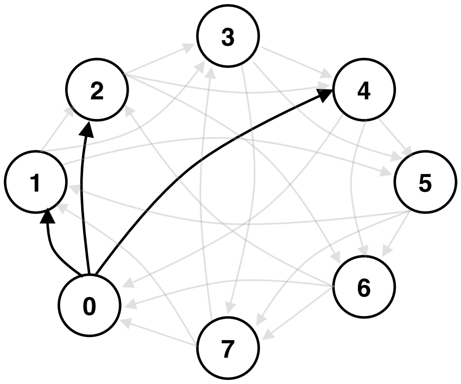
Supplementary Material
Stochastic Gradient Push for Distributed Deep Learning

A. Communication Topology

Directed exponential graph. For the SGP experiments we use a time-varying directed graph to represent the inter-node connectivity. Thinking of the nodes as being ordered sequentially, according to their rank, $0, \ldots, n-1$,[3] each node periodically communicates with peers that are $2^0, 2^1, \ldots, 2^{\lfloor \log_2(n-1) \rfloor}$ hops away. Fig. A.1 shows an example of a directed 8-node exponential graph. Node 0's 2^0-hop neighbour is node 1; node 0's 2^1-hop neighbour is node 2; and node 0's 2^2-hop neighbour is node 4.



(a) Directed Exponential Graph highlighting node 0's out-neighbours

Figure A.1: Example of an 8-node exponential graph used in experiments

In the one-peer-per-node experiments, each node cycles through these peers, transmitting, only, to a single peer from this list at each iteration. E.g., at iteration k, all nodes transmit messages to their 2^0-hop neighbours, at iteration $k+1$ all nodes transmit messages to their 2^1-hop neighbours, an so on, eventually returning to the beginning of the list before cycling through the peers again. This procedure ensures that each node only sends and receives a single message at each iteration. By using full-duplex communication, sending and receiving can happen in parallel.

In the two-peer-per-node experiments, each node cycles through the same set of peers, transmitting to two peers from the list at each iteration. E.g., at iteration k, all nodes transmit messages to their 2^0-hop and 2^1-hop neighbours, at iteration $k+1$ all nodes transmit messages to their 2^1-hop and 2^2 neighbours, an so on, eventually returning to the beginning of the list before cycling through the peers again. Similarly, at each iteration, each node also receives, in a full-duplex manner, two messages from some peers that are unknown to the receiving node ahead of time. Thereby performing the send and receive operations in parallel.

Definition of $\boldsymbol{P}^{(k)}$. We choose the mixing matrices such that they are column stochastic (all columns sum to 1), and conform to the graph topology described above. Recall that each node i can choose its mixing weights (i^{th} column of $\boldsymbol{P}^{(k)}$), independently of the other nodes in the network. To minimize the number of floating point operations in each iteration, we choose to use uniform mixing weights, meaning that nodes assign uniform message weights to all neighbours. In the one-peer-per-node experiments, each node sends a message to one neighbor, and "sends a message" to itself at every iteration, and so each column of $\boldsymbol{P}^{(k)}$ has exactly two non-zero entries, both of which are equal to $1/2$. The first set of non-zero entries corresponds to the diagonals. At all time steps k, the diagonal entries satisfy $p_{i,i}^{(k)} = 1/2$ for all i. The

[3] We use indices $0, \ldots, n-1$ rather than $1, \ldots, n$ only in this section, to simplify the discussion.

second set of non-zero entries correspond to the neighbor indices. At time step k, each node sends to a neighbor that is $h_k := 2^{k \bmod \lfloor \log_2(n-1) \rfloor}$ hops away. That is, at each time step k, each node i sends a message to node $(i + h_k) \bmod n$. Thus, we get that

$$p_{j,i}^{(k)} = \begin{cases} 1/2, & \text{if } j = (i + h_k) \bmod n \\ 0, & \text{otherwise.} \end{cases}$$

Note that, with this design, in fact each node sends to one peer and receives from one peer at every iteration, so the communication load is balanced across the network.

In the two-peer-per-node experiments, the definition is similar, but now there will be three non-zero entries in each column of $\boldsymbol{P}^{(k)}$, all of which will be equal to $1/3$; these are the diagonal, and the entries corresponding to the two neighbors to which the node sends at that iteration. In addition, each node will send two messages and receive two messages at every iteration, so the communication load is again balanced across the network.

Undirected exponential graph. For the D-PSGD experiments we use a time-varying undirected bipartite exponential graph to represent the inter-node connectivity. Odd-numbered nodes send messages to peers that are $2^1 - 1, 2^2 - 1, \ldots, 2^{\lfloor \log_2(n-1) \rfloor} - 1$ (even-numbered nodes), and wait to a receive a message back in return. Each odd-numbered node cycles through the peers in the list in a similar fashion to the one-peer-per-node SGP experiments. Even-numbered nodes wait to receive a message from some peer (unknown to the receiving node ahead of time), and send a message back in return.

We adopt these graphs to be consistent with the experimental setup used in Lian et al. (2017) and Lian et al. (2018).

Note also that these graphs are all regular, in that all nodes have the same number of in-coming and out-going connections.

Decentralized averaging errors. To further motivate our choice of using the directed exponential graph with SGP, let us forget about optimization for a moment and focus on the problem of distributed averaging, described in Section 2, using the PushSum algorithm. Recall that each node i starts with a vector $\boldsymbol{y}_i^{(0)}$, and the goal of the agents is to compute the average $\overline{\boldsymbol{y}} = \frac{1}{n} \sum_i \boldsymbol{y}_i^{(0)}$. Then, since $\boldsymbol{y}_i^{(k+1)} = \sum_{j=1}^n p_{i,j}^{(k)} \boldsymbol{y}_j^{(k)}$, after k steps we have

$$\boldsymbol{Y}^{(k)} = \boldsymbol{P}^{(k-1)} \boldsymbol{P}^{(k-2)} \cdots \boldsymbol{P}^{(1)} \boldsymbol{P}^{(0)} \boldsymbol{Y}^{(0)},$$

where $\boldsymbol{Y}^{(k)}$ is a $n \times d$ matrix with $\boldsymbol{y}_i^{(k)}$ as its ith row.

Let $\boldsymbol{P}^{(k-1:0)} = \boldsymbol{P}^{(k-1)} \boldsymbol{P}^{(k-2)} \cdots \boldsymbol{P}^{(1)} \boldsymbol{P}^{(0)}$. The worst-case rate of convergence can be related to the second-largest singular value of $\boldsymbol{P}^{(k-1:0)}$ (Nedić et al., 2018). In particular, after k iterations we have

$$\sum_i \|\boldsymbol{y}_i^{(k)} - \overline{\boldsymbol{y}}\|_2^2 \leq \lambda_2(\boldsymbol{P}^{(k-1:0)}) \sum_i \|\boldsymbol{y}_i^{(0)} - \overline{\boldsymbol{y}}\|_2^2,$$

where $\lambda_2(\boldsymbol{P}^{(k-1:0)})$ denotes the second largest singular value of $\boldsymbol{P}^{(k-1:0)}$.

For the scheme proposed above, cycling deterministically through neighbors in the directed exponential graph, one can verify that after $k = \lfloor \log_2(n-1) \rfloor$ iterations, we have $\lambda_2(\boldsymbol{P}^{(k-1:0)}) = 0$, so all nodes exactly have the average. Intuitively, this happens because the directed exponential graph has excellent mixing properties: from any starting node in the network, one can get to any other node in at most $\log_2(n)$ hops. For $n = 32$ nodes, after 5 iterations averaging has converged using this strategy. In comparison, if one were to cycle through edges of the complete graph (where every node is connected to every other node), then for $n = 32$, after 5 consecutive iterations one would have still have $\lambda_2(\boldsymbol{P}^{(k-1:0)}) \approx 0.6$; i.e., nodes could be much further from the average (and hence, much less well-synchronized).

Similarly, one could consider designing the matrices $\boldsymbol{P}^{(k)}$ in a stochastic manner, where each node randomly samples one neighbor to send to at every iteration. If each node samples a destination uniformly from its set of neighbors in the directed exponential graph, then $\mathbb{E}\lambda_2(\boldsymbol{P}^{(k-1:0)}) \approx 0.4$, and if each node randomly selected a destination uniformly among all other nodes in the network (i.e., randomly from neighbors in the complete graph), then $\mathbb{E}\lambda_2(\boldsymbol{P}^{(k-1:0)}) \approx 0.2$. Thus, random schemes are still not as effective at quickly averaging as deterministically cycling through neighbors in the directed exponential graph. Moreover, with randomized schemes, we are no longer guaranteed that each node receives the same number of messages at every iteration, so the communication load will not be balanced as in the deterministic scheme.

The above discussion focused only on approximate distributed averaging, which is a key step within decentralized optimization. When averaging occurs less quickly, this also impacts optimization. Specifically, since nodes are less well-synchronized

Algorithm 2 Overlap Stochastic Gradient Push (SGP)

```
Require: Initialize $\tau \geq 0$, count_since_last $= 0$, $\gamma > 0$, $\boldsymbol{x}_i^{(0)} = \boldsymbol{z}_i^{(0)} \in \mathbb{R}^d$ and $w_i^{(0)} = 1$ for all nodes $i \in \{1, 2, \ldots, n\}$
 1: for $k = 0, 1, 2, \cdots, K$, at node $i$, do
 2:   Sample new mini-batch $\xi_i^{(k)} \sim \mathcal{D}_i$ from local distribution
 3:   Compute mini-batch gradient at $\boldsymbol{z}_i^{(k)}$: $\nabla \boldsymbol{F}_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)})$
 4:   $\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} - \gamma \nabla \boldsymbol{F}_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)})$
 5:   if $k$ mod $\tau = 0$ then
 6:     Non-blocking send $\left(p_{j,i}^{(k)} \boldsymbol{x}_i^{(k+\frac{1}{2})}, p_{j,i}^{(k)} w_i^{(k)}\right)$ to out-neighbors
 7:     $\boldsymbol{x}_i^{(k+1)} = p_{i,i} \boldsymbol{x}_i^{(k+1/2)}$
 8:     $w_i^{(k+1)} = p_{i,i} w_i^{(k)}$
 9:   else
10:     $\boldsymbol{x}_i^{(k+1)} = \boldsymbol{x}_i^{(k+1/2)}$
11:     $w_i^{(k+1)} = w_i^{(k)}$
12:   end if
13:   if count_since_last $= \tau$ then
14:     Block until $\left(p_{i,j}^{(k-\tau)} \boldsymbol{x}_j^{(k-\tau+\frac{1}{2})}, p_{i,j}^{(k-\tau)} w_j^{(k-\tau)}\right)$ is received for all in-neighbors $j$
15:     count_since_last $\leftarrow 0$
16:   else
17:     count_since_last $\leftarrow$ count_since_last $+1$
18:   end if
19:   if Receive buffer non-empty then
20:     for $\left(p_{i,j}^{(k')} \boldsymbol{x}_j^{(k'+\frac{1}{2})}, p_{i,j}^{(k')} w_j^{(k')}\right)$ in the receive buffer do
21:       $\boldsymbol{x}_i^{(k+1)} \leftarrow \boldsymbol{x}_i^{(k+1)} + p_{i,j}^{(k')} \boldsymbol{x}_j^{(k'+\frac{1}{2})}$
22:       $w_i^{(k+1)} \leftarrow p_{i,j}^{(k')} w_j^{(k')}$
23:     end for
24:   end if
25:   $\boldsymbol{z}_i^{(k+1)} = \boldsymbol{x}_i^{(k+1)} / w_i^{(k+1)}$
26: end for
```

(i.e., further from a consensus), each node will be evaluating its local mini-batch gradient at a different point in parameter space. Averaging these points (rather than updates based on mini-batch gradients evaluated at the same point) can be seen as injecting additional noise into the optimization process, and in our experience this can lead to worse performance in terms of train error.

B. Overlap SGP

Although SGP does not use network-wide collective communication primitives like ALLREDUCE, the implementation of Alg. 1 requires using blocking sends and receives; *i.e.*, nodes do not proceed to until they have received messages from all neighbors at that iteration. In this section we present the pseudocode of Overlap-SGP (OSGP) in Alg. 2 that overlaps gradient computation with communication to hide the communication cost. In line 25 in Algorithm 2, nodes compute the de-biased estimate of their model parameters. In lines 19 to 24, nodes aggregate all messages received in that iteration. Lines 13 to 18 ensure that the message delays are bounded, and that the nodes remain synchronized. In particular, Algorithm 2 is *synchronous* because of lines 13 to 18. If a node hasn't received a message from its in-neighbours in τ iterations, it will block and wait to received said messages. Note that if $\tau = 0$, vanilla SGP, then nodes block and wait to receive all incoming messages in each iteration. In lines 5 to 6, nodes send messages to their neighbours every τ iterations. Once again, note that if $\tau = 0$, vanilla SGP, then nodes send messages to their neighbours every iteration. In lines 2 to 4 the nodes take a stochastic gradient step. If $\tau = 1$ (1-overlap SGP), nodes transmit messages to their neighbours in every iteration, but don't wait to receive messages until the subsequent iteration.

We provide a lot of detail in Algorithm 2 to make it easier to implement the method; however, in essence, τ-overlap SGP is simply vanilla SGP with delayed communication. *i.e.*, where nodes only send a message to their neighbours every τ iterations, and can receive messages at any time in-between communication intervals.

[3] We define $(k \bmod 0) := 0$.

C. Implementation Details

In all of our experiments, we minimize the number of floating-point operations performed in each iteration, k, by using the mixing weights

$$p_{j,i}^{(k)} = 1/\left|\mathcal{N}_i^{\text{out}(k)}\right|$$

for all $i, j = 1, 2, \ldots, n$. In words, each node assigns mixing weights uniformly to all of its out-neighbors in each iteration. Recalling our convention that each node is an in- and out-neighbor of itself, it is easy to see that this choice of mixing-weight satisfies the column-stochasticity property. It may very well be that there is a different choice of mixing-weights that lead to better spectral properties of the gossip algorithm; however we leave this exploration for future work. We denote node i's uniform mixing weights at iteration k by $p_i^{(k)}$ — dropping the other subscript, which identifies the receiving node.

To leverage the highly efficient NVLink interconnect within each server, we treat each machine as one node in all of our experiments. In our implementation of SGP, each node computes a local mini-batch in parallel using all 8 GPUs via a local ALLREDUCE, which is efficiently implemented via the NVIDIA Collective Communications Library. Then inter-node averaging is accomplished using PUSHSUM either over Ethernet or InfiniBand. In the InfiniBand experiments, we leverage GPUDirect to directly send/receive messages between GPUs on different nodes and avoid transferring the model back to host memory. In the Ethernet experiments this is not possible, so the model is transferred to host memory after the local ALLREDUCE, and then PUSHSUM messages are sent over Ethernet.

To maximize the utility of the resources available on each server, each node (occupying a single server exclusively) runs two threads, a gossip thread and a computation thread. The computation thread executes the main logic used to train the local model on the GPUs available to the node, while the communication thread is used for inter-node network I/O. In particular, the communication thread is used to gossip messages between nodes. When using Ethernet-based communication, the nodes communicate their parameter tensors over CPUs. When using InifiniBand-based communication, the nodes communicate their parameter tensors using GPUDirect RDMA, thereby avoiding superfluous device to pinned-memory transfers of the model parameters.

Each node initializes its model on one of its GPUs, and initializes its scalar push-sum weight to 1. At the start of training, each node also allocates a *send*- and a *receive*- communication-buffer in pinned memory on the CPU (or equivalently on a GPU in the case of GPUDirect RDMA communication).

In each iteration, the communication thread waits for the send-buffer to be filled by the computation thread; transmits the message in the send-buffer to its out-neighbours; and then aggregates any newly-received messages into the receive-buffer.

In each iteration, the computation thread blocks to retrieve the aggregated messages (in the non-overlap case) in the receive-buffer; directly adds the received parameters to its own model parameters; and directly adds the received push-sum weights to its own push-sum weight. The computation thread then converts the model parameters to the *de-biased* estimate by dividing by the push-sum weight; executes a forward-backward pass of the *de-biased model* in order to compute a stochastic mini-batch gradient; converts the model parameters back to the *biased estimate* by multiplying by the push-sum weight; and applies the newly-computed stochastic gradients to the biased model. The updated model parameters are then multiplied by the mixing weight, $p_i^{(k)}$, and asynchronously copied back into the send-buffer for use by the communication thread. The push-sum weight is also multiplied by the same mixing weight and concatenated into the send-buffer.

In short, gossip is performed on the biased model parameters (push-sum numerators); stochastic gradients are computed using the de-biased model parameters; stochastic gradients are applied back to the biased model parameters; and then the biased-model and the push-sum weight are multiplied by the same uniform mixing-weight and copied back into the send-buffer.

C.1. Hyperparameters

For the ImageNet experiments, we follow the experimental protocol described in (Goyal et al., 2017). When we "apply the stochastic gradients" to the biased model parameters, we actually carry out an SGD step with nesterov momentum (see Alg. 3). For the $32, 64$, and 128 GPU experiments we use the same exact learning-rate, schedule, momentum, and weight decay as those suggested in (Goyal et al., 2017) for SGD. In particular, we use a reference learning-rate of 0.1 with respect to a 256 sample batch, and scale this linearly with the batch-size; we decay the learning-rate by a factor of 10 at epochs $30, 60, 80$; we use a Nesterov momentum parameter of 0.9, and we use weight decay 0.0001.

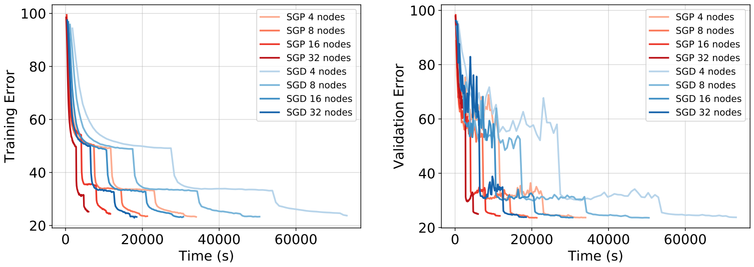
Algorithm 3 Stochastic Gradient Push with Momentum

```
Require: Initialize $\gamma > 0$, $m \in (0,1)$, $\boldsymbol{x}_i^{(0)} = \boldsymbol{z}_i^{(0)} \in \mathbb{R}^d$ and $w_i^{(0)} = 1$ for all nodes $i \in [n]$
for $k = 0, 1, 2, \cdots, K$, at node $i$, do
    Sample new mini-batch $\xi_i^{(k)} \sim \mathcal{D}_i$ from local distribution
    Compute mini-batch gradient at $\boldsymbol{z}_i^{(k)}$: $\nabla \boldsymbol{F}_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)})$
    $\boldsymbol{u}_i^{(k+1)} = m\boldsymbol{u}_i^{(k)} + \nabla \boldsymbol{F}_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)})$
    $\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} - \gamma(m\boldsymbol{u}_i^{(k+1)} + \nabla \boldsymbol{F}_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)}))$
    Send $\left(p_{j,i}^{(k)}\boldsymbol{x}_i^{(k+\frac{1}{2})}, p_{j,i}^{(k)}w_i^{(k)}\right)$ to out-neighbors;
    receive $\left(p_{i,j}^{(k)}\boldsymbol{x}_j^{(k+\frac{1}{2})}, p_{i,j}^{(k)}w_j^{(k)}\right)$ from in-neighbors
    $\boldsymbol{x}_i^{(k+1)} = \sum_{j \in \mathcal{N}_i^{\text{in}(k)}} p_{i,j}^{(k)}\boldsymbol{x}_j^{(k+\frac{1}{2})}$
    $w_i^{(k+1)} = \sum_{j \in \mathcal{N}_i^{\text{in}(k)}} p_{i,j}^{(k)}w_j^{(k)}$
    $\boldsymbol{z}_i^{(k+1)} = \boldsymbol{x}_i^{(k+1)}/w_i^{(k+1)}$
end for
```

For the machine translation experiment, we follow (Vaswani et al., 2017) and combine Stochastic Gradient Push with the Adam preconditioner. In particular, we make use of the FAIRSEQ code (Gehring et al., 2017), and train the transformer networks via SGP by replacing the built-in PyTorch parallel SGD model wrapper with our SGP model wrapper.

D. Additional Experiments

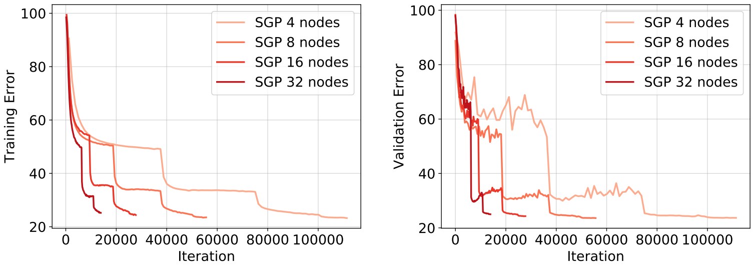
D.1. Additional Training Curves



(a) Train

(b) Validation

Figure D.1: Training on Ethernet 10Gbit/s



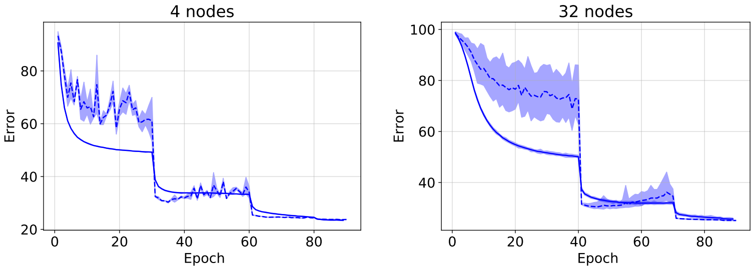
(a) Train

(b) Validation

Figure D.2: Training/Validation accuracy per iteration for SGP (Ethernet 10Gbit/s). Each time we double the number of node in the network, we half the total number of iterations.

The curves in Figure D.1 show the time-wise train- and validation-accuracies for the different runs performed on Ethernet 10Gbit/s. Figure D.2 reports the iteration-wise training and validation accuracy of SGP when using 10Gbits/s Ethernet.

D.2. Discrepancy across different nodes

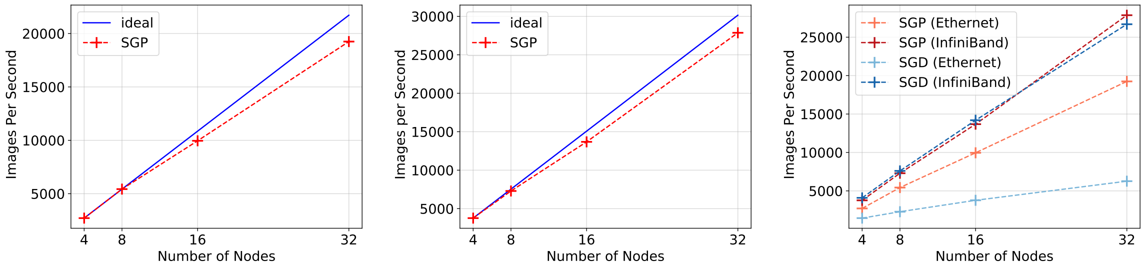


(a) Discrepancy on 4 nodes (b) Discrepancy on 32 nodes

Figure D.3: Resnet50, trained with SGP, training and validation errors for 4 and 32 nodes experiments. The solid and dashed lines in each figure show the mean training and validation error, respectively, over all nodes. The shaded region shows the maximum and minimum error attained at different nodes in the same experiment. Although there is non-trivial variability across nodes early in training, all nodes eventually converge to similar validation errors, achieving consensus in the sense that they represent the same function.

Here, we investigate the performance variability across nodes during training for SGP. In figure D.3, we report the minimum, maximum and mean error across the different nodes for training and validation. In an initial training phase, we observe that nodes have different validation errors; their local copies of the Resnet-50 model diverge. As we decrease the learning, the variability between the different nodes diminish and the nodes eventually converging to similar errors. This suggests that all models ultimately represent the same function, achieving consensus.

D.3. SGP Scaling Analysis



(a) ETH 10Gbit/s (b) InfiniBand 100Gbit/s (c) Scaling of SGP and SGP

Figure D.4: SGP throughput on Ethernet (a) and InfiniBand (b). SGP exhibits 88.6% scaling efficiency on Ethernet 10Gbit/s and 92.4% on InfiniBand. Comparison of SGD vs SGP throughput in Figure (c) shows that SGP exhibit better scaling and is more robust to high-latency interconnect.

Figure D.4 highlights SGP input images throughput as we scale up the number of cluster node on both Ethernet 10Gbit/s and Infiniband 100Gbit/s. SGP exhibits 88.6% scaling efficiency on Ethernet 10Gbit/s and 92.4% on InfiniBand and stay close to the ideal scaling in both cases. In addition Figure (c) shows that SGP exhibit better scaling as we increase the network size and is more robust to high-latency interconnect.

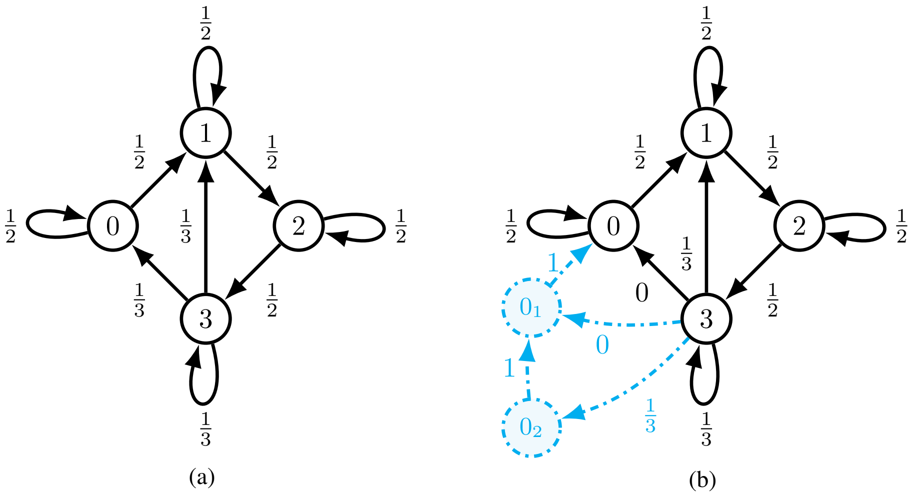


Figure E.1: (a) Example of a 4-node network, with mixing-weights drawn on edges. (b) Example of a 4-node network, augmented with virtual nodes and edges, with mixing-weights draw on edges. The virtual nodes/edges are used to model the fact that messages from node 3 to node 0 can experience a delay of at most 2 iterations. In this particular example, we model the fact that node 3 sends a message to node 0 with a delay of 2 iterations. All virtual nodes always forward all of their messages to their out-neighbor.

E. Proofs of Theoretical Guarantees

Our convergence rate analysis is divided into three main parts. In the first one (subsection E.1) we present upper bounds for three important expressions that appear in our computations. In subsection E.2 we focus on proving the important for our analysis Lemma 8 based on which we later build the proofs of our main Theorems. Finally in the third part (subsection E.3) we provide the proofs for Theorems 1 and 2.

Preliminary results. In our analysis, two preliminary results are extensively used. We state them here for future reference.

- Let $a, b \in \mathbb{R}$. Since $(a-b)^2 \geq 0$, it holds that

$$2ab \leq a^2 + b^2. \tag{4}$$

  Thus, $\|\boldsymbol{x}\| \, \|\boldsymbol{y}\| \leq (\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2)/2$.

- Let $r \in (0,1)$ then from the summation of geometric sequence and for any $K \leq \infty$ it holds that

$$\sum_{k=0}^{K} r^k \leq \sum_{k=0}^{\infty} r^k = \frac{1}{1-r}. \tag{5}$$

Modeling message delays. To model message delays we follow the procedure used in Assran & Rabbat (2018) (which we will reiterate here). In essence, we augment the communication topology (and the mixing matrices) with virtual nodes that store messages that were transmitted, but not yet received. Similar graph augmentations have been used in Charalambous et al. (2015) and Hadjicostis & Charalambous (2014).

We commence by presenting a brief example of the delay-model before formalizing the discussion. Figure E.1 (a) shows an example of a 4-node network at some arbitrary iteration k. Suppose each node communicates to each of its out-neighbors with uniform mixing weights. These mixing weights are labeled on the corresponding edges in Figure E.1 (a). Then, the mixing matrix $\boldsymbol{P}^{(k)} \in \mathbb{R}^{4\times 4}$ is given by

$$\boldsymbol{P}^{(k)} = \begin{pmatrix} 1/2 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/3 \end{pmatrix}.$$

Column indices correspond to sending nodes, and row indices correspond to receiving nodes. Recall that sending nodes choose the mixing weights (columns of $\boldsymbol{P}^{(k)}$) used to pre-weight outgoing messages. Note that the matrix $\boldsymbol{P}^{(k)}$ is column stochastic (all columns sum to 1) — the crucial requirement of our analysis. Thus at time $k+1$, we have the following parameter updates

$$\begin{aligned} \boldsymbol{x}_0^{(k+1)} &= \frac{1}{2}\boldsymbol{x}_0^{(k)} + \frac{1}{3}\boldsymbol{x}_3^{(k)} \\ \boldsymbol{x}_1^{(k+1)} &= \frac{1}{2}\boldsymbol{x}_0^{(k)} + \frac{1}{2}\boldsymbol{x}_1^{(k)} + \frac{1}{3}\boldsymbol{x}_3^{(k)} \\ \boldsymbol{x}_2^{(k+1)} &= \frac{1}{2}\boldsymbol{x}_1^{(k)} + \frac{1}{2}\boldsymbol{x}_2^{(k)} \\ \boldsymbol{x}_3^{(k+1)} &= \frac{1}{2}\boldsymbol{x}_2^{(k)} + \frac{1}{3}\boldsymbol{x}_3^{(k)}. \end{aligned}$$

In particular, each node updates its variables with the most recent information from its in-neighbours. Similar equations can be written for the push-sum weights w.

Now suppose that node 3 sends messages to its neighbors, nodes 0 and 1, at iteration k, but the message to node 0 doesn't arrive until iteration $k+2$. To model this delay, we augment the graph topology with virtual nodes 0_1, 0_2 (cf. Figure E.1 (b)). The virtual nodes are initialized with parameters $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ and push-sum weight $w^{(0)} = 0$. Given this model, node 3 can send its pre-weighted message to virtual node 0_2 (instead of node 0) at iteration k, while the rest of communication proceeds business as usual. At the subsequent iteration, $k+1$, node 0_2 forwards this message to node 0_1. Subsequently, at iteration, $k+2$, node 0_1 forwards this message to node 0, thereby modeling a 2-iteration message delay. The corresponding mixing matrix at iteration k is given by

$$\boldsymbol{P}^{(k)} = \left(\begin{array}{c|cccc|cc} & & & & & 0_1 & 0_2 \\ & 1/2 & 0 & 0 & 1/2 & 1 & 0 \\ & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ & 0 & 1/2 & 1/3 & 0 & 0 & 0 \\ & 0 & 0 & 1/3 & 1/2 & 0 & 0 \\ \hline 0_1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0_2 & 0 & 0 & 1/3 & 0 & 0 & 0 \end{array}\right).$$

Note that we have added two extra rows and columns corresponding to the virtual nodes 0_1 and 0_2. As intended, node 2 sends a message to node 0_2 (instead of node 0) at iteration k. Node 0_2 always forwards any and all information it receives to node 0_1, and node 0_1 always forwards any and all information it receives to node 0. Since all virtual nodes are initialized with parameters $\boldsymbol{x}^{(k)} = \boldsymbol{0}$ and push-sum weight $w^{(0)} = 0$, they do not have any impact on the final consensus value. The sole purpose of the virtual nodes is to store messages that are in-transit (transmitted but not yet received).

If the message delays at every node are upper-bounded by τ, then we can generalize this procedure, and add τ virtual nodes for every (non-virtual) node in the network. Thus, the augmented graph has $n(\tau+1)$ nodes in total. The corresponding augmented mixing matrix, $\boldsymbol{P}^{(k)} \in \mathbb{R}^{n(\tau+1)\times n(\tau+1)}$, in block matrix form is written as

$$\boldsymbol{P}^{(k)} = \begin{array}{c} \begin{array}{ccccc} & (0_1, 1_1 \ldots) & (0_2, 1_2, \ldots) & & (0_\tau, 1_\tau, \ldots) \end{array} \\ \begin{pmatrix} \widetilde{\boldsymbol{P}}_0^{(k)} & \boldsymbol{I} & \mathbf{0} & \ldots & \mathbf{0} \\ \widetilde{\boldsymbol{P}}_1^{(k)} & \mathbf{0} & \boldsymbol{I} & & \vdots \\ \vdots & \vdots & & \ddots & \mathbf{0} \\ \widetilde{\boldsymbol{P}}_{\tau-1}^{(k)} & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{I} \\ \widetilde{\boldsymbol{P}}_{\tau}^{(k)} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}$$

where each block is of size $n \times n$. In particular, if node i sends a message to node j with weight $p_{j,i}^{(k)}$ at iteration k, and that message is received with delay r (*i.e.*, received at iteration $k + r$), then

$$[\boldsymbol{P}_r^{(k)}]_{j,i} = p_{j,i}^{(k)},$$

otherwise

$$[\boldsymbol{P}_r^{(k)}]_{j,i} = 0.$$

The off-diagonal of block identity matrices $\boldsymbol{I}$ denote the fact that the virtual nodes always forward all of their messages to the next node in the delay daisy-chain. It is straightforward to verify that these augmented mixing matrices are still column stochastic at all iterations k. We refer the curious reader to Assran & Rabbat (2018); Charalambous et al. (2015); Hadjicostis & Charalambous (2014) for a deeper discussion of the augmented delay model.

Matrix Representation. In Algorithm 1, SGP was presented from node i's perspective (for all $i \in [n]$). However, we can actually write the SGP update at each iteration from a global viewpoint. To see this, first define the following matrices, for all $r = 1, 2, \ldots \tau$,

$$\mathbf{X}_r^{(k)} = \left[\boldsymbol{x}_{1_r}^{(k)}, \boldsymbol{x}_{2_r}^{(k)}, \ldots, \boldsymbol{x}_{n_r}^{(k)}\right] \in \mathbb{R}^{d\times n}.$$

The matrix $\mathbf{X}_r^{(k)}$ denotes a concatenation of all the delay-r nodes' parameters at iteration k. For the purpose of notational consistency, we let the matrix $\mathbf{X}_0^{(k)}$ denote the concatenation of all the non-virtual nodes' parameters. We generalize this notation to other variables as well. In block-matrix form, we can define the augmented parameter matrix

$$\mathbf{X}^{(k)} = [\mathbf{X}_0^{(k)}, \mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_\tau^{(k)}] \in \mathbb{R}^{d\times n(\tau+1)},$$

which denotes a concatenation of *all* (virtual and non-virtual) nodes' parameters at iteration k. Recall that the we initialize all virtual nodes with parameters $\boldsymbol{x}^{(k)} = \mathbf{0}$ and push-sum weight $w^{(0)} = 0$. Additionally, since the virtual nodes are only used to model delays, and do not compute any gradient updates, we use the convention that $\boldsymbol{z}^{(k)} = 0$, $\xi^{(k)} = 0$, and $\nabla F(\boldsymbol{z}^{(k)}; \xi^{(k)}) = \mathbf{0}$ for all virtual nodes at all times k. Therefore, we define the augmented de-biased parameter matrix and stochastic-seed matrix as follows

$$\mathbf{Z}^{(k)} = [\mathbf{Z}_0^{(k)}, \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{d\times n(\tau+1)}; \quad \xi^{(k)} = [\boldsymbol{\xi}_0^{(k)}, \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{n(\tau+1)}.$$

Similarly, we define the augmented stochastic-gradient matrix as

$$\nabla \boldsymbol{F}(\mathbf{Z}^{(k)}; \boldsymbol{\xi}^{(k)}) = [\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)}; \boldsymbol{\xi}_0^{(k)}), \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{d\times n(\tau+1)},$$

where the block matrix $\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)}; \boldsymbol{\xi}_0^{(k)})$ denotes the concatenation of all non-virtual nodes' stochastic gradients at iteration k. Precisely

$$\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)}, \boldsymbol{\xi}_0^{(k)}) = \left[\nabla F_1(\boldsymbol{z}_1^{(k)}; \xi_1^{(k)}), \nabla F_2(\boldsymbol{z}_2^{(k)}; \xi_2^{(k)}), \ldots, \nabla F_n(\boldsymbol{z}_n^{(k)}; \xi_n^{(k)})\right] \in \mathbb{R}^{d\times n}.$$

We also define the augmented expected gradient matrix (with respect to local node data distributions) as

$$\nabla \boldsymbol{F}(\mathbf{Z}^{(k)}) = [\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)}), \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{d\times n(\tau+1)},$$

where the block matrix $\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)})$ denotes the concatenation of all non-virtual nodes' expected stochastic gradients at iteration k. Precisely

$$\nabla \boldsymbol{F}_0(\mathbf{Z}_0^{(k)}) = \Big[\mathbb{E}_{\xi_1^{(k)} \sim \mathcal{D}_1}[\nabla F_1(\boldsymbol{z}_1^{(k)}; \xi_1^{(k)})], \mathbb{E}_{\xi_2^{(k)} \sim \mathcal{D}_2}[\nabla F_2(\boldsymbol{z}_2^{(k)}; \xi_2^{(k)})], \ldots, \mathbb{E}_{\xi_n^{(k)} \sim \mathcal{D}_n}[\nabla F_n(\boldsymbol{z}_n^{(k)}; \xi_n^{(k)})]\Big] \in \mathbb{R}^{d \times n}.$$

For notational convenience, we simply write $\nabla f_i(\boldsymbol{z}_i^{(k)}) := \mathbb{E}_{\xi_i^{(k)} \sim \mathcal{D}_i}[\nabla F_i(\boldsymbol{z}_i^{(k)}; \xi_i^{(k)})]$. Using the above matrices, the 6^{th} step of SGP in Algorithm 1 (lines 19 to 24 in OSGP Algorithm 2) can be expressed from a global perspective as follows

$$\mathbf{X}^{(k+1)} = \left(\mathbf{X}^{(k)} - \gamma \nabla \boldsymbol{F}(\mathbf{Z}^{(k)}, \boldsymbol{\xi}^{(k)})\right) [\boldsymbol{P}^{(k)}]^T, \tag{6}$$

where $[\boldsymbol{P}^{(k)}]^T \in \mathbb{R}^{n(\tau+1) \times n(\tau+1)}$ is the transpose of the augmented mixing matrix.

Lastly, let $\overline{\boldsymbol{n}} := n(\tau + 1)$, and let $\overline{\boldsymbol{x}}^{(k)} = (1/n)\mathbf{X}^{(k)}\mathbf{1}_{\overline{\mathrm{n}}}$ denote the average of all nodes' parameters at iteration k. Note that this definition incorporates parameters that are in-transit.

Bound for the mixing matrices. Next we state a known result from the control literature studying gossip-based optimization which allows us to bound the distance between the de-biased parameters at each node and the node-wise average.

Recall that we have assumed that the sequence of communication topologies is B-strongly connected. A directed graph is called *strongly connected* if every pair of vertices is connected with a directed path (*i.e.*, following the direction of edges). A sequence of directed graphs is called B-strongly connected if the graph with edge set $\bigcup_{k=lB}^{(l+1)B-1} E^{(k)}$ is strongly connected, for every $l \geq 0$. Recall that we have also assumed that the upper bound on the message delays is τ iterations. In particular, we assume all messages reach their destination within τ-iterations from transmission. *i.e.*, a message in-transit does not get dropped when the communication topology changes.

If the maximum message delay is τ, and all non-zero mixing weights are at least ϵ large, and the diameter of the graph with edge set $\bigcup_{k=lB}^{(l+1)B-1} E^{(k)}$ has diameter at most Δ, then the product

$$\boldsymbol{A}^{(k)} := \boldsymbol{P}^{(k+(\tau+1)\Delta B-1)} \cdots \boldsymbol{P}^{(k+1)} \boldsymbol{P}^{(k)}$$

has no non-zero entries in the first n-rows (corresponding to non-virtual agents). Moreover, every entry in the first n-rows of $\boldsymbol{A}^{(k)}$ is at least $\epsilon^{(\tau+1)\Delta B}$.

If we further assume that all nodes have at most D out-neighbors in any iteration, and that all nodes always assign mixing weights uniformly, then $\epsilon = D^{-1}$, and every entry in the first n-rows of $\boldsymbol{A}^{(k)}$ is at least $D^{-(\tau+1)\Delta B}$.

Lemma 3. *Suppose that Assumption 3 (mixing connectivity) holds. Let $\lambda = 1 - nD^{-(\tau+1)\Delta B}$ and let $q = \lambda^{1/((\tau+1)\Delta B+1)}$. Then there exists a constant*

$$C < \frac{2\sqrt{d} D^{(\tau+1)\Delta B}}{\lambda^{\frac{(\tau+1)\Delta B+2}{(\tau+1)\Delta B+1}}},$$

where d is the dimension of $\overline{\boldsymbol{x}}^{(k)}$, ${z_i}^{(k)}$, and ${x_i}^{(0)}$, such that, for all $i = 1, 2, \ldots, n$ (non-virtual nodes) and $k \geq 0$,

$$\left\|\overline{\boldsymbol{x}}^{(k)} - {z_i}^{(k)}\right\|_2 \leq Cq^k \left\|{x_i}^{(0)}\right\|_2 + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|\nabla F_i(z_i^{(s)}; \xi_i^{(s)})\right\|_2.$$

This particular lemma follows after a small adaptation to Theorem 1 in Assran & Rabbat (2018) and its proof is based on Wolfowitz (1963). Similar bounds appear in a variety of other papers, including Nedić & Olshevsky (2016).

E.1. Important Upper Bounds

Lemma 4 (Bound of stochastic gradient)**.** *We have the following inequality under Assumptions 1 and 2:*

$$\mathbb{E}\left\|\nabla f_i(\boldsymbol{z}_i^{(k)})\right\|^2 \leq 3L^2 \mathbb{E}\left\|\boldsymbol{z}_i^{(k)} - \overline{\boldsymbol{x}}^{(k)}\right\|^2 + 3\zeta^2 + 3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2$$

Proof.

$$\begin{aligned}
\mathbb{E}\left\|\nabla f_i(\boldsymbol{z}_i^{(k)})\right\|^2 &\leq 3\mathbb{E}\left\|\nabla f_i(\boldsymbol{z}_i^{(k)}) - \nabla f_i(\overline{\boldsymbol{x}}^{(k)})\right\|^2 + 3\mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{x}}^{(k)}) - \nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 + 3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\overset{\text{L-smooth}}{\leq} 3L^2\mathbb{E}\left\|\boldsymbol{z}_i^{(k)} - \overline{\boldsymbol{x}}^{(k)}\right\|^2 + 3\mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{x}}^{(k)}) - \nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 + 3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\overset{\text{Bounded Variance}}{\leq} 3L^2\mathbb{E}\left\|\boldsymbol{z}_i^{(k)} - \overline{\boldsymbol{x}}^{(k)}\right\|^2 + 3\zeta^2 + 3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2
\end{aligned}$$

□

Lemma 5. *Let Assumptions 1-3 hold. Then,*

$$\begin{aligned}
Q_i^{(k)} = \mathbb{E}\left\|\overline{\boldsymbol{x}}^{(k)} - z_i{}^{(k)}\right\|^2 &\leq \left(\gamma^2\frac{4C^2}{(1-q)^2} + \gamma\frac{q^kC^2}{1-q}\right)\sigma^2 + \left(\gamma^2\frac{12C^2}{(1-q)^2} + \gamma\frac{q^k3C^2}{1-q}\right)\zeta^2 \\
&+ \left(\gamma^2\frac{12L^2C^2}{1-q} + \gamma q^k 3L^2C^2\right)\sum_{j=0}^{k} q^{k-j}Q_i^{(j)} \\
&+ \left(\gamma^2\frac{12C^2}{1-q} + \gamma q^k 3C^2\right)\sum_{j=0}^{k} q^{k-j}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(j)})\right\|^2 \\
&+ \left(q^{2k}C^2 + \gamma q^k\frac{2C^2}{1-q}\right)\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2. \qquad (7)
\end{aligned}$$

Proof.

$$\begin{aligned}
Q_i^{(k)} &= \mathbb{E}\left\|\overline{\boldsymbol{x}}^{(k)} - \boldsymbol{z}_i{}^{(k)}\right\|^2 \\
&\overset{Lemma\ 3}{\leq} \mathbb{E}\left(Cq^k\left\|\boldsymbol{x}_i{}^{(0)}\right\| + \gamma C\sum_{s=0}^{k} q^{k-s}\left\|\nabla F_i(\boldsymbol{z}_i^{(s)};\xi_i^{(s)})\right\|\right)^2 \\
&= \mathbb{E}\left(Cq^k\left\|\boldsymbol{x}_i{}^{(0)}\right\| + \gamma C\sum_{s=0}^{k} q^{k-s}\left\|\nabla F_i(\boldsymbol{z}_i^{(s)};\xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)}) + \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|\right)^2 \\
&\leq \mathbb{E}\left(\underbrace{Cq^k\left\|\boldsymbol{x}_i{}^{(0)}\right\|}_{a} + \underbrace{\gamma C\sum_{s=0}^{k} q^{k-s}\left\|\nabla F_i(\boldsymbol{z}_i^{(s)};\xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}^{(s)})\right\|}_{b} + \underbrace{\gamma C\sum_{s=0}^{k} q^{k-s}\left\|\nabla f_i(\boldsymbol{z}_i^{(s)})\right\|}_{c}\right)^2
\end{aligned}$$

$$(8)$$

Thus, using the above expressions of a, b and c we have that $Q_i^{(k)} \leq \mathbb{E}(a^2 + b^2 + c^2 + 2ab + 2bc + 2ac)$. Let us now obtain bounds for all of these quantities:

$$a^2 = C^2\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2 q^{2k}$$

$$\begin{aligned}
b^2 &= \gamma^2C^2\sum_{j=0}^{k} q^{2(k-j)}\left\|\nabla F_i(\boldsymbol{z}_i^{(j)};\xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \\
&\quad \underbrace{+\,2\gamma^2C^2\sum_{j=0}^{k}\sum_{s=j+1}^{k} q^{2k-j-s}\left\|\nabla F_i(\boldsymbol{z}_i^{(j)};\xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|\left\|\nabla F_i(\boldsymbol{z}_i^{(s)};\xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|}_{b_1}
\end{aligned}$$

$$c^2 = \gamma^2C^2\sum_{j=0}^{k} q^{2(k-j)}\left\|\nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 + \underbrace{2\gamma^2C^2\sum_{j=0}^{k}\sum_{s=j+1}^{k} q^{2k-j-s}\left\|\nabla f_i(\boldsymbol{z}_i^{(j)})\right\|\left\|\nabla f_i(\boldsymbol{z}_i^{(s)})\right\|}_{c_1}$$

$$
\begin{aligned}
2ab &= 2\gamma C^2 q^k \left\|{\boldsymbol{x}_i}^{(0)}\right\| \sum_{s=0}^{k} q^{k-s} \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\| \\
2ac &= 2\gamma C^2 q^k \left\|{\boldsymbol{x}_i}^{(0)}\right\| \sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i(\boldsymbol{z}_i^{(s)})\right\| \\
2bc &= 2\gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-j-s} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\| \left\|\nabla f_i(\boldsymbol{z}_i^{(s)})\right\|.
\end{aligned}
$$

The expression b_1 is bounded as follows:

$$
\begin{aligned}
b_1 &= \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=j+1}^{k} q^{2k-j-s} 2 \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\| \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\| \\
&\overset{(4)}{\le} \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=j+1}^{k} q^{2k-s-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \\
&\quad + \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=j+1}^{k} q^{2k-s-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|^2 \\
&\le \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-s-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \\
&\quad + \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-s-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|^2 \\
&= \gamma^2 C^2 \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \sum_{s=0}^{k} q^{k-s} \\
&\quad + \gamma^2 C^2 \sum_{s=0}^{k} q^{k-s} \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|^2 \sum_{j=0}^{k} q^{k-j} \\
&\overset{(5)}{\le} \frac{1}{1-q} \gamma^2 C^2 \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \\
&\quad + \frac{1}{1-q} \gamma^2 C^2 \sum_{s=0}^{k} q^{k-s} \left\|\nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)})\right\|^2 \\
&= \frac{2}{1-q} \gamma^2 C^2 \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2. \qquad (9)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
b^2 &= \gamma^2 C^2 \sum_{j=0}^{k} q^{2(k-j)} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 + b_1 \\
&\le \frac{\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 + b_1 \\
&\overset{(9)}{\le} \frac{3\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)})\right\|^2 \qquad (10)
\end{aligned}
$$

where in the first inequality above we use the fact that for $q \in (0, 1)$, we have $q^k < \frac{1}{1-q}, \forall k > 0$.

By identical construction we have

$$c^2 \leq \frac{3\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\| \nabla f_i(\boldsymbol{z}_i^{(j)}) \right\|^2 .$$

Now let us bound the products $2ab$, $2ac$ and $2bc$.

$$\begin{aligned} 2ab &= \gamma C^2 q^k \sum_{s=0}^{k} q^{k-s} 2 \left\| \boldsymbol{x}_i{}^{(0)} \right\| \left\| \nabla F_i(\boldsymbol{z}_i^{(s)}; \xi_i^{(s)}) - \nabla f_i(\boldsymbol{z}_i^{(s)}) \right\| \\ &\overset{(4)}{\leq} \gamma C^2 q^k \sum_{j=0}^{k} q^{k-j} \left\| \nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)}) \right\|^2 + \gamma C^2 q^k \sum_{j=0}^{k} q^{k-j} \left\| \boldsymbol{x}_i{}^{(0)} \right\|^2 \\ &\overset{(5)}{\leq} \gamma C^2 q^k \sum_{j=0}^{k} q^{k-j} \left\| \nabla F_i(\boldsymbol{z}_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(\boldsymbol{z}_i^{(j)}) \right\|^2 + \frac{\gamma C^2 \left\| x_i{}^{(0)} \right\|^2}{1-q} q^k \end{aligned} \tag{11}$$

By similar procedure,

$$2ac \leq \quad \gamma C^2 q^k \textstyle\sum_{s=0}^{k} q^{k-s} \left\| \nabla f_i(z_i^{(s)}) \right\|^2 + \frac{\gamma C^2 \left\| x_i{}^{(0)} \right\|^2}{1-q} q^k \tag{12}$$

Finally,

$$\begin{aligned} 2bc &= \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-j-s} 2 \left\| \nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)}) \right\| \left\| \nabla f_i(z_i^{(s)}) \right\| \\ &\overset{(4)}{\leq} \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-j-s} \left\| \nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)}) \right\|^2 + \gamma^2 C^2 \sum_{j=0}^{k} \sum_{s=0}^{k} q^{2k-j-s} \left\| \nabla f_i(z_i^{(s)}) \right\|^2 , \\ &= \gamma^2 C^2 \sum_{j=0}^{k} q^{k-j} \left\| \nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)}) \right\|^2 \sum_{s=0}^{k} q^{k-s} + \gamma^2 C^2 \sum_{s=0}^{k} q^{k-s} \left\| \nabla f_i(z_i^{(s)}) \right\|^2 \sum_{j=0}^{k} q^{k-j} , \\ &\overset{(5)}{\leq} \frac{\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\| \nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)}) \right\|^2 + \frac{\gamma^2 C^2}{1-q} \sum_{s=0}^{k} q^{k-s} \left\| \nabla f_i(z_i^{(s)}) \right\|^2 \end{aligned} \tag{13}$$

By combining all of the above bounds together we obtain:

$$\begin{aligned} Q_i^{(k)} &\leq \mathbb{E}(a^2 + b^2 + c^2 + 2ab + 2bc + 2ac) \\ &\leq \mathbb{E} \frac{4\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\| \nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)}) \right\|^2 \\ &+ \mathbb{E} \frac{4\gamma^2 C^2}{1-q} \sum_{j=0}^{k} q^{k-j} \left\| \nabla f_i(z_i^{(j)}) \right\|^2 \\ &+ C^2 \left\| x_i{}^{(0)} \right\|^2 q^{2k} \\ &+ \frac{2\gamma C^2 \left\| x_i{}^{(0)} \right\|^2}{1-q} q^k \\ &+ \mathbb{E} \gamma C^2 q^k \sum_{j=0}^{k} q^{k-j} \left\| \nabla f_i(z_i^{(j)}) \right\|^2 \end{aligned}$$

$$+ \quad \mathbb{E}\gamma C^2 q^k \sum_{j=0}^{k} q^{k-j} \left\|\nabla F_i(z_i^{(j)}; \xi_i^{(j)}) - \nabla f_i(z_i^{(j)})\right\|^2 . \tag{14}$$

After grouping terms together and using the upper bound of Lemma 4, we obtain

$$\begin{aligned}
Q_i^{(k)} \quad &\leq \quad \left(\gamma^2 \frac{4C^2}{(1-q)^2} + \gamma \frac{q^k C^2}{1-q}\right)\sigma^2 + \left(q^{2k}C^2 + \gamma q^k \frac{2C^2}{1-q}\right)\left\|{x_i}^{(0)}\right\|^2 . \\
&+ \quad \left(\gamma^2 \frac{4C^2}{1-q} + \gamma q^k C^2\right)\sum_{j=0}^{k} q^{k-j}\mathbb{E}\left\|\nabla f_i(z_i^{(j)})\right\|^2 \\
&\overset{Lemma\ 4}{\leq} \quad \left(\gamma^2 \frac{4C^2}{(1-q)^2} + \gamma \frac{q^k C^2}{1-q}\right)\sigma^2 + \left(q^{2k}C^2 + \gamma q^k \frac{2C^2}{1-q}\right)\left\|{x_i}^{(0)}\right\|^2 \\
&+ \quad \left(\gamma^2 \frac{12C^2}{(1-q)^2} + \frac{\gamma q^k 3C^2}{1-q}\right)\zeta^2 \\
&+ \quad \left(\gamma^2 \frac{12L^2C^2}{1-q} + \gamma q^k 3L^2C^2\right)\sum_{j=0}^{k} q^{k-j} Q_i^{(j)} \\
&+ \quad \left(\gamma^2 \frac{12C^2}{1-q} + \gamma q^k 3C^2\right)\sum_{j=0}^{k} q^{k-j}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(j)})\right\|^2
\end{aligned} \tag{15}$$

This completes the proof. □

Having found a bound for the quantity $Q_i^{(k)}$, let us now present a lemma for bounding the quantity $\sum_{k=0}^{K-1} M^{(k)}$ where $K > 1$ is a constant and $M^{(k)}$ is the average $Q_i^{(k)}$across all (non-virtual) nodes $i \in [n]$. That is, $M^{(k)} = \frac{1}{n}\sum_{i=1}^{n} Q_i^{(k)}$.

Lemma 6. *Let Assumptions 1-3 hold and let us define* $D_2 = 1 - \dfrac{\gamma^2 12 L^2 C^2}{(1-q)^2} - \dfrac{\gamma 3 L^2 C^2}{(1-q)^2}$ *. Then,*

$$\begin{aligned}
\sum_{k=0}^{K-1} M^{(k)} \quad \leq \quad & \left(\gamma^2 \frac{4C^2}{(1-q)^2 D_2}\right)\sigma^2 K + \left(\gamma \frac{C^2}{(1-q)^2 D_2}\right)\sigma^2 \\
& + \left(\gamma^2 \frac{12C^2}{(1-q)^2 D_2}\right)\zeta^2 K + \left(\frac{\gamma 3C^2}{(1-q)^2 D_2}\right)\zeta^2 \\
& + \left(\frac{C^2}{(1-q)^2 D_2} + \gamma \frac{2C^2}{(1-q)^2 D_2}\right)\frac{\sum_{i=1}^{n}\left\|{x_i}^{(0)}\right\|^2}{n} \\
& + \left(\gamma^2 \frac{12C^2}{(1-q)^2 D_2} + \gamma \frac{3C^2}{(1-q)^2 D_2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2
\end{aligned} \tag{16}$$

Proof. Using the bound for $Q_i^{(k)}$ let us first bound its average across all nodes $M^{(k)}$

$$\begin{aligned}
M^{(k)} &= \frac{1}{n}\sum_{i=1}^{n} Q_i^{(k)} \\
&\overset{Lemma~5}{\leq} \left(\gamma^2 \frac{4C^2}{(1-q)^2} + \gamma \frac{q^k C^2}{1-q}\right)\sigma^2 + \left(\gamma^2 \frac{12C^2}{(1-q)^2} + \frac{\gamma q^k 3C^2}{1-q}\right)\zeta^2 \\
&+ \left(\gamma^2 \frac{12C^2}{1-q} + \gamma q^k 3C^2\right)\sum_{j=0}^{k} q^{k-j}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(j)})\right\|^2 \\
&+ \left(\gamma^2 \frac{12L^2C^2}{1-q} + \gamma q^k 3L^2C^2\right)\sum_{j=0}^{k} q^{k-j} M^{(j)} \\
&+ \left(q^{2k}C^2 + \gamma q^k \frac{2C^2}{1-q}\right)\frac{\sum_{i=1}^{n}\left\|{x_i}^{(0)}\right\|^2}{n}. \qquad (17)
\end{aligned}$$

At this point note that for any $\lambda \in (0,1)$, non-negative integer $K \in \mathbb{N}$, and non-negative sequence $\{\beta^{(j)}\}_{j=0}^{k}$, it holds that

$$\begin{aligned}
\sum_{k=0}^{K}\sum_{j=0}^{k}\lambda^{k-j}\beta^{(j)} &= \beta^{(0)}\left(\lambda^K + \lambda^{K-1} + \cdots + \lambda^0\right) + \beta^{(1)}\left(\lambda^{K-1} + \lambda^{K-2} + \cdots + \lambda^0\right) + \cdots + \beta^{(K)}\left(\lambda^0\right) \\
&\leq \frac{1}{1-\lambda}\sum_{j=0}^{K}\beta^{(j)}. \qquad (18)
\end{aligned}$$

Similarly,

$$\sum_{k=0}^{K}\lambda^k\sum_{j=0}^{k}\lambda^{k-j}\beta^{(j)} = \sum_{k=0}^{K}\sum_{j=0}^{k}\lambda^{2k-j}\beta^{(j)} \leq \sum_{k=0}^{K}\sum_{j=0}^{k}\lambda^{2(k-j)}\beta^{(j)} \overset{(18)}{\leq} \frac{1}{1-\lambda^2}\sum_{j=0}^{K}\beta^{(j)} \qquad (19)$$

Now by summing from $k=0$ to $K-1$ and using the bounds of (18) and (19) we obtain:

$$\begin{aligned}
\sum_{k=0}^{K-1} M^{(k)} \leq &\left(\gamma^2 \frac{4C^2}{(1-q)^2}\right)\sigma^2 K + \left(\gamma \frac{C^2}{(1-q)^2}\right)\sigma^2 \\
&+ \left(\gamma^2 \frac{12C^2}{(1-q)^2}\right)\zeta^2 K + \left(\frac{\gamma 3C^2}{1-q}\right)\zeta^2 \\
&+ \left(\frac{C^2}{1-q^2} + \gamma \frac{2C^2}{(1-q)^2}\right)\frac{\sum_{i=1}^{n}\left\|{x_i}^{(0)}\right\|^2}{n} \\
&+ \left(\gamma^2 \frac{12C^2}{(1-q)^2} + \gamma \frac{3C^2}{1-q^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&+ \left(\gamma^2 \frac{12L^2C^2}{(1-q)^2} + \gamma \frac{3L^2C^2}{1-q^2}\right)\sum_{k=0}^{K-1} M^{(k)}.
\end{aligned}$$

By rearranging:

$$\begin{aligned}\left(1-\gamma^2\frac{12L^2C^2}{(1-q)^2}-\gamma\frac{3L^2C^2}{1-q^2}\right)\sum_{k=0}^{K-1}M^{(k)} \le{}& \left(\gamma^2\frac{4C^2}{(1-q)^2}\right)\sigma^2K+\left(\gamma\frac{C^2}{(1-q)^2}\right)\sigma^2\\ &+\left(\gamma^2\frac{12C^2}{(1-q)^2}\right)\zeta^2K+\left(\frac{\gamma3C^2}{(1-q)^2}\right)\zeta^2\\ &+\left(\frac{C^2}{1-q^2}+\gamma\frac{2C^2}{(1-q)^2}\right)\frac{\sum_{i=1}^n\left\|{x_i}^{(0)}\right\|^2}{n}\\ &+\left(\gamma^2\frac{12C^2}{(1-q)^2}+\gamma\frac{3C^2}{1-q^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2\end{aligned}$$

Note that since $q\in(0,1)$ it holds that $\frac{1}{1-q^2}\le\frac{1}{(1-q)^2}$.[4] Thus,

$$\begin{aligned}\left(1-\gamma^2\frac{12L^2C^2}{(1-q)^2}-\gamma\frac{3L^2C^2}{(1-q)^2}\right)\sum_{k=0}^{K-1}M^{(k)} \le{}& \left(\gamma^2\frac{4C^2}{(1-q)^2}\right)\sigma^2K+\left(\gamma\frac{C^2}{(1-q)^2}\right)\sigma^2\\ &+\left(\gamma^2\frac{12C^2}{(1-q)^2}\right)\zeta^2K+\left(\frac{\gamma3C^2}{(1-q)^2}\right)\zeta^2\\ &+\left(\frac{C^2}{(1-q)^2}+\gamma\frac{2C^2}{(1-q)^2}\right)\frac{\sum_{i=1}^n\left\|{x_i}^{(0)}\right\|^2}{n}\\ &+\left(\gamma^2\frac{12C^2}{(1-q)^2}+\gamma\frac{3C^2}{(1-q)^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2\end{aligned}$$

Dividing both sides with $D_2=1-\dfrac{\gamma^2 12L^2C^2}{(1-q)^2}-\dfrac{\gamma3L^2C^2}{(1-q)^2}$ completes the proof. □

E.2. Towards the proof of the main Theorems

The goal of this section is the presentation of Lemma 8. It is the main lemma of our convergence analysis and based on which we build the proofs of Theorems 1 and 2.

Let us first state a preliminary lemma that simplifies some of the expressions that involve expectations with respect to the random variable $\xi_i^{(t)}$.

Lemma 7. *Under the definition of our problem and the Assumptions 1-3 we have that:*

(i)

$$\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n\nabla F_i(z_i^{(k)};\xi_i^{(k)})}{n}\right\|^2=\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})}{n}\right\|^2+\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n\nabla f_i(z_i^{(k)})}{n}\right\|^2$$

(ii)

$$\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n\left[\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})\right]}{n}\right\|^2\le\frac{\sigma^2}{n}$$

[4] This step is used to simplified the expressions involve the parameter q. One can still obtain similar results by keeping the expression $\frac{1}{1-q^2}$ in the definition of D_2.

Proof.

$$\begin{aligned}
\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla F_i(z_i^{(k)};\xi_i^{(k)})}{n}\right\|^2 &= \mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})}{n}+\frac{\sum_{i=1}^n \nabla f_i(z_i^{(k)})}{n}\right\|^2 \\
&= \mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})}{n}\right\|^2 \\
&\quad +\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla f_i(z_i^{(k)})}{n}\right\|^2 \\
&\quad +2\left\langle\frac{\sum_{i=1}^n \mathbb{E}_{\xi_i^{(k)}}\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})}{n},\frac{\sum_{i=1}^n \nabla f_i(z_i^{(k)})}{n}\right\rangle \\
&= \mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})}{n}\right\|^2 \\
&\quad +\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla f_i(z_i^{(k)})}{n}\right\|^2. \qquad (20)
\end{aligned}$$

where in the last equality the inner product becomes zero from the fact that $\mathbb{E}_{\xi_i^{(k)}}\nabla F_i(z_i^{(k)};\xi_i^{(k)})=\nabla f_i(z_i^{(k)})$.

$$\begin{aligned}
&\mathbb{E}_{\xi_i^{(k)}}\left\|\frac{\sum_{i=1}^n \nabla F_i(z_i^{(k)};\xi_i^{(k)})-\sum_{i=1}^n \nabla f_i(z_i^{(k)})}{n}\right\|^2 \\
&=\frac{1}{n^2}\mathbb{E}_{\xi_i^{(k)}}\left\|\sum_{i=1}^n\left[\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})\right]\right\|^2 \\
&=\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{\xi_i^{(k)}}\left\|\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})\right\|^2 \\
&\quad+\frac{2}{n^2}\sum_{i\neq j}\left\langle\mathbb{E}_{\xi_i^{(k)}}\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)}),\mathbb{E}_{\xi_j^{(k)}}\nabla F_j(z_j^{(k)};\xi_j^{(k)})-\nabla f_j(z_j^{(k)})\right\rangle \\
&=\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{\xi_i^{(k)}}\left\|\nabla F_i(z_i^{(k)};\xi_i^{(k)})-\nabla f_i(z_i^{(k)})\right\|^2 \\
&\overset{\text{Bounded Variance}}{\leq}\frac{1}{n^2}\sum_{i=1}^n\sigma^2=\frac{\sigma^2}{n}, \qquad (21)
\end{aligned}$$

□

Before presenting the proof of next lemma let us define the conditional expectation

$$\mathbb{E}[\cdot|\mathcal{F}_k]:=\mathbb{E}_{\xi_i^{(k)}\sim\mathcal{D}_i\forall i\in[n]}[\cdot]=\mathbb{E}_{\xi_i^{(k)}\forall i\in[n]}[\cdot].$$

The expectation in this expression is *only* with respect to the random choices $\xi_i^{(k)}$ for all nodes $i\in[n]$ at the k^{th} iteration. In addition, we should highlight that the choices of random variables $\xi_i^k\sim\mathcal{D}_i$, $\xi_j^k\sim\mathcal{D}_j$ at the step t of the algorithm, are independent for any two nodes $i\neq j\in[n]$. This is also true in the case that the two nodes follow the same distribution $\mathcal{D}=\mathcal{D}_i=\mathcal{D}_j$.

Lemma 8. *Let Assumptions 1-3 hold and let*

$$D_1=\frac{1}{2}-\frac{L^2}{2}\left(\frac{12\gamma^2C^2+3\gamma C^2}{(1-q)^2D_2}\right)\quad \textit{and}\quad D_2=1-\frac{\gamma^2 12L^2C^2}{(1-q)^2}-\frac{\gamma 3L^2C^2}{(1-q)^2}.$$

Here $C > 0$ and $q \in (0, 1)$ are the two non-negative constants defined in Lemma 3. Let $\{\mathbf{X}_k\}_{k=0}^{\infty}$ be the random sequence produced by (6) (Matrix representation of Algorithm 1). Then,

$$
\begin{aligned}
&\frac{1}{K}\left(D_1\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 + \frac{1-L\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right)\\
\le&\frac{f(\overline{\boldsymbol{x}}^{(0)})-f^*}{\gamma K} + \frac{L\gamma\sigma^2}{2n} + \frac{4L^2\gamma^2C^2\sigma^2+12L^2\gamma^2C^2\zeta^2}{2(1-q)^2D_2} + \frac{\gamma L^2C^2\sigma^2+3L^2\gamma C^2\zeta^2}{2K(1-q)^2D_2}\\
&+\left(\frac{L^2C+2L^2\gamma C^2}{2(1-q)^2D_2K}\right)\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n}.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
f\left(\overline{\boldsymbol{x}}^{(k+1)}\right) = f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) \quad &\overset{(6)}{=} \quad f\left(\frac{\mathbf{X}^{(k)}[\mathbf{P}^{(k)}]^\top\mathbf{1}_{\overline{\boldsymbol{n}}} - \gamma\nabla\boldsymbol{F}(\mathbf{Z}^{(k)},\boldsymbol{\xi}^{(k)})[\mathbf{P}^{(k)}]^\top\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\\
&= \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n} - \frac{\gamma\nabla\boldsymbol{F}(\mathbf{Z}^{(k)},\boldsymbol{\xi}^{k})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\\
&\overset{L-smooth}{\le} \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \gamma\left\langle\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right), \frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)},\boldsymbol{\xi}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\rangle\\
&\qquad + \frac{L\gamma^2}{2}\left\|\frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)},\boldsymbol{\xi}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2 \qquad (22)
\end{aligned}
$$

Taking expectations of both sides conditioned on $\mathcal{F}_k$:

$$
\begin{aligned}
\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)|\mathcal{F}_k\right] \quad &\le \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \gamma\left\langle\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right), \frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\rangle\\
&\qquad + \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)},\boldsymbol{\xi}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2|\mathcal{F}_k\right]\\
&\overset{Lemma\ 7[i]}{=} \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \gamma\left\langle\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right), \frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\rangle\\
&\qquad + \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\sum_{i=1}^n\nabla F_i(z_i^{(k)};\xi_i^{(k)}) - \sum_{i=1}^n\nabla f_i(z_i^{(k)})}{n}\right\|^2|\mathcal{F}_k\right]\\
&\qquad + \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\sum_{i=1}^n\nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2|\mathcal{F}_k\right]\\
&\overset{Lemma\ 7[ii]}{\le} \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \gamma\left\langle\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right), \frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\rangle\\
&\qquad + \frac{L\gamma^2\sigma}{2n} + \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\sum_{i=1}^n\nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2|\mathcal{F}_k\right]\\
&= \quad f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \frac{\gamma}{2}\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2 - \frac{\gamma}{2}\left\|\frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2,\\
&\qquad + \frac{\gamma}{2}\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right) - \frac{\nabla\boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2 + \frac{L\gamma^2\sigma^2}{2n}\\
&\qquad + \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\sum_{i=1}^n\nabla f_i(z_i^{(k)})}{n}\right\|^2|\mathcal{F}_k\right] \qquad (23)
\end{aligned}
$$

where in the last step above we simply expand the inner product.

Taking expectations with respect to $\mathcal{F}_k$ and using the tower property, we get

$$\begin{aligned}
\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right] &\leq \mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right], \\
&+ \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)-\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right]+\frac{L\gamma^2\sigma^2}{2n} \\
&+ \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2\right] \\
&= \mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]-\frac{\gamma-L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right], \\
&+ \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)-\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right]+\frac{L\gamma^2\sigma^2}{2n} \qquad (24)
\end{aligned}$$

Let us now focus on find an upper bound for the quantity $\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)-\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right]$.

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)-\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right] &= \mathbb{E}\left[\left\|\nabla f\left(\overline{\boldsymbol{x}}\right)-\frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_i^n \nabla f_i\left(\overline{\boldsymbol{x}}\right)-\frac{\sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{\sum_i^n \nabla f_i\left(\overline{\boldsymbol{x}}\right)-\sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{(k)})}{n}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_i^n \left[\nabla f_i\left(\overline{\boldsymbol{x}}\right)-\nabla f_i(\boldsymbol{z}_i^{(k)})\right]\right\|^2\right] \\
&\overset{Jensen}{\leq} \frac{1}{n}\sum_i^n \mathbb{E}\left[\left\|\nabla f_i\left(\overline{\boldsymbol{x}}\right)-\nabla f_i(\boldsymbol{z}_i^{(k)})\right\|^2\right] \\
&\overset{L-smooth}{\leq} \frac{L^2}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\overline{\boldsymbol{x}}-\boldsymbol{z}_i^{(k)}\right\|^2\right] \\
&= \frac{L^2}{n}\sum_{i=1}^n Q_i^{(k)} \qquad (25)
\end{aligned}$$

Thus we have that:

$$\begin{aligned}
\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right] &\leq \mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]-\frac{\gamma-L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right], \\
&+ \frac{\gamma L^2}{2n}\sum_{i=1}^n Q_i^{(k)}+\frac{L\gamma^2\sigma^2}{2n} \qquad (26)
\end{aligned}$$

By rearranging:

$$\begin{aligned}\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]+\frac{\gamma-L\gamma^2}{2}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right] \quad &\leq \quad \mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]\\ &+\quad \frac{L\gamma^2\sigma^2}{2n}+\frac{\gamma L^2}{2n}\sum_{i=1}^{n}Q_i^{(k)} \qquad (27)\end{aligned}$$

Let us now sum from $k=0$ to $k=K-1$:

$$\begin{aligned}\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]+\frac{\gamma-L\gamma^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right] \quad &\leq \quad \sum_{k=0}^{K-1}\left[\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k+1)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]\right]\\ &+\quad \sum_{k=0}^{K-1}\frac{L\gamma^2\sigma^2}{2n}+\frac{\gamma L^2}{2n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}Q_i^{(k)}\\ &\leq\quad \mathbb{E}\left[f\left(\frac{\mathbf{X}^{(0)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]-\mathbb{E}\left[f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right]\\ &+\quad \frac{LK\gamma^2\sigma^2}{2n}+\frac{\gamma L^2}{2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}Q_i^{(k)}\\ &\leq\quad f(\overline{\boldsymbol{x}}^{(0)})-f^*\\ &+\quad \frac{LK\gamma^2\sigma^2}{2n}+\frac{\gamma L^2}{2}\sum_{k=0}^{K-1}\underbrace{\frac{1}{n}\sum_{i=1}^{n}Q_i^{(k)}}_{M_k} \qquad (28)\end{aligned}$$

For the last inequality above, recall that we let f^* denote the global infimum of our problem.

Using the bound for the expression $\sum_{k=0}^{K-1}M_k$ from Lemma 6 we obtain:

$$\begin{aligned}&\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}^{(k)}\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right)\right\|^2\right]+\frac{\gamma-L\gamma^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right]\\ \leq& f(\overline{\boldsymbol{x}}^{(0)})-f^*+\frac{LK\gamma^2\sigma^2}{2n}\\ &+\frac{\gamma L^2}{2}\frac{4\gamma^2C^2\sigma^2K+\gamma C^2\sigma^2}{(1-q)^2D_2}+\frac{\gamma L^2}{2}\frac{12\gamma^2C^2\zeta^2K+3\gamma C^2\zeta^2}{(1-q)^2D_2}\\ &+\frac{\gamma L^2}{2}\left(\frac{12\gamma^2C^2+3\gamma C^2}{(1-q)^2D_2}\right)\sum_{k=0}^{K}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2\\ &+\frac{\gamma L^2}{2}\left(\frac{C^2+2\gamma C^2}{(1-q)^2D_2}\right)\frac{\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n}.\end{aligned}$$

By rearranging and dividing all terms by γK we obtain:

$$\begin{aligned}&\frac{1}{K}\left(\left[\frac{1}{2}-\frac{L^2}{2}\left(\frac{12\gamma^2C^2+3\gamma C^2}{(1-q)^2D_2}\right)\right]\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2+\frac{1-L\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}^{(k)})\mathbf{1}_{\overline{\boldsymbol{n}}}}{n}\right\|^2\right)\\ \leq&\frac{f(\overline{\boldsymbol{x}}^{(0)})-f^*}{\gamma K}+\frac{L\gamma\sigma^2}{2n}+\frac{4L^2\gamma^2C^2\sigma^2+12L^2\gamma^2C^2\zeta^2}{2(1-q)^2D_2}+\frac{\gamma L^2C^2\sigma^2+3L^2\gamma C^2\zeta^2}{2K(1-q)^2D_2}\\ &+\left(\frac{L^2C^2+2L^2\gamma C^2}{2(1-q)^2D_2K}\right)\frac{\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n}.\end{aligned}$$

By defining $D_1 = \left[\frac{1}{2} - \frac{L^2}{2}\left(\frac{12\gamma^2C^2+3\gamma C^2}{(1-q)^2D_2}\right)\right]$ the proof is complete. □

E.3. Proofs of Main Theorems

Having present all of the above Lemmas we are now ready to provide the proofs of main Theorems 1 and 2.

E.3.1. Proof of Theorem 1

Let $\gamma \leq \min\left\{\frac{(1-q)^2}{60L^2C^2}, 1\right\}$. Then:

$$D_2 = 1 - \frac{\gamma^2 12L^2C^2}{(1-q)^2} - \frac{\gamma 3L^2C^2}{(1-q)^2} \overset{(\gamma^2<\gamma)}{\geq} 1 - \frac{\gamma 15L^2C^2}{(1-q)^2} \geq 1 - \frac{1}{4} \geq \frac{1}{2}$$

and

$$D_1 = \frac{1}{2} - \frac{L^2}{2}\left(\frac{12\gamma^2C^2 + 3\gamma C^2}{(1-q)^2D_2}\right) \overset{(\gamma^2<\gamma)}{\geq} \frac{1}{2} - \frac{15\gamma C^2L^2}{2(1-q)^2D_2} \geq \frac{1}{2} - \frac{1}{8D_2} \geq \frac{1}{4}$$

By substituting the above bounds into the result of Lemma 8 and by removing the second term of left hand side we obtain:

$$\begin{aligned}
\frac{1}{4}\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2}{K} &= \frac{1}{K}\left(\frac{1}{4}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 + \frac{1-L\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{\nabla \boldsymbol{F}(\mathbf{Z}_k)\mathbf{1_n}}{n}\right\|^2\right) \\
&\leq \frac{f(\overline{\boldsymbol{x}}^{(0)}) - f^*}{\gamma K} + \frac{L\gamma\sigma^2}{2n} + \frac{4L^2\gamma^2C^2\sigma^2 + 12L^2\gamma^2C^2\zeta^2}{(1-q)^2} + \frac{\gamma L^2C^2\sigma^2 + 3L^2\gamma C^2\zeta^2}{K(1-q)^2} \\
&+ \left(\frac{L^2C + 2L^2\gamma C^2}{(1-q)^2K}\right)\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} \qquad (29)
\end{aligned}$$

Let us now substitute in the above expression $\gamma = \sqrt{\frac{n}{K}}$. This can be done due to the lower bound (see equation 3) on the total number of iterations K where guarantees that $\sqrt{\frac{n}{K}} \leq \min\left\{\frac{(1-q)^2}{60L^2C^2}, 1\right\}$.

$$\begin{aligned}
\frac{1}{4}\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2}{K} &\leq \frac{f(\overline{\boldsymbol{x}}^{(0)}) - f^*}{\gamma K} + \frac{L\gamma\sigma^2}{2n} + \gamma^2\frac{4L^2C^2\sigma^2 + 12L^2C^2\zeta^2}{(1-q)^2} + \gamma\frac{L^2C^2\sigma^2 + 3L^2C^2\zeta^2}{K(1-q)^2} \\
&+ \frac{L^2C}{(1-q)^2K}\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} + \gamma\frac{2L^2C^2}{(1-q)^2K}\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} \\
&\overset{\gamma=\sqrt{\frac{n}{K}}}{=} \frac{f(\overline{\boldsymbol{x}}^{(0)}) - f^*}{\sqrt{nK}} + \frac{L\sigma^2}{2\sqrt{nK}} + \frac{n}{K}\frac{4L^2C^2\sigma^2 + 12L^2C^2\zeta^2}{(1-q)^2} + \sqrt{\frac{n}{K}}\frac{L^2C^2\sigma^2 + 3L^2C^2\zeta^2}{K(1-q)^2} \\
&+ \frac{L^2C^2}{(1-q)^2K}\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} + \sqrt{\frac{n}{K}}\frac{2L^2C^2}{(1-q)^2K}\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} \\
&= \frac{f(\overline{\boldsymbol{x}}^{(0)}) - f^* + \frac{L}{2}\sigma^2}{\sqrt{nK}} + \frac{L^2C^2}{K(1-q)^2}\left[(4\sigma^2 + 12\zeta^2)n + \frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n}\right] \\
&+ \frac{\sqrt{n}L^2C^2}{\sqrt{K}(1-q)^2K}\left[\sigma^2 + 3L^2C^2\zeta^2 + 2\frac{\sum_{i=1}^n\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n}\right] \qquad (30)
\end{aligned}$$

Using again the assumption on the lower bound (3) of the total number of iterations K, the last two terms of the above expression are bounded by the first term. Thus,

$$\frac{1}{4}\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2}{K} \leq 3\frac{f(\overline{\boldsymbol{x}}^{(0)}) - f^* + \frac{L}{2}\sigma^2}{\sqrt{nK}} \qquad (31)$$

E.3.2. Proof of Theorem 2

Proof. From Lemma 6 we have that:

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1} M^{(k)} \quad \leq \quad & \left(\gamma^2 \frac{4C^2}{(1-q)^2 D_2}\right)\sigma^2 + \left(\gamma \frac{C^2}{(1-q)^2 D_2}\right)\frac{\sigma^2}{K} \\
& + \left(\gamma^2 \frac{12C^2}{(1-q)^2 D_2}\right)\zeta^2 + \left(\frac{\gamma 3C^2}{(1-q)^2 D_2}\right)\frac{\zeta^2}{K} \\
& + \left(\frac{C^2}{(1-q)^2 D_2 K} + \gamma \frac{2C^2}{(1-q)^2 D_2 K}\right)\frac{\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} \\
& + \left(\gamma^2 \frac{12C^2}{(1-q)^2 D_2} + \gamma \frac{3C^2}{(1-q)^2 D_2}\right)\frac{\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2}{K} \qquad (32)
\end{aligned}
$$

Using the assumptions of Theorem 1 and stepsize $\gamma = \sqrt{\frac{n}{K}}$:

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1} M^{(k)} \quad \leq \quad & \left(\frac{n}{K}\frac{4C^2}{(1-q)^2 D_2}\right)\sigma^2 + \left(\sqrt{\frac{n}{K}}\frac{C^2}{(1-q)^2 D_2}\right)\frac{\sigma^2}{K} \\
& + \left(\frac{n}{K}\frac{12C^2}{(1-q)^2 D_2}\right)\zeta^2 + \left(\frac{\sqrt{\frac{n}{K}}3C^2}{(1-q)^2 D_2}\right)\frac{\zeta^2}{K} \\
& + \left(\frac{C^2}{(1-q)^2 D_2 K} + \sqrt{\frac{n}{K}}\frac{2C^2}{(1-q)^2 D_2 K}\right)\frac{\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n} \\
& + \left(\frac{n}{K}\frac{12C^2}{(1-q)^2 D_2} + \sqrt{\frac{n}{K}}\frac{3C^2}{(1-q)^2 D_2}\right)\frac{12\left[f(\overline{\boldsymbol{x}}^{(0)}) - f^* + \frac{L}{2}\sigma^2\right]}{\sqrt{nK}} \\
= \quad & \frac{1}{K}\left[\frac{4nC^2\sigma^2}{(1-q)^2 D_2} + \frac{12nC^2\zeta^2}{(1-q)^2 D_2} + \frac{C^2\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{n(1-q)^2 D_2} + \frac{3\sqrt{n}C^2 12\left[f(\overline{\boldsymbol{x}}^{(0)}) - f^* + \frac{L}{2}\sigma^2\right]}{\sqrt{n}(1-q)^2 D_2}\right] \\
+ \quad & \frac{1}{K\sqrt{K}}\left[\frac{n\sigma^2 C^2}{(1-q)^2 D_2} + \frac{\frac{n}{3}C^2\zeta^2}{(1-q)^2 D_2} + \frac{2C^2\sum_{i=1}^{n}\left\|\boldsymbol{x}_i{}^{(0)}\right\|^2}{(1-q)^2 D_2\sqrt{n}} + \frac{144\sqrt{n}C^2\left[f(\overline{\boldsymbol{x}}^{(0)}) - f^* + \frac{L}{2}\sigma^2\right]}{(1-q)^2 D_2}\right] \\
= \quad & O\left(\frac{1}{K} + \frac{1}{K\sqrt{K}}\right) \qquad (33)
\end{aligned}
$$

where the Big O notation swallows all constants of our setting $\left(n, L, \sigma, \zeta, C, q, \sum_{i=1}^{n}\left\|x_i{}^{(0)}\right\|^2 \text{and} f(\overline{\boldsymbol{x}}^{(0)}) - f^*\right)$.

Now using the above upper bound equation 33 and result of Theorem 1 we obtain:

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\boldsymbol{z}_i^k)\right\|^2 &= \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\boldsymbol{z}_i^k)+\nabla f(\overline{\boldsymbol{x}}^{(k)})-\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\leq \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}2\mathbb{E}\left\|\nabla f(\boldsymbol{z}_i^k)-\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2+2\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&= \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}2\mathbb{E}\left\|\nabla f(\boldsymbol{z}_i^k)-\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}2\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&= 2\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\boldsymbol{z}_i^k)-\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2+2\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\overset{L-smooth}{=} 2L^2\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\boldsymbol{z}_i^k-\overline{\boldsymbol{x}}^{(k)}\right\|^2+2\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\overset{(33)+(31)}{\leq} O\left(\frac{1}{\sqrt{nK}}+\frac{1}{K}+\frac{1}{K^{3/2}}\right) \qquad (34)
\end{aligned}$$

where again the Big O notation swallows all constants of our setting $\left(n, L, \sigma, \zeta, C, q, \sum_{i=1}^{n}\left\|x_i{}^{(0)}\right\|^2 \text{ and} f(\overline{\boldsymbol{x}}^{(0)})-f^*\right)$.

□