 8.1 based on ω, we derive the inequality in the statement of Theorem 5.1.

9. Proof of Theorem 5.2 : Quantized Push-sum with Convex Objectives

First we write the iterations of Algorithm 2 in matrix notation to derive the following:

$$\begin{cases} Q(t) = Q\left(X(t) - \widehat{X}(t)\right) \\ \widehat{X}(t+1) = \widehat{X}(t) + Q(t) \\ W(t+1) = X(t) + (A - I)\widehat{X}(t+1) \\ \mathbf{y}(t+1) = A\mathbf{y}(t) \\ \mathbf{z}_i(t+1) = \frac{\mathbf{w}_i(t+1)}{y_i(t+1)} \\ X(t+1) = W(t+1) - \alpha\,\partial F(Z(t+1), \zeta_{t+1} \end{cases} \tag{19}$$

Similar to (14), we can rewrite the iterations of $X(t)$ to obtain the following expression:

$$\begin{aligned}\left\|X(t+1)-\boldsymbol{\phi}\mathbf{1}^TX(t)\right\|^2 \le 3C^2\lambda^{2t}\left\|X(1)\right\|^2 + 3C^2\alpha^2\left\|\sum_{s=0}^{t}\lambda^s\left\|\partial F(Z(t-s+1),\zeta_{t-s+1})\right\|\right\|^2 \\ + 3C^2\left\|\sum_{s=0}^{t-1}\lambda^s\left\|\widehat{X}(t-s+1)-X(t-s)\right\|\right\|^2\end{aligned} \tag{20}$$

For the last term, we expand the summation as well as simplifying the resulting expressions to obtain:

$$\begin{aligned}&\left\|\sum_{s=0}^{t-1}\lambda^s\left\|\widehat{X}(t-s+1)-X(t-s)\right\|\right\|^2 \\ &= \sum_{s=0}^{t-1}\lambda^{2s}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 + \sum_{s\neq s'}\lambda^s\cdot\lambda^{s'}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|\left\|\widehat{X}\left(t-s'+1\right)-X\left(t-s'\right)\right\|\end{aligned}$$

Using the relation $x\cdot y\le x^2/2+y^2/2$ for all $x,y\in\mathbb{R}$, as well as the inequality $\lambda^{2s}\le\frac{\lambda^s}{1-\lambda}$, from the equation above the following yields:

$$\begin{aligned}&\left\|\sum_{s=0}^{t-1}\lambda^s\left\|\widehat{X}(t-s+1)-X(t-s)\right\|\right\|^2 \\ &\le \sum_{s=0}^{t-1}\lambda^{2s}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 + 1/2\sum_{s\neq s'}\lambda^{s+s'}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 \\ &\quad + 1/2\sum_{s\neq s'}\lambda^s\cdot\lambda^{s'}\left\|\widehat{X}(t-s'+1)-X(t-s')\right\|^2 \\ &\le \sum_{s=0}^{t-1}\left(\lambda^{2s}+\frac{\lambda^s}{1-\lambda}\right)\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 \\ &\le \sum_{s=0}^{t-1}\frac{2\lambda^s}{1-\lambda}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2.\end{aligned}$$

Using a similar approach as above, we can bound the second term in the RHS of (20) as follows:

$$\left\|\sum_{s=0}^{t}\lambda^s\left\|\partial F(Z(t-s+1),\zeta_{t-s+1})\right\|\right\|^2 \le \sum_{s=0}^{t}\frac{2\lambda^s}{1-\lambda}\left\|\partial F(Z(t-s+1),\zeta_{t-s+1})\right\|^2.$$

Thus from (20) with the assumption that $X(1)=0$ (as in Assumption 4) we find that

$$\begin{aligned}&\mathbb{E}\left\|X(t+1)-\boldsymbol{\phi}\mathbf{1}^TX(t)\right\|^2 \\ &\le 6C^2\mathbb{E}\sum_{s=0}^{t-1}\frac{\lambda^s}{1-\lambda}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 + 6C^2\alpha^2\sum_{s=0}^{t-1}\frac{\lambda^s}{1-\lambda}\mathbb{E}\left\|\partial F(Z(t-s+1),\zeta_{t-s+1})\right\|^2 \\ &\le 6C^2\sum_{s=0}^{t-1}\frac{\lambda^s}{1-\lambda}\mathbb{E}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2 + \frac{6C^2\alpha^2nD^2}{(1-\lambda)^2},\end{aligned} \tag{21}$$

where we used Assumption 6 to bound the second term in the RHS of (21). Moreover, similar to the chain of inequalities in

(15) derived for Algorithm 11, we derive the following inequality here in the presence of stochastic gradients:

$$\begin{aligned}
&\mathbb{E}\Big\|X(t+1)-\widehat{X}(t+2)\Big\|^2\\
&\leq 3\,\omega^2\left(1+\gamma^2\right)\mathbb{E}\Big\|X(t)-\widehat{X}(t+1)\Big\|^2+3\,\omega^2\gamma^2\mathbb{E}\Big\|X(t)-\phi\mathbf{1}^TX(t-1)\Big\|^2\\
&\qquad+3\,\omega^2\alpha^2\mathbb{E}\Big\|\partial F(Z(t+1),\zeta_{t+1})\Big\|^2.
\end{aligned}\tag{22}$$

Let $\lambda_2:=\omega^2(1+\gamma^2)$. Then, from (22) as well as bounding stochastic gradients according to Assumption 6, it directly follows that

$$\mathbb{E}\Big\|X(t+1)-\widehat{X}(t+2)\Big\|^2\leq 3\lambda_2\left(\mathbb{E}\Big\|X(t)-\widehat{X}(t+1)\Big\|^2+\mathbb{E}\Big\|X(t)-\boldsymbol{\phi}\mathbf{1}^TX(t-1)\Big\|^2+n\alpha^2D^2\right).\tag{23}$$

Let $R(t+1):=\mathbb{E}\Big\|X(t+1)-\boldsymbol{\phi}\mathbf{1}^TX(t)\Big\|^2$ and $U(t+1):=\mathbb{E}\Big\|X(t+1)-\widehat{X}(t+2)\Big\|^2$. Then noting (21) and (23), we can rewrite the corresponding errors as follows:

$$\begin{cases}\mathbb{E}R(t+1)\leq\frac{6C^2n\alpha^2D^2}{(1-\lambda)^2}+\frac{6C^2}{1-\lambda}\sum_{s=0}^{t-1}\lambda^sU(t-s),\\ \mathbb{E}U(t+1)\leq 3\lambda_2\left(U(t)+R(t)+n\alpha^2D^2\right).\end{cases}\tag{24}$$

Next, we state the following lemma which shows that the error of quantization i.e. $U(t)$ decays proportionately with α^2 .

Lemma 9.1. *Under Assumption 4, the inequalities in (24) satisfy the following for all $\lambda_2\leq(\frac{1}{6}+\frac{C^2}{(1-\lambda)^2})^{-1}$ and all iterations $t\geq 1$*

$$U(t)\leq\xi_2\,\alpha^2,\tag{25}$$

where $\xi_2=6\lambda_2\left(nD^2+\frac{6nC^2D^2}{(1-\lambda)^2}\right)$.

Proof. First we write the inequalities in (24) based on $U(\cdot)$ to obtain:

$$U(t+1)\leq 3\lambda_2\Big(U(t)+n\alpha^2D^2+\frac{6nC^2\alpha^2D^2}{(1-\lambda)^2}+\frac{6C^2}{1-\lambda}\sum_{s=0}^{t-2}\lambda^sU(t-s-1)\Big).$$

Thus, by the assumption of induction we obtain the following for $U(t+1)$:

$$\begin{aligned}
U(t+1)&\leq 3\lambda_2\left(\xi_2\alpha^2+n\alpha^2D^2+\frac{6C^2\alpha^2nD^2}{(1-\lambda)^2}+\frac{6C^2\xi_2\alpha^2}{(1-\lambda)}\sum_{s=0}^{t-2}\lambda^s\right)\\
&\leq 3\lambda_2\alpha^2\xi_2\left(1+\frac{6C^2}{(1-\lambda)^2}\right)+3\lambda_2\alpha^2\left(nD^2+\frac{6nC^2D^2}{(1-\lambda)^2}\right)\\
&\leq\frac{\xi_2\,\alpha^2}{2}+3\lambda_2\alpha^2\left(nD^2+\frac{6nC^2D^2}{(1-\lambda)^2}\right)=\xi_2\,\alpha^2,
\end{aligned}\tag{26}$$

where the last two steps follow from the assumptions for λ_2 and ξ_2, respectively. Note that based on iterations of the algorithm and Assumption 4 we conclude that $\|U(1)\|=0$. This completes the proof of the lemma. □

Lemma 9.2. *Under Assumptions 1-7, if $\lambda_2\leq(\frac{1}{6}+\frac{C^2}{(1-\lambda)^2})^{-1}$, the following relation holds for the consensus error of Algorithm 2 for all $t\geq 1$:*

$$\mathbb{E}\left\|\mathbf{z}_i(t+1)-\frac{\boldsymbol{1}^TX(t)}{n}\right\|^2\leq\frac{6\,C^2\alpha^2}{\delta^2(1-\lambda)^2}(2nD^2+\xi_2),$$

where $\xi_2=6\lambda_2\left(nD^2+\frac{6nC^2D^2}{(1-\lambda)^2}\right)$.

Proof. Using the update rule in Algorithm 19 and similar to (18) we derive the following for the vector corresponding to consensus error of node i :

$$\mathbf{z}_i\,(t+1) - \frac{\mathbf{1}^T X(t)}{n} =$$
$$\frac{\Big[\sum_{s=0}^{t-1} A^s(A-I)(\widehat{X}(t-s+1)-X(t-s)) - \alpha\sum_{s=0}^{t} A^s \partial F\Big(Z(t-s+1),\zeta_{t-s+1}\Big)\Big]_i}{\Big[(A^t-\boldsymbol{\phi}\mathbf{1}^T)\mathbf{1}\Big]_i + \boldsymbol{\phi}_i n}$$
$$+\frac{\alpha\mathbf{1}^T\sum_{s=0}^{t-1}\partial F\Big(Z(t-s+1),\zeta_{t-s+1}\Big)\Big(\Big[\Big(A^t-\boldsymbol{\phi}\mathbf{1}^T\Big)\mathbf{1}\Big]_i+\boldsymbol{\phi}_i n\Big)}{n\Big(\Big[(A^t-\boldsymbol{\phi}\mathbf{1}^T)\mathbf{1}\Big]_i+\boldsymbol{\phi}_i n\Big)}.$$

Note that by Proposition 2.1 for all $t \geq 1$ we have $[(A^t - \boldsymbol{\phi}\mathbf{1}^T)\mathbf{1}]_i + \boldsymbol{\phi}_i n = [A^t\mathbf{1}]_i \geq \delta$, which yields the following for squared norm of consensus error:

$$\mathbb{E}\left\|\mathbf{z}_i(t+1)-\frac{\mathbf{1}^T X(t)}{n}\right\|^2 \leq \frac{3}{\delta^2}\mathbb{E}\left\|\sum_{s=0}^{t}\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big)\right\|^2$$
$$+\frac{3\alpha^2}{\delta^2}\mathbb{E}\left\|\sum_{s=0}^{t}\Big[A^s-\boldsymbol{\phi}\mathbf{1}^T\Big]_i\partial F\Big(Z(t-s+1),\zeta_{t-s+1}\Big)\right\|^2 \tag{27}$$
$$+\frac{3\alpha^2}{n^2\delta^2}\mathbb{E}\left\|\mathbf{1}^T\left(\sum_{s=0}^{t-1}\Big[A^t-\boldsymbol{\phi}\mathbf{1}^T\Big]_i\partial F\Big(Z\left(t-s+1\right),\zeta_{t-s+1}\Big)\right)\right\|^2.$$

By expanding the first term in the RHS of (27) we derive

$$\left\|\sum_{s=0}^{t}\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big)\right\|^2 = \sum_{s=0}^{t}\left\|\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big)\right\|^2$$
$$+\sum_{s\neq s'}^{t}\left\langle\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big),\Big[A^s(A-I)\Big]_i\left(\widehat{X}\left(t-s'+1\right)-X\left(t-s'\right)\right)\right\rangle$$
$$\leq\sum_{s=0}^{t}\left\|\Big[A^s(A-I)\Big]_i\right\|^2\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2$$
$$+\sum_{s\neq s'}^{t}\left\|\Big[A^s(A-I)\Big]_i\right\|\left\|\widehat{X}(t-s+1)-X(t-s)\right\|\left\|\Big[A^s(A-I)\Big]_i\right\|\left\|\left(\widehat{X}\left(t-s'+1\right)-X\left(t-s'\right)\right)\right\|.$$

Using the relation $x \cdot y \leq x^2/2 + y^2/2$ for all $x, y \in \mathbb{R}$, this inequality reduces to the following:

$$\left\|\sum_{s=0}^{t}\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big)\right\|^2$$
$$\leq\sum_{s=0}^{t}\left\|\Big[A^s(A-I)\Big]_i\right\|^2\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2$$
$$+\frac{1}{2}\sum_{s\neq s'}^{t}\left\|\Big[A^s(A-I)\Big]_i\right\|\left\|\Big[A^{s'}(A-I)\Big]_i\right\|\left(\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2+\left\|\widehat{X}(t-s'+1)-X(t-s')\right\|^2\right).$$

Next we use Proposition 2.1 to yield that

$$\begin{aligned}
&\left\|\sum_{s=0}^{t}\Big[A^s(A-I)\Big]_i\Big(\widehat{X}(t-s+1)-X(t-s)\Big)\right\|^2 \leq \\
&C^2\sum_{s=0}^{t}\lambda^{2s}\Big\|\widehat{X}(t-s+1)-X(t-s)\Big\|^2 + C^2\sum_{s\neq s'}^{t}\lambda^{s+s'}\Big\|\widehat{X}(t-s+1)-X(t-s)\Big\|^2\\
&\leq C^2\sum_{s=0}^{t}\Big(\lambda^{2s}+\frac{\lambda^s}{1-\lambda}\Big)\Big\|\widehat{X}(t-s+1)-X(t-s)\Big\|^2\\
&\leq \frac{2C^2}{1-\lambda}\sum_{s=0}^{t}\lambda^s\Big\|\widehat{X}(t-s+1)-X(t-s)\Big\|^2,
\end{aligned}$$

where we used $\lambda^{2s}\leq\frac{\lambda^s}{1-\lambda}$ to derive the last inequality. Using the same approach for the second term in the RHS of (27), we derive the following upper bound:

$$\mathbb{E}\left\|\sum_{s=0}^{t}\Big[A^s-\boldsymbol{\phi}\mathbf{1}^T\Big]_i\partial F\Big(Z(t-s+1),\zeta_{t-s+1}\Big)\right\|^2 \leq \frac{2C^2}{1-\lambda}\sum_{s=0}^{t}\lambda^s\mathbb{E}\Big\|\partial F(Z(t-s+1),\zeta_{t-s+1})\Big\|^2.$$

To bound the third term in the RHS of (27), we use the same method,as well as the fact that $\|A^t-\boldsymbol{\phi}\mathbf{1}^T\|\leq\lambda^t\leq\lambda^s$ for all $s\leq t$ to deduce that

$$\mathbb{E}\left\|\mathbf{1}^T\left(\sum_{s=0}^{t-1}\Big[A^t-\boldsymbol{\phi}\mathbf{1}^T\Big]_i\partial F\Big(Z\left(t-s+1\right),\zeta_{t-s+1}\Big)\right)\right\|^2 \leq \frac{2nC^2}{1-\lambda}\sum_{s=0}^{t}\lambda^s\mathbb{E}\Big\|\partial F(Z(t-s+1),\zeta_{t-s+1})\Big\|^2.$$

Replacing these back in (27) gives

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{z}_i(t+1)-\frac{\mathbf{1}^TX(t)}{n}\right\|^2\\
&\leq\frac{6C^2}{\delta^2(1-\lambda)}\sum_{s=0}^{t}\lambda^s\mathbb{E}\Big\|\widehat{X}(t-s+1)-X(t-s)\Big\|^2\\
&\quad+\left(\frac{6\alpha^2C^2}{\delta^2(1-\lambda)}+\frac{6\alpha^2C^2}{n\delta^2(1-\lambda)}\right)\sum_{s=0}^{t}\lambda^s\mathbb{E}\Big\|\partial F(Z(t-s+1),\zeta_{t-s+1})\Big\|^2. \qquad (28)
\end{aligned}$$

Note that $U(t-s):=\mathbb{E}\left\|\widehat{X}(t-s+1)-X(t-s)\right\|^2\leq\xi_2\alpha^2$ by Lemma 9.1. Also by bounded stochastic gradient property in Assumption 6, we have $\mathbb{E}\left\|\partial F(Z(t-s+1),\zeta_{t-s+1})\right\|^2\leq nD^2$. Therefore we conclude the following from (28):

$$\mathbb{E}\left\|\mathbf{z}_i(t+1)-\frac{\mathbf{1}^TX(t)}{n}\right\|^2\leq\frac{6\,\xi_2\,C^2\alpha^2}{\delta^2(1-\lambda)^2}+\left(\frac{6\alpha^2C^2}{\delta^2(1-\lambda)^2}+\frac{6\alpha^2C^2}{n\delta^2(1-\lambda)^2}\right)nD^2. \qquad (29)$$

Simplifying relations with $1+1/n\leq 2$, yields the desired inequality in the statement of the lemma. □

We continue with proving the next lemma which relates the consensus error as stated in Lemma 9.2 to the error of global objective function f evaluated at the average of parameters $\mathbf{x}_i(t)$ of all nodes.

Lemma 9.3. *For all $t\geq 1$, iterations of Algorithm 19 satisfy:*

$$\begin{aligned}
\mathbb{E}\left\|\bar{X}(t+1)-\mathbf{z}^\star\right\|^2\leq\mathbb{E}\left\|\bar{X}(t)-\mathbf{z}^\star\right\|^2-\left(2\alpha-\frac{8\alpha^2L}{n}\right)\mathbb{E}(f\left(\bar{X}(t)\right)-f(\mathbf{z}^\star))\\
+\frac{2\alpha L+4L^2\alpha^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}_i(t+1)-\bar{X}(t)\right\|+\frac{2\sigma^2\alpha^2}{n}.
\end{aligned}$$

Proof. First we recall that A is column stochastic; thus, it yields that $\mathbf{1}^T\left(A-I\right)=0$. Therefore, iterations of Algorithm 19 yield that

$$\bar{X}(t+1)=\bar{X}(t)-\frac{\alpha}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1}).$$

Thus,

$$\begin{aligned}\mathbb{E}\left\|\bar{X}(t+1)-\mathbf{z}^\star\right\|^2=&\mathbb{E}\left\|\bar{X}(t)-\mathbf{z}^\star\right\|^2-\frac{2\alpha}{n}\sum_{i=1}^{n}\mathbb{E}\left\langle\nabla f_i\left(\mathbf{z}_i(t+1)\right),\bar{X}(t)-\mathbf{z}^\star\right\rangle\\&+\alpha^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1})\right\|^2.\end{aligned}\tag{30}$$

We derive an upper bound for the third term by adding and subtracting $\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i(t+1))$:

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1})\right\|^2&\le 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1})-\nabla f_i(\mathbf{z}_i(t+1))\right)\right\|^2\\&\quad+2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i(t+1))\right\|^2\\&\le\frac{2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1})-\nabla f_i(\mathbf{z}_i(t+1))\right\|^2\\&\quad+2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i(t+1))\right\|^2\le\frac{2\sigma^2}{n}+2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i(t+1))\right\|^2.\end{aligned}\tag{31}$$

For the last term in (31) we have

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i(t+1))\right\|^2&\le 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(\mathbf{z}_i(t+1))-\nabla f_i(\bar{X}(t)))\right\|^2+2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(\bar{X}(t))-\nabla f_i(\mathbf{z}^\star))\right\|^2\\&\le\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{z}_i(t+1))-\nabla f_i(\bar{X}(t))\right\|^2+2\mathbb{E}\left\|\nabla f(\bar{X}(t))-\nabla f(\mathbf{z}^\star)\right\|^2\\&\le\frac{2L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}_i(t+1)-\bar{X}(t)\right\|^2+\frac{4L}{n}(\mathbb{E}f(\bar{X}(t))-f(\mathbf{z}^\star)).\end{aligned}$$

Replacing this back in (31) we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{z}_i(t+1),\zeta_{i,t+1})\right\|^2\le\frac{2\sigma^2}{n}+\frac{4L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}_i(t+1)-\bar{X}(t)\right\|^2+\frac{8L}{n}(\mathbb{E}f(\bar{X}(t))-f(\mathbf{z}^\star)).\tag{32}$$

Next, we will achieve a bound for the second term in the RHS of (30). By adding and subtracting $\mathbf{z}_i(t+1)$ we have

$$\mathbb{E}\left\langle\nabla f_i(\mathbf{z}_i(t+1)),\bar{X}(t)-\mathbf{z}^\star\right\rangle=\mathbb{E}\left\langle\nabla f_i(\mathbf{z}_i(t+1)),\bar{X}(t)-\mathbf{z}_i(t+1)\right\rangle+\mathbb{E}\left\langle\nabla f_i(\mathbf{z}_i(t+1)),\mathbf{z}_i(t+1)-\mathbf{z}^\star\right\rangle,\tag{33}$$

where after using L-smoothness of the function f_i for the first term and convexity of the function f_i for the second term of (33) we deduce that

$$\begin{aligned}\mathbb{E}\left\langle\nabla f_i(\mathbf{z}_i(t+1)),\bar{X}(t)-\mathbf{z}^\star\right\rangle&\ge\mathbb{E}f_i(\bar{X}(t))-\mathbb{E}f_i(\mathbf{z}_i(t+1))-\frac{L}{2}\mathbb{E}\left\|\bar{X}(t)-\mathbf{z}_i(t+1)\right\|^2+\mathbb{E}f_i(\mathbf{z}_i(t+1))-f_i(\mathbf{z}^\star)\\&=\mathbb{E}f_i(\bar{X}(t))-f_i(\mathbf{z}^\star)-\frac{L}{2}\mathbb{E}\left\|\bar{X}(t)-\mathbf{z}_i(t+1)\right\|^2.\end{aligned}$$

Using this and recalling that $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$, we derive the following:

$$-\frac{2\alpha}{n}\sum_{i=1}^n \mathbb{E}\left\langle \nabla f_i(\mathbf{z}_i(t+1)), \bar{X}(t) - \mathbf{z}^\star \right\rangle \leq -2\alpha(\mathbb{E}\, f(\bar{X}(t)) - f(\mathbf{z}^\star)) + \frac{2\alpha L}{n}\sum_{i=1}^n \mathbb{E}\left\|\bar{X}(t) - \mathbf{z}_i(t+1)\right\|^2.$$

By replacing this result and (32) in (30) we derive the desired inequality in the statement of the lemma. □

We continue with rearranging and summing both sides of Lemma 9.3 for $t = 1, \cdots, T$ to find the following:

$$\begin{aligned}&\left(2\alpha - \frac{8\alpha^2 L}{n}\right)\sum_{t=1}^T (\mathbb{E}f(\bar{X}(t)) - f(\mathbf{z}^\star))\\ &\leq \left\|\bar{X}(1) - \mathbf{z}^\star\right\|^2 + \frac{2\alpha L + 4\alpha^2 L^2}{n}\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\left\|\mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2 + \sum_{t=1}^T \frac{2\sigma^2\alpha^2}{n}.\end{aligned}$$

Scaling both sides, as well as noting the Assumption 4 i.e. $X(1) = 0$, we find that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T (\mathbb{E}f(\bar{X}(t)) - f(\mathbf{z}^\star))\\ &\leq \frac{1}{2T\alpha(1 - \frac{4\alpha L}{n})}\left\|\mathbf{z}^\star\right\|^2 + \frac{2\alpha L^2 + L}{nT(1 - \frac{4\alpha L}{n})}\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\left\|\mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2 + \frac{\alpha\sigma^2}{n(1 - \frac{4\alpha L}{n})}.\end{aligned}$$

Using convexity of $f(\cdot)$, the above inequality simplifies to the following:

$$\begin{aligned}&\mathbb{E}\, f\left(\frac{1}{T}\sum_{t=1}^T \bar{X}(t)\right) - f(\mathbf{z}^\star)\\ &\leq \frac{1}{2T\alpha(1 - \frac{4\alpha L}{n})}\left\|\mathbf{z}^\star\right\|^2 + \frac{2\alpha L^2 + L}{nT(1 - \frac{4\alpha L}{n})}\sum_{t=1}^T\sum_{i=1}^n \mathbb{E}\left\|\mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2 + \frac{\alpha\sigma^2}{n(1 - \frac{4\alpha L}{n})},\end{aligned}$$

which after replacing the consensus error from Lemma 9.2, further simplifies into

$$\begin{aligned}&\mathbb{E}f\left(\frac{1}{T}\sum_{t=1}^T \bar{X}(t)\right) - f(\mathbf{z}^\star)\\ &\leq \frac{1}{2T\alpha(1 - \frac{4\alpha L}{n})}\left\|\mathbf{z}^\star\right\|^2 + \frac{6\alpha^2 C^2\left(2nD + \xi_2\right)}{\delta^2(1-\lambda)^2}\cdot\frac{2\alpha L^2 + L}{1 - \frac{4\alpha L}{n}} + \frac{\alpha\sigma^2}{n(1 - \frac{4\alpha L}{n})}.\end{aligned} \tag{34}$$

The inequality in (34) guarantees the convergence of time average of $\bar{X}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(t)$ to the optimal point $\mathbf{z}^\star$. However computing the average of $\mathbf{x}_i(t)$ between workers in every iteration is time consuming since it can not be done in the decentralized setting. Next we show that the time average of the local variables, $\mathbf{z}_i(t)$ converges to an optimal $\mathbf{z}^\star$, for every node i . First, By L-smoothness of the function $f(\cdot)$ as well as the inequality $\langle \mathbf{x}, \mathbf{y}\rangle \leq \frac{1}{2}\|\mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{y}\|^2$, we derive for all $i \in [n]$ it holds that

$$\begin{aligned}&f\left(\frac{1}{T}\sum_{t=1}^T \mathbf{z}_i(t+1)\right) - f\left(\frac{1}{T}\sum_{t=1}^T \bar{X}(t)\right)\\ &\leq \left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{z}_i(t+1) - \bar{X}(t), \nabla f\left(\frac{1}{T}\sum_{t=1}^T \bar{X}(t)\right)\right\rangle + \frac{L}{2T^2}\left\|\sum_{t=1}^T \mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2\\ &\leq \frac{1}{2T^2}\left\|\sum_{t=1}^T \mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2 + \frac{1}{2}\left\|\nabla f\left(\frac{1}{T}\sum_{t=1}^T \bar{X}(t)\right)\right\|^2 + \frac{L}{2T^2}\left\|\sum_{t=1}^T \mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2.\end{aligned} \tag{35}$$

Note that for the optimal solution $\mathbf{z}^\star$ it holds that $\|\nabla f(\mathbf{x})\|^2 = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{z}^\star)\|^2 \leq 2L\left(f(\mathbf{x}) - f(\mathbf{z}^\star)\right)$. Using this inequality for the second term in (35) we conclude that

$$
\begin{aligned}
&\mathbb{E}\, f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_i(t+1)\right) - \mathbb{E}\, f\left(\frac{1}{T}\sum_{t=1}^{T}\bar{X}(t)\right) \\
&\leq \left(\frac{1}{2T} + \frac{L}{2T}\right)\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{z}_i(t+1) - \bar{X}(t)\right\|^2 + L\left(\mathbb{E}\, f\left(\frac{1}{T}\sum_{t=1}^{T}\bar{X}(t)\right) - f(\mathbf{z}^\star)\right).
\end{aligned} \tag{36}
$$

By replacing the consensus error as derived in (29) and combining the inequalities (34) and (36), we get the convergence error of the time average of local variables $\mathbf{z}_i$:

$$
\begin{aligned}
&\mathbb{E}f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_i(t+1)\right) - f(\mathbf{z}^\star) \\
&\leq \frac{L+1}{2T\alpha(1-\frac{4\alpha L}{n})}\|\mathbf{z}^\star\|^2 + \left(\frac{(2\alpha L^2 + L)(L+1)}{1-\frac{4\alpha L}{n}} + \frac{1+L}{2}\right)\left(\frac{6\alpha^2 C^2\left(2nD + \xi_2\right)}{\delta^2(1-\lambda)^2}\right) + \frac{\alpha\sigma^2(L+1)}{n(1-\frac{4\alpha L}{n})}.
\end{aligned} \tag{37}
$$

We choose $\alpha = \frac{\sqrt{n}}{8L\sqrt{T}}$ in (37), which results in $\left(1 - \frac{4\alpha L}{n}\right)^{-1} \leq 2$ for all $T \geq 1$ and $n \geq 1$. Furthermore,

$$
\begin{aligned}
&\mathbb{E}f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_i(t+1)\right) - f(\mathbf{z}^\star) \\
&\leq \frac{8L(L+1)}{\sqrt{nT}}\|\mathbf{z}^\star\|^2 + \frac{\sigma^2(L+1)}{4\,L\sqrt{nT}} + \frac{C^2 n\left((L+1)\left(\frac{L\sqrt{n}}{2\sqrt{T}} + L + 1\right)\right)\left(\xi_2 + 2nD^2\right)}{10\,T\delta^2(1-\lambda)^2 L^2}.
\end{aligned} \tag{38}
$$

Replacing ξ_2 and rewriting the condition on λ_2 with ω (as required by Lemmas 9.1 and 9.2) completes the proof.

10. Proof of Theorem 5.3 : Quantized Push-sum with Non-convex Objectives

Using the L-smoothness of the global objective function which is implied by Assumption 5, we have for all $t \geq 1$:

$$
\begin{aligned}
\mathbb{E}f\left(\frac{\mathbf{1}^T X(t+1)}{n}\right) &= \mathbb{E}\, f\left(\frac{\mathbf{1}^T X(t)}{n} - \alpha\frac{\mathbf{1}^T \partial F(Z(t+1, \zeta_{t+1}))}{n}\right) \\
&= \mathbb{E}\, f\left(\frac{\mathbf{1}^T X(t)}{n}\right) - \alpha\,\mathbb{E}\left\langle \nabla f\left(\frac{\mathbf{1}^T X(t)}{n}\right), \frac{\mathbf{1}^T \partial f(Z(t+1))}{n}\right\rangle \\
&\quad + \frac{\alpha^2 L}{2}\,\mathbb{E}\left\|\frac{\mathbf{1}^T \partial F(Z(t+1, \zeta_{t+1}))}{n}\right\|^2.
\end{aligned} \tag{39}
$$

For the last term in the RHS of (39) we add and subtract $\frac{\mathbf{1}^T}{n}\partial f(Z(t+1))$ to yield

$$
\begin{aligned}
\mathbb{E}\left\|\frac{\mathbf{1}^T \partial F(Z(t+1), \zeta_{t+1})}{n}\right\|^2 &= \mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1), \zeta_{t+1}) - \partial f(Z(t+1))\Big) + \frac{\mathbf{1}^T}{n}\partial f(Z(t+1))\right\|^2 \\
&= \mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1), \zeta_{t+1}) - \partial f(Z(t+1))\Big)\right\|^2 + \mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\partial f(Z(t+1))\right\|^2 \\
&\quad + 2\,\mathbb{E}\left\langle \frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1), \zeta_{t+1}) - \partial f(Z(t+1))\Big), \frac{\mathbf{1}^T}{n}\partial f\left(Z(t+1)\right)\right\rangle.
\end{aligned}
$$

Since stochastic gradients of all nodes are unbiased estimators of the local gradients, the last term is zero in expectation. Thus,

$$\begin{aligned}\mathbb{E}\left\|\frac{\mathbf{1}^T\partial F(Z(t+1),\zeta_{t+1})}{n}\right\|^2 \leq &\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1),\zeta_{t+1})-\partial f(Z(t+1))\Big)\right\|^2+\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\partial f(Z(t+1))\right\|^2\\ &+2\mathbb{E}\left\langle\mathbb{E}_{\zeta_{t+1}}\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1),\zeta_{t+1})-\partial f(Z(t+1))\Big),\frac{\mathbf{1}^T}{n}\partial f\left(Z(t+1)\right)\right\rangle\\ =&\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1),\zeta_{t+1})-\partial f(Z(t+1))\Big)\right\|^2+\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\partial f(Z(t+1))\right\|^2.\end{aligned}$$

Next, by expanding and using the fact that stochastic gradients are computed independently among different nodes, we show that the first term in the equation above is bounded:

$$\begin{aligned}&\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\Big(\partial F(Z(t+1),\zeta_{t+1})-\partial f(Z(t+1))\Big)\right\|^2=\frac{1}{n^2}\mathbb{E}\left\|\sum_{i=1}^n\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right)\right\|^2\\ &=\frac{1}{n^2}\mathbb{E}\sum_{i=1}^n\left\|\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right)\right\|^2\\ &+\frac{1}{n^2}\mathbb{E}\sum_{i\neq i'}\left\langle\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right),\nabla F_{i'}\left(\mathbf{z}_{i'}(t+1),\zeta_{i',t+1}\right)-\nabla f_{i'}\left(\mathbf{z}_{i'}(t+1)\right)\right\rangle\\ &=\frac{1}{n^2}\mathbb{E}\sum_{i=1}^n\left\|\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right)\right\|^2\\ &+\frac{1}{n^2}\mathbb{E}\sum_{i\neq i'}\left\langle\mathbb{E}_{\zeta_{i,t+1}}\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right),\nabla F_{i'}\left(\mathbf{z}_{i'}(t+1),\zeta_{i',t+1}\right)-\nabla f_{i'}\left(\mathbf{z}_{i'}(t+1)\right)\right\rangle\\ &=\frac{1}{n^2}\mathbb{E}\sum_{i=1}^n\left\|\nabla F_i\left(\mathbf{z}_i(t+1),\zeta_{i,t+1}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right)\right\|^2\leq\frac{\sigma^2}{n},\end{aligned}$$

where we recall Assumption 7 in the last inequality. Next, we rewrite (39) using the new terms as follows:

$$\begin{aligned}\mathbb{E}f\left(\frac{\mathbf{1}^TX(t+1)}{n}\right)\leq\mathbb{E}f\left(\frac{\mathbf{1}^TX(t)}{n}\right)-\alpha\,\mathbb{E}\left\langle\nabla f\left(\frac{\mathbf{1}^TX(t)}{n}\right),\frac{\mathbf{1}^T\partial f(Z(t+1))}{n}\right\rangle\\ +\frac{\alpha^2L}{2}\left(\frac{\sigma^2}{n}+\mathbb{E}\left\|\frac{\mathbf{1}^T}{n}\partial f(Z(t+1))\right\|^2\right).\end{aligned}\tag{40}$$

Moreover, using the relation $\langle\mathbf{x},\mathbf{y}\rangle=\frac{1}{2}\|\mathbf{x}\|^2+\frac{1}{2}\|\mathbf{y}\|^2-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$, we find the following for the second term in the RHS of (40):

$$\begin{aligned}&\mathbb{E}\left\langle\nabla f\left(\frac{\mathbf{1}^TX(t)}{n}\right),\frac{\mathbf{1}^T\partial f(Z(t+1))}{n}\right\rangle\\ &=\frac{1}{2}\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^TX(t)}{n}\right)\right\|^2+\frac{1}{2}\mathbb{E}\left\|\frac{\mathbf{1}^T\partial f(Z(t+1))}{n}\right\|^2-\frac{1}{2}\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^TX(t)}{n}\right)-\frac{\mathbf{1}^T\partial f(Z(t+1))}{n}\right\|^2.\end{aligned}\tag{41}$$

Using L-lipschitz assumption of local gradients (Assumption 5), the last term in (41) reduces to

$$\begin{aligned}\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^TX(t)}{n}\right)-\frac{\mathbf{1}^T\partial f(Z(t+1))}{n}\right\|^2&\leq\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i\left(\frac{\mathbf{1}^TX(t)}{n}\right)-\nabla f_i\left(\mathbf{z}_i(t+1)\right)\right\|^2\\ &\leq\frac{L^2}{n}\sum_{i=1}^n\mathbb{E}\left\|\frac{\mathbf{1}^TX(t)}{n}-\mathbf{z}_i(t+1)\right\|^2,\end{aligned}$$

where we used $\nabla f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x})$ in the first step. Replacing this in (41) and finally substituting the resulting expression of (41) in (40) yields

$$\begin{aligned}\mathbb{E} f\left(\frac{\mathbf{1}^T X(t+1)}{n}\right) \leq \mathbb{E} f\left(\frac{\mathbf{1}^T X(t)}{n}\right) &+ \frac{\alpha^2\sigma^2 L}{2n} \\ &+ \frac{\alpha^2 L - \alpha}{2}\,\mathbb{E}\left\|\frac{\mathbf{1}^T \partial f(Z(t+1))}{n}\right\|^2 - \frac{\alpha}{2}\,\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^T X(t)}{n}\right)\right\|^2 \\ &+ \frac{\alpha L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{z}_i(t+1) - \frac{\mathbf{1}^T X(t)}{n}\right\|^2.\end{aligned} \tag{42}$$

By replacing the value of consensus error from Lemma 9.2 in (42), we obtain the following:

$$\begin{aligned}\mathbb{E} f\left(\frac{\mathbf{1}^T X(t+1)}{n}\right) \leq \mathbb{E} f\left(\frac{\mathbf{1}^T X(t)}{n}\right) &+ \frac{\alpha^2\sigma^2 L}{2n} \\ &+ \frac{\alpha^2 L - \alpha}{2}\,\mathbb{E}\left\|\frac{\mathbf{1}^T \partial f(Z(t+1))}{n}\right\|^2 - \frac{\alpha}{2}\,\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^T X(t)}{n}\right)\right\|^2 \\ &+ \frac{6\alpha^3 C^2 L^2}{\delta^2(1-\lambda)^2}\left(nD^2 + D^2 + \xi_2\right).\end{aligned} \tag{43}$$

Thus by rearranging terms in (43) and averaging both sides from $t = 1$ to $t = T$ we conclude that

$$\begin{aligned}&\frac{1-\alpha L}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\frac{\mathbf{1}^T \partial f(Z(t+1))}{n}\right\|^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f\left(\frac{\mathbf{1}^T X(t)}{n}\right)\right\|^2 \\ &\leq \frac{2\left(f\left(\frac{\mathbf{1}^T X(1)}{n}\right) - f^\star\right)}{\alpha T} + \frac{\alpha\sigma^2 L}{n} + \frac{12\alpha^2 L^2 C^2}{\delta^2(1-\lambda)^2}\left(\xi_2 + nD^2 + D^2\right).\end{aligned} \tag{44}$$

By choosing $\alpha = \frac{\sqrt{n}}{L\sqrt{T}}$ and noting that $X(1) = 0$ we derive the desired statement of the theorem. Note that $T \geq 4n$ implies that $\frac{1-\alpha L}{T} \geq \frac{1}{2T}$ guaranteeing that the first term in the LHS of (44) is positive. The consensus error in (10) is concluded from Lemma 9.2 with the given choice of step-size as in the statement of the theorem. This completes the proof of the theorem.