DRACO: Decentralized Asynchronous Federated Learning over Row-Stochastic Wireless Networks

Eunjeong Jeong, *Member, IEEE*, and Marios Kountouris, *Fellow, IEEE*.

***Abstract*—Emerging technologies and use cases, such as smart Internet of Things (IoT), Internet of Agents, and Edge AI, have generated significant interest in training neural networks over fully decentralized, serverless networks. A major obstacle in this context is ensuring stable convergence without imposing stringent assumptions, such as identical data distributions across devices or synchronized updates. In this paper, we introduce DRACO, a novel framework for decentralized asynchronous Stochastic Gradient Descent (SGD) over row-stochastic gossip wireless networks. Our approach leverages continuous communication, allowing edge devices to perform local training and exchange model updates along a continuous timeline, thereby eliminating the need for synchronized timing. Additionally, our algorithm decouples communication and computation schedules, enabling complete autonomy for all users while effectively addressing straggler issues. Through a thorough convergence analysis, we show that DRACO achieves high performance in decentralized optimization while maintaining low variance across users even without predefined scheduling policies. Numerical experiments further validate the effectiveness of our approach, demonstrating that controlling the maximum number of received messages per client significantly reduces redundant communication costs while maintaining robust learning performance.**

***Index Terms*—Collaborative intelligence, continuous time models, decentralized learning, asynchronous learning, federated learning, row-stochastic matrices, networked intelligent systems, peer-to-peer networks.**

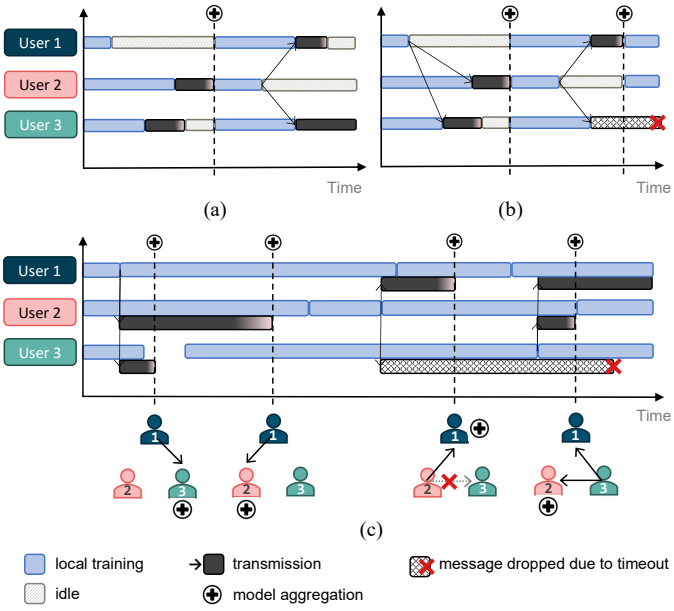


Fig. 1. A schematic view of DRACO's timelines with comparisons. (a) Synchronous FL; (b) asynchronous FL with transmission delay deadline; (c) (in DRACO) fully asynchronous FL with delay deadline, but the iteration count is continuous.

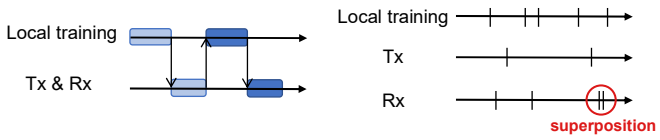
I. INTRODUCTION

RECENT advancements in machine learning and wireless connectivity have paved the way for various innovative applications across various sectors such as the Internet of Things (IoT), consumer robotics, autonomous transportation, and edge computing. These systems promise reduced latency, more efficient bandwidth usage, and enhanced privacy through local data processing. However, achieving these benefits requires overcoming the significant challenge of maintaining reliable learning in dynamic and often unstable network environments. This unreliability calls for innovative approaches that can seamlessly adapt to decentralized architectures while ensuring both robustness and scalability.

In this work, we focus on enhancing communication efficiency in federated learning (FL) [1], particularly within fully decentralized (serverless) environments where collaboration occurs without a central coordinator [2]–[6]. Asynchronous learning, which allows independent training and communication, is essential in such networks [7]–[11]. While both decentralization and asynchrony offer significant advantages in flexibility and resource efficiency, their combination introduces additional challenges in achieving global consensus without incurring substantial synchronization overhead [12].

A key challenge in decentralized asynchronous learning is the uncertainty of convergence. Gossip-based communication, which relies on probabilistic message exchanges, has emerged as a promising alternative to rigid scheduling [13]–[17], offering better adaptability to fluctuating network conditions compared to predefined schedules [18]. However, most existing methods depend on idealized assumptions that limit their applicability to specific network types. In particular, many approaches assume doubly stochastic communication weights [19], implying symmetric information exchange and equal influence between neighbors, which is an unrealistic assumption in directed and unreliable wireless networks. Furthermore, even more recent methods designed for directed graphs often presume that users are aware of successful

E. Jeong is with the Department of Computer and Information Science, Linköping University, Sweden. M. Kountouris is with the Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Department of Computer Science and Artificial Intelligence, University of Granada, Spain, and with the Communication Systems Dept., EURECOM, France. The first author conducted the research during her employment at EURECOM, France. The work of M. Kountouris has been supported in part by the Horizon Europe JU SNS project ROBUST-6G (Grant Agreement no. 101139068).



(a) Sequential iterations

(b) Decoupled iterations with superposition windows

Fig. 2. (a) Sequential computation and communication over a doubly stochastic network; (b) Timelines of DRACO with decoupled computation and communication over a row-stochastic network. If two messages arrive at the same agent with a negligibly small time gap (in red circle), they are considered simultaneous and are used for the same model aggregation step. The concept of the superposition window is elaborated in Section II-II-B.

outbound message deliveries [20]–[22]. This strong requirement may not hold when the system has packet losses or unidirectional links [23].

To address these limitations, we introduce DRACO, a novel framework for decentralized asynchronous FL. DRACO is built on two core principles: (i) continuous, fully asynchronous operation without reliance on global iteration counters, and (ii) decoupled communication and computation through gradient pushing. By eliminating synchronized timing, DRACO enables nodes to transmit updates at flexible, non-integer time instants. Although this introduces variability in node progress, it significantly reduces idle time and enhances learning efficiency. Additionally, DRACO mitigates the effects of stale gradients by discarding excessively delayed messages, ensuring timely and effective optimization.

DRACO also tackles the communication overhead caused by frequent updates in sequential iterations by decoupling communication and computation, as illustrated in Fig. 2b, allowing independent management of local training and message exchanges. The separated action timelines enable nodes to adjust their transmission schedules dynamically, enhancing network efficiency and preventing performance oscillations from redundant messages. Another advantage of decoupling is that the integrated learning process is less likely to stagnate when local training and transmission occur independently. A fully asynchronous network with all users actively engaged, as shown in Fig. 1c, serves as an exemplary case. In such environments, if users always forward their updated models after aggregating local updates from their one-hop neighbors, as in Fig. 2a, the resulting communication overhead can be nearly twice as high as in push-based collaboration. Furthermore, the content delivered between nodes is often duplicated or overwritten. Thus, separating the two schedules differentiates DRACO from conventional approaches that mandate a sequential or predetermined order for gradient updates and gossip communications.

In summary, this paper presents DRACO as a practical and scalable solution for decentralized asynchronous learning over wireless networks. Our main contributions are as follows:

- Asynchronous and Continuous Learning: DRACO eliminates rigid transmission schedules, enabling seamless adaptation to dynamic network conditions.
- Realistic Asynchrony Management: DRACO treats asynchrony and serverless operation as inherent system characteristics, ensuring resilience to network variability.
- Efficient Message Control: DRACO optimizes the trade-off between learning efficiency and communication overhead by dynamically managing message transmissions and stale information, while enhancing stability and accelerating convergence.

We demonstrate through rigorous theoretical analysis and extensive numerical experiments that DRACO achieves high decentralized optimization performance with low user variance, even under unreliable communication conditions.

A. Related Works

Asynchronous decentralized learning In synchronous learning systems, all participants should wait for the slowest learner (straggler) before proceeding to the next global round. As depicted in Fig. 1b, asynchronous learning with a transmission delay deadline effectively reduces the overall training time of synchronous systems by excluding users whose updates arrive after a predetermined deadline [15], [24]. This approach is applied not only to asynchronous settings but to synchronous learning through partial participation [25]. Both asynchronous learning and partially participating synchronous learning face the challenge of variance reduction since only a subset of local updates is considered in each training round [26]–[29]. Despite fewer average participation in model aggregation per user compared to synchronous methods, asynchronous learning performs as well as its counterpart, especially in solving large-scale multi-user optimization problems [30]. Nevertheless, this approach requires users to start their computations simultaneously to synchronize the global phase, leading to idle times when a message arrives before the start of the next iteration. Additionally, sufficient local storage is necessary to manage multiple messages queued in the receive buffer until the next round.

Randomized communication over serverless and directed networks Recent studies in decentralized learning have explored algorithms implementable for networks modeled by directed graphs, where the connectivity matrix is not necessarily doubly stochastic. This adaptation is often necessary when neither full-duplex nor half-duplex systems can ensure stable gradient transmissions. Techniques, such as push-sum [20], [31]–[36], push-pull [37], [38], and random walk [39]–[41], have been proposed to improve decentralized optimization on directed graphs. Meanwhile, row-stochastic communication [42] significantly reduces both the number of communication rounds and storage requirements on edge devices; hence, this benefits tackling complex problems, specifically those involving small-scale neural networks [43]. Among random communication protocols, the gossip protocol is well known for its rapid information spread, but is also criticized for its high network resource consumption [44]. Consequently, asynchronous gossip learning in such contexts needs innovative approaches to manage information flow among edge devices [45].

Distributed optimization over directed graphs While many existing studies assume doubly stochastic weight matrices, this condition restricts their applicability to arbitrarily

directed communication networks, thus limiting the practical scope of these algorithms [23]. The study of distributed optimization over directed graphs has a well-established history in control theory [43], [46]–[48], where the relevant researchers developed fundamental methodologies to handle asymmetric and unbalanced communication topologies. More recently, FL research has begun to address challenges related to asymmetric connectivity, employing row-stochastic matrices [49] and accommodating time-varying directed graph structures [50]. Building on these advances, some investigations have further incorporated personalization techniques [51] to better manage the inherent data heterogeneity across participants in directed settings [52]. Nevertheless, a significant open question remains: *how resilient are decentralized learning systems when faced with unreliable communication links?*, which is a critical concern for real-world wireless and edge computing deployments.

Decoupling communication and computation Unlike traditional methods that align gradient computation and communication either sequentially or in parallel, decoupling these processes significantly accelerates peer-to-peer averaging by releasing clients from waiting for others [53]–[55]. In AD-OGP [56], the authors replaced global communication slots with an event-based aggregation system, including activities such as prediction and local updating. This unified timeline of events is particularly well-suited for environments where users train locally at different computational speeds. However, the event types of AD-OGP are restricted to "prediction" and "local computation", overlooking the impact of transmission delay. The authors assume that message delay, defined as the time gap between the latest prediction and the local updating event within a user, provides no insight into how long it takes for a message to reach a neighboring node. Despite the growing interest in approaches for effective timeline integration, only a few studies have explored decoupled model averaging over unreliable wireless networks, where issues such as packet loss or delays are prevalent.

II. SYSTEM MODEL

We consider the following optimization task over N clients, whose goal is to minimize

$$f(\mathbf{x}) := \frac{1}{N}\sum_{i\in\mathcal{U}} f_i(\mathbf{x}) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is an d-dimensional model parameter and $\mathcal{U} = \{1, \cdots, N\}$ is the set of network users. In a serverless network, there is no global model $\mathbf{x}_t$; instead, each agent i holds $\mathbf{x}_t^{(i)}$, which serves as a reference for the globally acquired model. Therefore, we rephrase the objective function as

$$\mathbf{x}^* = \inf_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{N} f_i(\mathbf{x}^{(i)}) \ . \tag{2}$$

To tackle the minimization problem described in (1) or (2), we adopt a decentralized stochastic gradient descent (DSGD) approach. In this approach, individual devices iteratively enhance their local models $\mathbf{x}^{(i)}$ and subsequently share these estimates with their neighboring nodes, which in turn could vary over time.

A. Absence of a global belief

The underlying assumption regarding the global consensus is that each user cannot reach a global "true parameter", denoted by $\mathbf{x}^*$, by local updates only. The global model $\mathbf{x}$ should be a vector combined with the beliefs (pseudo-global model) at each agent, said $\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}\}$. However, in practice, none of the agents work as a central server or aggregator, which can obtain a centralized global model. We therefore adopt a virtual global model $\bar{\mathbf{x}}$ that could have been acquired through the superposition of all beliefs if the network had an entirely authorized server, i.e.,

$$\bar{\mathbf{x}} = \mathbb{E}_{i\in\mathcal{U}}[\mathbf{x}^{(i)}] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}.$$

Therefore, when P seconds have elapsed since the initial moment t_0, we rewrite the virtual global model difference between these two time instances as

$$\bar{\mathbf{x}}_{t_0+P} - \bar{\mathbf{x}}_{t_0} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_{t_0+P}^{(i)} - \mathbf{x}_{t_0}^{(i)}\right).$$

B. Communication system

In our work, the processes of computation and communication are decoupled; hence, when to train locally and when to transmit the updates are determined independently by each user. Since there can be infinite instants between any two close events on the continuous timeline, each message will likely arrive at a different moment. Thus, practically speaking, there is no aggregation during the entire process, even though two updates arrive at the same destination node by a narrow margin of time. In this regard, we introduce a *superposition window*, which is analogous to congestion windows in TCP (Transmission Control Protocol) [57]. Like a TCP window, the superposition window in DRACO controls the flow of received updates by grouping the messages for one aggregation. This superposing function leads to lower computation costs since the clients avoid renewing the local reference model every time a message arrives.

This paper investigates the effect of unreliable wireless communications and controlled transmissions on the performance of DRACO. Specifically, we consider scenarios with time-invariant connectivity graphs, which represent stable network topologies during the learning process. This simplified assumption facilitates a focused analysis of how the frequency of successful message receptions and the underlying structure of the communication network affect convergence in the learning process. Unlike traditional fixed-topology models, we explicitly account for the inherent unreliability of wireless channels. In our model, successful message delivery between connected nodes is probabilistic and not guaranteed, influenced by factors such as physical distance, interference, and channel capacity limitations. User nodes are randomly distributed, and their geographical positions directly affect communication probabilities. To provide a more comprehensive understanding

beyond standard fixed-topology analyses, we examine the effect of the frequency of successful message receptions within a defined unification period (detailed in Section III-A). This approach provides a more nuanced assessment of the learning process under various wireless channel conditions.

A weighted graph at a certain instance is mathematically defined as an $N \times N$-sized matrix where each element indicates whether node i transmits its message to one of its neighbors j or not. It follows a conditional probability distribution if a communication event exists on client i. Transmission incidents are defined as

$$q_k^i = \begin{cases} 1, & \text{if } i \text{ broadcasts } \Delta^{(i)} \text{ at } k \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$$q_k^{ij} = \begin{cases} 1, & \text{if } j \text{ receives } i\text{'s message sent at } k \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where k is the index of an event. We define the neighborhood of user i, denoted by $\mathcal{N}_t(i) = \{j | q_t^{ij} = 1\}$, as the set of all users j that have an edge going from i to j at time t. It is also possible to denote the neighbor set concerning event k, such as $\mathcal{N}_k(i)$. Following the notation in [25], these participation indicators are normalized across all moments, i.e., $\sum_{j \in \mathcal{U} \setminus \{i\}} q_t^{ij} = 1$ for $q_t^{ij} \geq 0$ and for all $i,\ t$. In addition to the definition, we define $\rho < 1$ that satisfies $\sum_{j \in \mathcal{U} \setminus \{i\}} (q_t^{ij})^2 \leq \rho^2$ for all $i,\ t$.

C. Local gradient computations

Each user performs stochastic gradient computations by iterating B batches of the local training datasets. Δ represents the local update of the model, defined as the difference between the model's state before the mini-batch training and its state after completing training on B batches of training samples.

Assumption 1. *(Exponential local gradient computation time.)* The computation time τ_i of the stochastic gradient $g_i(\mathbf{x}) \in \mathbb{R}^d$ at user i is exponentially distributed, i.e., $\tau_i \sim exp(\lambda_i)$.

In point processes, one can consider a PPP along the real line by examining the point count within a specific interval $[t_0, t_0 + P]$ [58]. For a homogeneous PPP with rate parameter $\lambda > 0$, the likelihood that the count of points, denoted by $num(t_0, t_0 + P]$, equals a certain integer m can be described by the following expression:

$$\Pr\{num(t_0, t_0 + P] = m\} = \frac{(\lambda P)^m}{m!} e^{-\lambda P}.$$

This formula calculates the probability of exactly m occurrences within the interval based on the Poisson distribution, where λP represents the expected number of points in the interval and $e^{-\lambda P}$ adjusts for the total rate of occurrences over the span.

III. DRACO: PROPOSED DECENTRALIZED ASYNCHRONOUS LEARNING

The primary rationale behind the proposed algorithm is to answer the following question: *How can we issue instructions to each user in the absence of a global time loop in the network?* To resolve this issue, we design the system such that each node focuses solely on its actions without considering the other nodes' training progress or channel conditions. Defining the algorithm within a unified time loop in asynchronous and fully decentralized networks presents several practical limitations in real-world systems. A significant challenge lies in the absence of a consistent global time reference, such as global iteration rounds, denoted as t, or timestamps marking the completion of each user's local computations, marked as k. This inconsistency arises because the total number of local training iterations varies across users, even when their updates are observed simultaneously. As a result, if the algorithm mandates exchanges every t_P seconds or every k_P global slots, some local models may fall behind in development due to completing fewer local training steps than others. To address this issue, we avoid defining the procedure as either sequential or simultaneous. Instead, our algorithm adopts a unified global loop where all users work in parallel. This global loop effectively encapsulates the learning process conducted by each user.

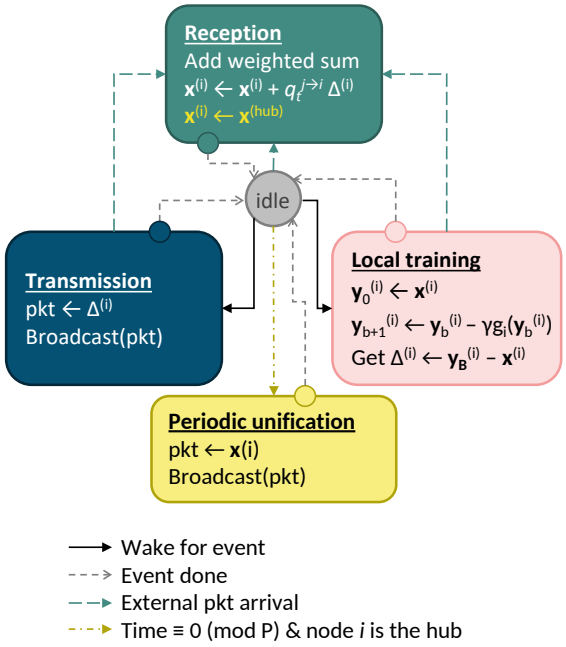


Fig. 3. The proposed algorithm (DRACO) in a chain graph illustrating the states of possible actions within each agent i.

At each instance, every user selects one of the following three statuses based on a probability distribution: (1) remaining idle, (2) transmitting a message to neighboring nodes, or (3) conducting local model training. Local computation involves batch training iterated B times to compute the update, termed Δ. During transmission, the user broadcasts its local update. If a node recognizes delivery from the other nodes, it switches to a fourth (4) status (receiving mode), renewing its reference model by aggregating the model updates from neighboring nodes. Unlike these four statuses, a node turns to the fifth (5) status when a periodic timeout occurs. As depicted in the yellow box in Figure 3, a temporary hub broadcasts its

Algorithm 1 User-centric algorithm of DRACO. A pseudo-algorithm for source code reproduction is provided in Appendix E.

for $i = 1, \cdots, N$ **do**
  while $t < T$ **do**
    $t \leftarrow \text{clock}()$
    if there is an event at time t **then**
      if the event is a gradient updating step **then**
        $\mathbf{y}_0^{(i)} \leftarrow \mathbf{x}^{(i)}$
        for $b = 0, \cdots, B-1$ **do**
          $\mathbf{y}_{b+1}^{(i)} \leftarrow \mathbf{y}_b^{(i)} - \gamma g_i(\mathbf{y}_b^{(i)})$ {*batch training*}
        end for
        $\Delta^{(i)} \leftarrow \mathbf{y}_B^{(i)} - \mathbf{x}^{(i)}$ {*local update evaluation*}
      else if the event is a transmission step **then**
        i sends $\Delta^{(i)}$ to its neighbors
        for $j \in \mathcal{N}(i)$ **do**
          j receives $\tilde{\Delta}^{(i)}$
          $\mathbf{x}^{(j)} \leftarrow \mathbf{x}^{(j)} + \sum_{j \neq i} q_t^{ij} \tilde{\Delta}^{(i)}$ {*aggregation*}
        end for
      end if
    end if
    if $t \equiv 0 \pmod{P}$ and i is the hub at t **then**
      i broadcasts $\mathbf{x}^{(i)}$
      for $j \in \mathcal{N}(i)$ **do**
        j receives $\tilde{\mathbf{x}}^{(i)}$
        $\mathbf{x}^{(j)} \leftarrow \tilde{\mathbf{x}}^{(i)}$ {*periodic unification*}
      end for
    end if
  end while
end for

reference model instead of a local update when the time is a multiple of the period P. The corresponding explanation as a form of algorithm is provided in Algorithm 1 on page 5.

Note that the 'idle' state is included since we assume that the agents alter their states instantaneously, i.e., without delay. The node's status is considered idle when it does none of the aforementioned steps. However, in practice, any activity takes time to complete, implying that each timestamp represents the moment that each action just finished. By this interpretation, the participants do not have an actual break time in practical scenarios, which is also applied to the experiments in Section V.

Notations. $\mathcal{U}$ represents the set of participants within the network with $\mathcal{Q} := \{q_t^{ij}\}$ for all $i, j \in \mathcal{U}$ and t. Also, $\sum_{j \neq i}$ represents the summations of variables attributed to any user other than user i, i.e., $j \in \mathcal{U} \setminus \{i\}$. A user i's local update at time t is symbolized as $\Delta_t^{(i)}$. When a user i sends $\Delta^{(i)}$ or $\mathbf{x}^{(i)}$, the recipient j receives $\tilde{\Delta}^{(i)}$ or $\tilde{\mathbf{x}}^{(i)}$, which are identical to the sender's original contents if the transmission is free from distortion. Throughout this manuscript, the term 'update' is used only as a noun that signifies the result derived from the difference between a local reference model and a newly obtained model through batch training. To avoid potential confusion, any instances in this paper that involve the action of updating are called alternatively, such as 'renew' or 'iterate on'.

A. Periodic Unification

Local models tend to diverge when the network does not use a central server because no one synchronizes their different learning stages. Like conventional FedAvg, periodic unification can effectively resolve the variance-reduction problem among local reference models. A countable upper bound for the number of messages per unit time is required for analysis because otherwise, the losses diverge to infinity. It is also reasonable to assume that it is finite because, in real-life applications, messages are countable even though the number of definable instances is infinite. Based on this, Assumption 2 and Definition 1 are introduced as follows.

Assumption 2. *(Finite number of messages during a unit time period) During every period P, the number of messages that each user receives is finite.*

Definition 1. *(Maximum number of receiving messages per user) Let $\psi_i(t_{start}, t_{end})$ indicate the function that counts the number of messages arrived at user i since time t_{start} until time t_{end}. For any $i \in \mathcal{U}$ and $m \in [0, 1, \cdots, \lfloor \frac{T}{P} \rfloor - 1]$,*

$$\psi_i(mP, (m+1)P) \leq \Psi \ ,$$

where Ψ is the maximum number of messages that a user permits to receive during the time duration $[mP, (m+1)P)$.

The Ψ term not only justifies the number of messages to be countable but also functions as a communication budget per period. Interestingly, when a decentralized network has a fixed communication budget per unit time, performing many consensus steps can effectively reduce the error, even though each gossiping step renders low precision. [59]

IV. CONVERGENCE ANALYSIS

In this section, we analyze the convergence performance of DRACO. Following the common practice in the literature, we make the subsequent assumptions along with the objective function.

Assumption 3. *(Lipschitz gradient.) For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and for any $i \in \mathcal{U}$, there is a nonnegative L that satisfies*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|. \tag{5}$$

Assumption 4. *(Unbiased stochastic gradient with bounded variance.) For all $\mathbf{x}$, i,*

$$\mathbb{E}[g_i(\mathbf{x})|\mathbf{x}] = \nabla f_i(\mathbf{x}) \text{ and } \mathbb{E}\left[\|g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2 | \mathbf{x}\right] \leq \sigma^2 \tag{6}$$

Assumption 5. *(Bounded gradient divergence.) For all $t \in [0, T)$ and $i \in \mathcal{U}$, the gradient divergence is bounded by ζ, i.e.,*

$$\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\mathbf{x}_t)\|^2 \leq \zeta^2. \tag{7}$$

From Assumption 5, an alternative deviation of local gradients is derived as in Lemma 1.

Lemma 1. *(Deviation of local gradients) When $N > 4$, for all $\mathbf{x}$, t,*

$$\Big\| \sum_{j \in \mathcal{U}} q_t^{ji} \left[\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_j(\mathbf{x}_t^{(j)}) \right] \Big\|^2 \leq \frac{2N\zeta^2}{N-4} \ .$$

Proof. The left side of the inequality above can be rephrased as

$$\Big\| \sum_{j\in\mathcal{U}} q_t^{ji} \big[\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_j(\mathbf{x}_t^{(j)})\big] \Big\|^2 = \Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \sum_{j\in\mathcal{U}} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2.$$

By adding and subtracting $\nabla f(\mathbf{x}_t)$, we have

$$\begin{aligned}
&\Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \sum_{j\in\mathcal{U}} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \\
&= \Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t) - \sum_{j=1}^{N} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \\
&= \Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\mathbf{x}_t) + \frac{1}{N} \sum_{i'=1}^{N} \nabla f_{i'}(\mathbf{x}_t^{(i')}) \\
&\quad - \frac{1}{N} \sum_{i'=1}^{N} \sum_{j=1}^{N} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \\
&\overset{(a)}{\leq} 2 \big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\mathbf{x}_t) \big\|^2 \\
&\quad + 2 \Big\| \frac{1}{N} \sum_{i'=1}^{N} \big[\nabla f_{i'}(\mathbf{x}_t^{(i')}) - \sum_{j=1}^{N} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)})\big] \Big\|^2 \\
&\overset{(b)}{\leq} 2\zeta^2 + \frac{2}{N} \sum_{i'=1}^{N} \Big\| \nabla f_{i'}(\mathbf{x}_t^{(i')}) - \sum_{j=1}^{N} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2,
\end{aligned}$$

where (a) uses $(\|\mathbf{z}_1 + \mathbf{z}_2\|^2)/2 \leq \|\mathbf{z}_1\|^2 + \|\mathbf{z}_2\|^2$; (b) is from the definition of ζ^2 in Assumption 5 on the first term and Jensen's inequality on the second term. By rearranging the second term of the right side of the inequality, we get

$$\begin{aligned}
&\Big(1 - \frac{2}{N}\Big) \Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \sum_{j\in\mathcal{U}} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \\
&\leq 2\zeta^2 + \frac{2}{N} \sum_{i'\in\mathcal{U}\setminus\{i\}} \Big\| \nabla f_{i'}(\mathbf{x}_t^{(i')}) - \sum_{j=1}^{N} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \\
&\leq 2\zeta^2 + \frac{2}{N} \sum_{i'=1}^{N} \Big\| \nabla f_{i'}(\mathbf{x}_t^{(i')}) - \sum_{j=1}^{N} q_t^{ji'} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2.
\end{aligned}$$

With another rearrangement to the left side, the inequality becomes

$$\Big(1 - \frac{4}{N}\Big) \Big\| \nabla f_i(\mathbf{x}_t^{(i)}) - \sum_{j\in\mathcal{U}} q_t^{ji} \nabla f_j(\mathbf{x}_t^{(j)}) \Big\|^2 \leq 2\zeta^2.$$

□

Considering all the above assumptions, we obtain an upper bound on the expectation of the original objective's gradient when $\mathcal{Q}$ is given in advance.

TABLE I
COMMUNICATION AND COMPUTATION COST COMPARISON ACROSS SCHEMES: NUMBER OF TRANSMISSIONS (# TX.) AND LOCAL MODEL TRAINING SESSIONS (# LR.) MEASURED FOR OVERALL NETWORK RESOURCE CONSUMPTION.

	# Tx.	# Lr.
sync-symm (Choco-SGD)	530	265
sync-push	346 ± 65	603
async-symm	281 ± 60	500
async-push (DIGEST)	574 ± 39	375
DRACO	207 ± 118	323 ± 110

Theorem 1. *Let* $\mathcal{F} := f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x})$. *Under all the aforementioned assumptions, we have*

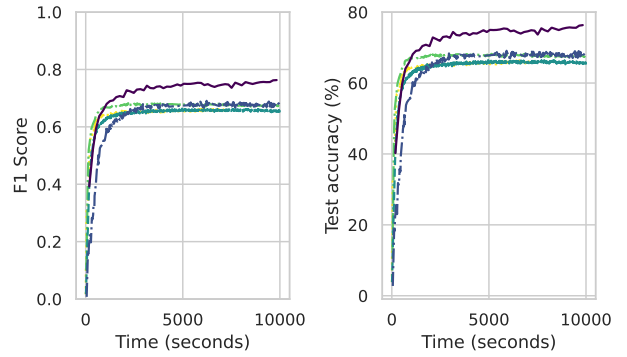
$$\begin{aligned}
&\min_t \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2 \big| \mathcal{Q}\big] \\
&\leq \mathcal{O}\Big(\frac{\mathcal{F}}{B\gamma\Psi} + \frac{\zeta^2}{N-4} + \sigma^2 + N\zeta^2 \\
&\qquad + BL^2\gamma^2\sigma^2 + \frac{L\gamma\rho^2\sigma^2}{N\Psi} \Big)
\end{aligned} \tag{8}$$

for $\gamma \leq \frac{1}{8BLN\Psi}$, $N > 4$, *and* $\Psi \geq 3$.

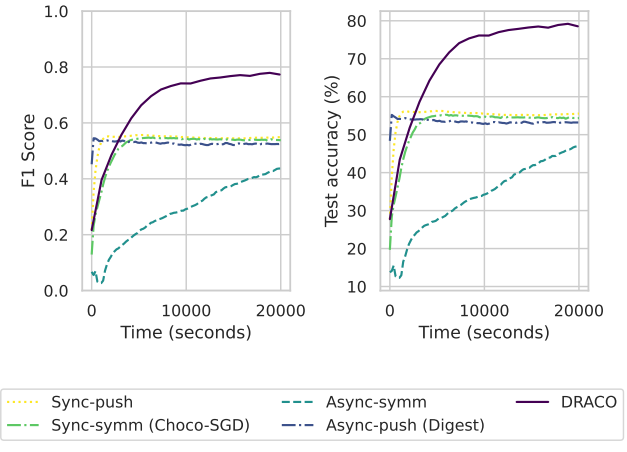
Remark. We begin, following a similar approach to [25], by deriving an inequality rooted in the smoothness of f_i. This inequality establishes a connection between two local losses from the same user at different timestamps, namely $f_i(\mathbf{x}_{t_0+P}^{(i)})$ and $f_i(\mathbf{x}_{t_0}^{(i)})$. Within this inequality, an inner product term unfolds into several components. Notably, it comprises three distinct subterms: one involving $\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2$ (refer to Lemma C.3), another featuring $\|\mathbf{x}_t^{(j)} - \mathbf{x}_{t_0}^{(j)}\|^2$ (see Lemma C.4) which is mainly derived from Algorithm 1, and a third term with $\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_j(\mathbf{x}_t^{(j)})\|^2$, of which the expectation has an upper bound (refer to Lemma 1). Our proof is novel in that it effectively converts and simplifies the terms on the continuous timeline into discrete values. Detailed proof is available in Appendix D. □

V. EXPERIMENTAL RESULTS

We conducted experiments with federated learning on two datasets: (1) the balanced EMNIST [60] dataset with 47 class labels for image classification tasks, and (2) the Poker hand dataset [61] for multi-class classification tasks, which is widely applied in automatic rule generation. Each user possesses 1000 local training samples arranged into training batches with 64 samples per batch. The default number of participants in each simulation is $N = 25$, otherwise it is specified accordingly. The sampling interval is 500 events, i.e., the evaluation of each local model is done under a test set whenever the 500$^{\text{th}}$ event is finished. The rate parameter of the exponential distribution in local gradient computation is $\lambda_i = 0.1$ for all users by default. In this study, we did not evaluate the impact on model compression, implying that the packet size is as large as the raw model. The convolutional neural network (CNN) architecture used in the simulations takes up 596776 B (0.57 MB) for feeding samples from EMNIST, and 51640 B (0.05 MB) from Poker hand, respectively. These values are used to quantify the message size.



(a) Balanced EMNIST



(b) Poker hand

Fig. 4. Performance comparison with the literature under (a) EMNIST dataset, and (b) Poker hand dataset.

We performed simulations using cycle and complete topologies, with a time-invariant $\mathcal{Q}$. The connectivity graph is fixed throughout the whole collaboration process. Each user, indexed i without losing generality, spends some time computing the local gradient following $exp(\lambda_i)$ as mentioned in Assumption 1. Whenever a local update is done at t, user i sends $\Delta_t^{(i)}$ to its neighbors $j \in \mathcal{N}(i)$, where $\mathcal{N}(i)$ indicates a set of user i's neighbors. Although a pre-defined topology outlines the intended communication paths between nodes, the inherent unreliability of wireless channels can significantly affect data transmission. Factors such as fading, interference, and physical obstructions can disrupt connectivity, resulting in packet losses and delays, thereby undermining the efficiency and reliability of communication within the network. We used parameters reported in [62] and [63] for the wireless communication settings. The radius of the field where the nodes can be scattered is $R = 500$ m. We fix the transmit power of each user as $P_i = 30$ dBm (1000 mW). We also set the path loss exponent $\alpha = 4$, the bandwidth $W = 10$ MHz, and the noise power density $N_0 = -174$ dBm/Hz. We assumed that two nodes interfere with each other during transmission if their distance is closer than $0.1R$. Due to those wireless communication characteristics, the mechanism for realizing DRACO is slightly different. Specifically, when user i has performed local training at time t, it broadcasts its update $\Delta_t^{(i)}$ to all $j \in \mathcal{U} \setminus \{i\}$. It takes

$$\Gamma_{ij} = \frac{\text{message size}}{W \cdot \log_2(1 + \text{SINR}_{i,j})} + \frac{\text{distance}(i,j)}{\text{lightspeed}}$$

seconds for the message to arrive at node j. Here, the signal-to-interference-plus-noise ratio (SINR) between the two nodes is defined as

$$\text{SINR}_{i,j} = \frac{P_i h_{ji} \text{distance}(j,i)^{-\alpha}}{\sum_{n \in \Phi_j} P_i h_{jn} \text{distance}(j,n)^{-\alpha} + z^2},$$

where $h_{ji} \sim exp(1)$ denotes the small-scale fading gain, Φ_j is a set of nodes interfering node j, and z^2 characterizes the variance of AWGN (Additive White Gaussian Noise). As long as the transmission duration Γ_{ij} is shorter than the predetermined threshold $\Gamma_{\max}$, user j succeeds to receive $\Delta^{(i)}$ at time $t + \Gamma_{ij}$. (i.e., $q^{ij}_{t+\Gamma_{ij}} = 1$.)

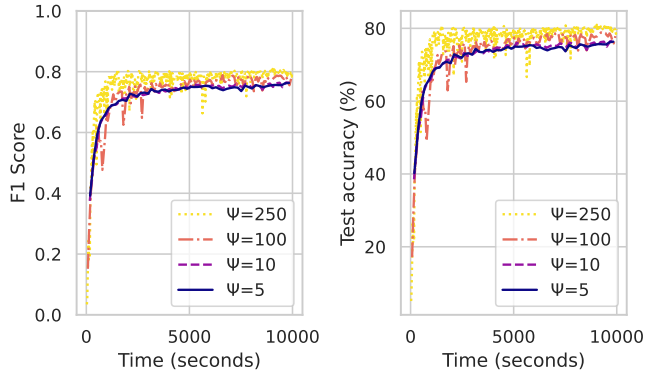
To simulate dynamic network conditions, we employed time-varying topologies utilizing the Gauss-Markov mobility model, a widely used approach for modeling the continuous movement of nodes. In this model, each node updates its velocity $v_i(t)$ and direction $\theta_i(t)$ at each timestep according to a weighted average of its previous velocity and direction, along with random perturbations. The update equations are given by:

$$v_i(t) = \alpha v_i(t-1) + (1-\alpha)\bar{v} + \sqrt{1-\alpha^2}\sigma_v w_{iv}(t)$$
$$\theta_i(t) = \alpha \theta_i(t-1) + (1-\alpha)\bar{\theta} + \sqrt{1-\alpha^2}\sigma_\theta w_{i\theta}(t)$$

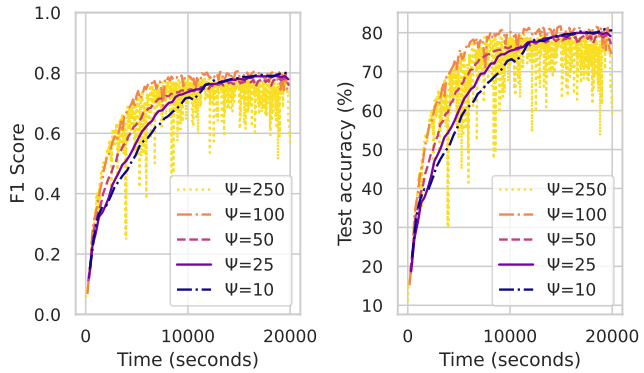
where $\alpha \in [0,1]$ is a tuning parameter that controls the degree of randomness, $\bar{v}$ and $\bar{\theta}$ represent the steady-state speed and direction, σ_v and σ_θ are the standard deviations for speed and direction, respectively, and $w_{iv}, w_{i\theta}$ are Gaussian noise terms. The value of α determines the dependency on previous states, with $\alpha = 0$ corresponding to a purely random walk, and $\alpha = 1$ corresponding to linear, deterministic motion. We set $\alpha = 0.5$ in our experiments to balance randomness and directed movement. The clients' physical coordinates (positions) were updated every $t_d = 1.0$ second, defining the interval between consecutive simulation timesteps. We modeled a scenario where mobile clients represent diverse types of entities, such as pedestrian smart devices or unmanned aerial vehicles (UAVs), by varying the maximum client speed between 1 m/s and 30 m/s.

The performance of DRACO is evaluated across different network topologies and datasets. Simulations with EMNIST employ a cycle topology, where each user is connected to two neighbors. In contrast, the Poker hand dataset employs a fully connected topology, where each user is directly connected to all others. DRACO's performance is compared against four benchmark methods:

- sync-symm: Synchronous learning with symmetric connectivity (Choco-SGD [64])
- sync-push: Synchronous learning with directed connectivity.
- async-symm: Asynchronous learning with symmetric connectivity (Decentralized Asynchronous SGD [15]).
- async-push: Asynchronous learning with directed connectivity (Digest [45]).



(a) EMNIST



(b) Poker hand

Fig. 5. Results for different upper bounds on the number of received messages per user. ($\Gamma_{\max} = 10$)

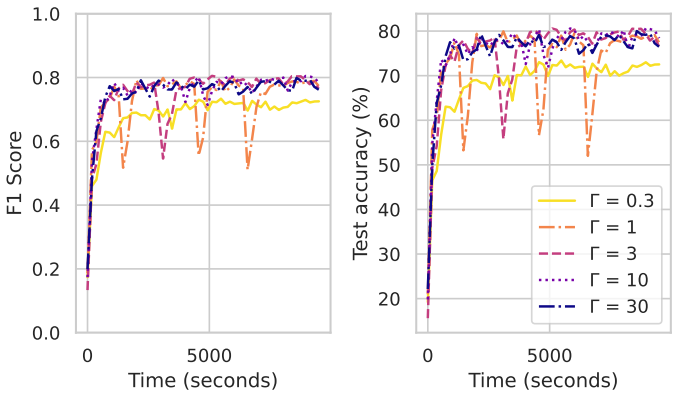


Fig. 6. Results for different transmission duration deadline (window$=0$, $P = 500$)

The term "Push" denotes the use of the push-sum algorithm for directed graphs.

The Poker hand dataset presents a unique challenge due to its imbalanced class distribution. To comprehensively assess model performance, we evaluated both test accuracy and F1-score, the latter accounting for both precision and recall.

While the choice of dataset had a minor impact on overall trends, the network topology significantly influenced performance. In the cycle topology, where each user exchanges information with only two neighbors, unreliable channels (e.g., due to fading) can lead to frequent client isolation. As shown in Fig. 4a, synchronous methods exhibited comparable performance, but async-symm underperformed async-push, despite using a doubly stochastic matrix. This performance gap highlights the sensitivity of async-symm to strict transmission deadlines, emphasizing the importance of well-designed scheduling in asynchronous learning. In the fully connected topology in Fig. 4b, where every user is connected to all others, the virtual global model can be trained more robustly, even when some edges are intermittently disrupted. While convergence speeds vary, all algorithms ultimately achieve similar performance. DRACO consistently outperformed competitors in both test accuracy and F1-score. This advantage stems from its parallel aggregation and unification mechanisms, which effectively mitigate the divergence of local models common in asynchronous decentralized learning. DRACO periodically unifies local reference models and regulates the number of received messages, enhancing robustness in continuous operation and fading environments.

To evaluate DRACO's communication and computation efficiency, we track the total number of transmissions and local training events per client throughout the entire learning process, and compare these metrics against several baseline algorithms (Table I). Notably, DRACO achieves the lowest overall number of transmissions per user, which is attributed to the Ψ term that effectively regulates the number of outbound transmission attempts. However, due to the uncoordinated nature of communication in DRACO, we observe a significant variance in transmission frequencies across clients, with some nodes transmitting nearly twice as many messages as others. This variability is also reflected in the number of local training events per client, stemming from DRACO's decoupled design between communication and computation phases. Despite this heterogeneity in individual client activity, the results presented in Fig. 4 show that DRACO achieved the highest test accuracy while simultaneously requiring fewer communication and computation resources compared to baseline algorithms from the literature, highlighting DRACO's resource efficiency without compromising performance.

During implementation, performance oscillations were observed when users received excessive redundant updates due to high transmission frequencies (large Ψ values in Fig. 5a and 5b). Conversely, excessively small Ψ values slowed learning by limiting the reception of crucial updates. These findings align with prior work [59] and the theoretical analysis presented in Theorem 1.

We vary the transmission deadline from 0.3 to 30 seconds in Fig. 6. As $\Gamma_{\max}$ increases, the number of communication events increases, as users are more willing to accept stale updates. This higher frequency of message exchanges generally accelerates convergence by providing more frequent opportunities for model updates and helps reduce spiking fluctuations in average performance. However, the improvement saturates around $\Gamma_{\max} = 10$, suggesting that the inclusion of excessively stale updates during model aggregation begins to degrade performance, ultimately leading to a longer overall training time.

Fig. 7 illustrates the performance of DRACO under non-

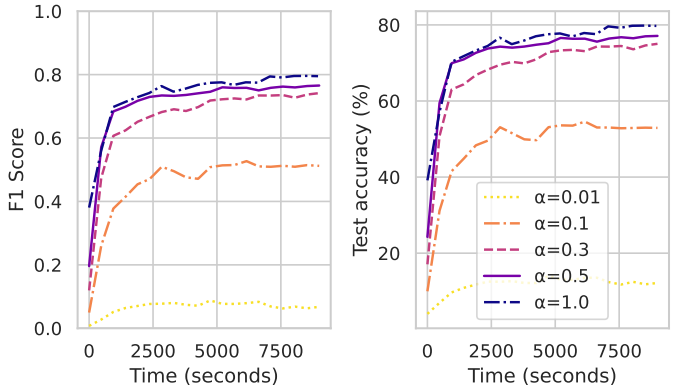
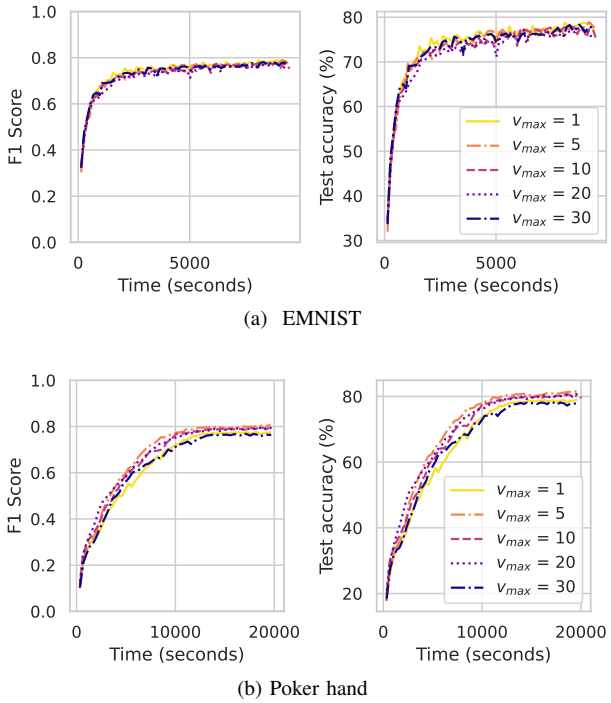


Fig. 7. Results for non-IID training data distribution with different Dirichlet coefficients (window= 0, $P = 500$, $\Psi = 100$)



(a) EMNIST

(b) Poker hand

Fig. 8. Performance across different maximum user speeds in mobile networks. ($t_d = 1.0$, $\Gamma_{\max} = 10$)

independently and identically distributed (non-IID) training data with varying Dirichlet distribution coefficients, denoted as α. Given the inherent high degree of non-IIDness in the Poker Hand dataset, where approximately 90% of the training samples belong to just two out of ten classes, we specifically manipulated the data heterogeneity using the EMNIST dataset for these experiments. To create diverse non-IID data distributions across clients, we set the Dirichlet coefficient α to values ranging from 0.01 to 1.0. Our results indicate that DRACO maintains robust accuracy for $\alpha \geq 0.3$. However, when α falls below 0.1, we observed a decline in performance for clients, evidenced by reductions in both test accuracy and F1-score. For reference, it is worth noting that a common setting in many federated learning studies utilizes a Dirichlet coefficient around $\alpha = 0.4$ when modeling standard non-IID environments.

To evaluate DRACO under dynamic network conditions, we conduct experiments using time-varying topologies generated with the Gauss-Markov mobility model (see Fig. 8). In this model, each client updates its speed and direction at each timestep based on a weighted combination of its previous state and random perturbations, resulting in smooth and realistic mobility patterns. The results demonstrate that DRACO maintains robust learning performance across various mobility scenarios. Both test accuracy and F1 score remain stable, even as the network topology evolves due to node movement. In high-mobility scenarios, particularly when the maximum speed exceeds 20 m/s, a slight increase in performance fluctuations over time is observed, which is expected due to more frequent link disruptions. Nevertheless, DRACO consistently outperforms baseline methods and demonstrates strong resilience in time-varying mobile wireless networks.

VI. CONCLUDING REMARKS AND FUTURE DIRECTIONS

We have studied decentralized asynchronous learning optimization through row-stochastic gossip communication networks and proposed a novel method called DRACO. Our technique facilitates the learning process, removing the need for global iteration counts. It presents local user performance defined on a continuous timeline. We provided practical instructions for each participant by decoupling training and transmission schedules, resulting in complete autonomy and simplified implementations in real-world applications. We analyzed the algorithm's convergence and provided experimental results that support the proposed framework's efficacy and feasibility.

In the remainder, we highlight some promising yet challenging directions that require further investigation.

- **Improve robustness against collisions.** Random access is known to have a higher probability of collision occurrence. However, it is cumbersome or impractical to predetermine the communication schedule because, while carrying out DRACO, the participants decide whether to transmit and/or train their local models without communication or agreement with the other users. Collision in a random access protocol can be alleviated by adapting classical approaches in wireless networks. These approaches include configuring a random backoff time after a collision for retransmission attempts, adopting collision detection mechanisms, or allowing clients to dynamically adjust the size of their messages or the transmission power. On the other hand, considering collisions from the resource allocation perspective, the system can assign different priority levels to clients based on factors such as their data urgency or historical collision rates.
- **Reception control** We manually selected the rate parameters for transmissions (λ_{ji}) because we assumed that the participants are not able to predict the frequency of message-receiving events, even in fixed $\mathcal{Q}$ cases. Nevertheless, there exist techniques that enable edge devices to

roughly estimate the ratio of successful message reception in advance. With this in mind, studying how to manage the reception events in realizing DRACO will be possible.

- **Bandwidth allocation.** In this paper, bandwidth is equally distributed to all users. If the users exchange their SINR information, as well as their weight updates, a bandwidth allocation algorithm can be added within the "for i" loop, as proposed in [63].
- **More realistic experiments with aggregation time threshold.** We can consider that each user has a predetermined threshold to aggregate its neighbors' local updates. The user can perform superposition on their local reference model only after the timeout occurs. For instance, each user j might have an upper bound on the number of $\Delta^{(i)}$'s that it can accept during its receiving period.
- **Enhance robustness against malicious users.** The current DRACO framework operates under the assumption of a network of honest clients collaboratively working toward a global consensus. A critical direction for future research is incorporating robust security mechanisms to detect and mitigate the impact of malicious participants. Strengthening DRACO's resilience against adversarial behavior would significantly broaden its applicability, enabling secure and reliable collaborative learning in fully decentralized environments. Robustness is crucial in settings characterized by asynchronous operation and flexible aggregation schedules, where vulnerabilities to malicious behavior are inherently higher. A more robust variant of DRACO would thus represent a key advancement toward trustworthy decentralized learning.

Future work could also include exploring how the system handles older or outdated updates, which could make it more reliable and efficient. We can also consider using different learning rates or adjustments across various devices, which could make the algorithm work better over a range of device capabilities. Additionally, one can adapt DRACO for more complex mobility scenarios, where distances and communication paths change over time in a three-dimensional space. This advance could make our approach more practical for real-world applications. Finally, the instructions can be simplified to make the procedure easier and more accessible in different settings. Addressing the abovementioned challenges could enhance DRACO's performance and usefulness, opening up new research avenues in asynchronous decentralized federated learning.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.), vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR, 20–22 Apr 2017.

[2] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[3] A. Lalitha, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *Third workshop on bayesian deep learning (NeurIPS)*, vol. 2, 2018.

[4] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "Braintorrent: A peer-to-peer environment for decentralized federated learning," 2019.

[5] A. Karras, C. Karras, K. C. Giotopoulos, D. Tsolis, K. Oikonomou, and S. Sioutas, "Peer to peer federated learning: Towards decentralized machine learning on edge devices," in *2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pp. 1–9, 2022.

[6] T. Qin, S. R. Etesami, and C. A. Uribe, "Decentralized federated learning for over-parameterized models," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5200–5205, 2022.

[7] G. Nadiradze, A. Sabour, P. Davies, S. Li, and D. Alistarh, "Asynchronous decentralized SGD with quantized and local updates," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 6829–6842, Curran Associates, Inc., 2021.

[8] R. Dai, L. Shen, F. He, X. Tian, and D. Tao, "DisPFL: Towards communication-efficient personalized federated learning via decentralized sparse training," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 4587–4604, PMLR, 17–23 Jul 2022.

[9] Y. Esfandiari, S. Y. Tan, Z. Jiang, A. Balu, E. Herron, C. Hegde, and S. Sarkar, "Cross-gradient aggregation for decentralized learning from non-iid data," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 3036–3046, PMLR, 18–24 Jul 2021.

[10] X. Liang, A. M. Javid, M. Skoglund, and S. Chatterjee, "Asynchrounous decentralized learning of a neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3947–3951, 2020.

[11] Y. Kanamori, Y. Yamasaki, S. Hosoai, H. Nakamura, and H. Takase, "An asynchronous federated learning focusing on updated models for decentralized systems with a practical framework," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1147–1154, 2023.

[12] E. T. Martínez Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.

[13] X. Zhang, X. Zhu, W. Bao, L. T. Yang, J. Wang, H. Yan, and H. Chen, "Distributed learning on mobile devices: A new approach to data mining in the internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10264–10279, 2021.

[14] I. Hegedűs, G. Danner, and M. Jelasity, "Decentralized learning works: An empirical comparison of gossip learning and federated learning," *Journal of Parallel and Distributed Computing*, vol. 148, pp. 109–124, 2021.

[15] E. Jeong, M. Zecchin, and M. Kountouris, "Asynchronous decentralized learning over unreliable wireless networks," in *ICC 2022 - IEEE International Conference on Communications*, pp. 607–612, 2022.

[16] L. Wulfert, N. Asadi, W.-Y. Chung, C. Wiede, and A. Grabmaier, "Adaptive decentralized federated gossip learning for resource-constrained iot devices," in *Proceedings of the 4th International Workshop on Distributed Machine Learning*, DistributedML '23, (New York, NY, USA), p. 27–33, Association for Computing Machinery, 2023.

[17] M. Even, A. Koloskova, and L. Massoulie, "Asynchronous SGD on graphs: a unified framework for asynchronous decentralized and federated optimization," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics* (S. Dasgupta, S. Mandt, and Y. Li, eds.), vol. 238 of *Proceedings of Machine Learning Research*, pp. 64–72, PMLR, 02–04 May 2024.

[18] M. Blot, D. Picard, M. Cord, and N. Thome, "Gossip training for deep learning," 2016.

[19] D. T. A. Nguyen, D. T. Nguyen, and A. Nedić, "Accelerated ab/push-pull methods for distributed optimization over time-varying directed networks," *IEEE Transactions on Control of Network Systems*, pp. 1–12, 2023.

[20] M. S. Assran and M. G. Rabbat, "Asynchronous gradient push," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 168–183, 2021.

[21] J. Zhang and K. You, "Asyspa: An exact asynchronous algorithm for convex optimization over digraphs," *IEEE Transactions on Automatic Control*, vol. 65, no. 6, pp. 2494–2509, 2020.

[22] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-

independent performance for strongly convex functions," *Journal of Machine Learning Research*, vol. 21, no. 58, pp. 1–47, 2020.

[23] R. Xin, C. Xi, and U. A. Khan, "FROST—fast row-stochastic optimization with uncoordinated step-sizes," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, pp. 1–14, 2019.

[24] H. Xing, O. Simeone, and S. Bi, "Federated learning over wireless device-to-device networks: Algorithms and convergence analysis," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3723–3741, 2021.

[25] S. Wang and M. Ji, "A unified analysis of federated learning with arbitrary client participation," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 19124–19137, Curran Associates, Inc., 2022.

[26] D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi, "Fedvarp: Tackling the variance due to partial client participation in federated learning," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* (J. Cussens and K. Zhang, eds.), vol. 180 of *Proceedings of Machine Learning Research*, pp. 906–916, PMLR, 01–05 Aug 2022.

[27] B. Li, M. N. Schmidt, T. S. Alstrøm, and S. U. Stich, "On the effectiveness of partial variance reduction in federated learning with heterogeneous data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3964–3973, June 2023.

[28] T. Qin, J. Yevale, and S. R. Etesami, "Communication-efficient local sgd for over-parametrized models with partial participation," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 2098–2103, 2023.

[29] Y. J. Cho, P. Sharma, G. Joshi, Z. Xu, S. Kale, and T. Zhang, "On the convergence of federated averaging with cyclic client participation," in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 5677–5721, PMLR, 23–29 Jul 2023.

[30] M. Even, H. Hendrikx, and L. Massoulié, "Asynchronous speedup in decentralized optimization," in *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

[31] M. G. Rabbat and K. I. Tsianos, "Asynchronous decentralized optimization in heterogeneous systems," in *53rd IEEE Conference on Decision and Control*, pp. 1125–1130, 2014.

[32] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 5453–5458, 2012.

[33] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 9324–9333, PMLR, 13–18 Jul 2020.

[34] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "Push-saga: A decentralized stochastic algorithm with variance reduction over directed graphs," *IEEE Control Systems Letters*, vol. 6, pp. 1202–1207, 2022.

[35] M. T. Toghani, S. Lee, and C. A. Uribe, "PARS-push: Personalized, asynchronous and robust decentralized optimization," *IEEE Control Systems Letters*, vol. 7, pp. 361–366, 2023.

[36] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 344–353, PMLR, 09–15 Jun 2019.

[37] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.

[38] Y.-G. Hsieh, Y. Laguel, F. Iutzeler, and J. Malick, "Push–pull with device sampling," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 7179–7194, 2023.

[39] X. Mao, K. Yuan, Y. Hu, Y. Gu, A. H. Sayed, and W. Yin, "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2513–2528, 2020.

[40] G. Ayache and S. El Rouayheb, "Random walk gradient descent for decentralized learning on graphs," in *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 926–931, 2019.

[41] H. Hendrikx, "A principled framework for the design and analysis of token algorithms," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.), vol. 206 of *Proceedings of Machine Learning Research*, pp. 470–489, PMLR, 25–27 Apr 2023.

[42] A. Nedić, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[43] D. Ghaderyan, N. S. Aybat, A. P. Aguiar, and F. L. Pereira, "A fast row-stochastic decentralized method for distributed optimization over directed graphs," *IEEE Transactions on Automatic Control*, vol. 69, no. 1, pp. 275–289, 2024.

[44] L. Giaretta and v. Girdzijauskas, "Gossip learning: Off the beaten path," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1117–1124, 2019.

[45] P. Gholami and H. Seferoglu, "Digest: Fast and communication efficient decentralized learning with local updates," *IEEE Transactions on Machine Learning in Communications and Networking*, pp. 1–1, 2024.

[46] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.

[47] M. Akbari, B. Gharesifard, and T. Linder, "Distributed online convex optimization on time-varying directed graphs," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 3, pp. 417–428, 2017.

[48] Z. Li, Z. Ding, J. Sun, and Z. Li, "Distributed adaptive convex optimization on directed graphs via continuous-time algorithms," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1434–1441, 2018.

[49] V. S. Mai and E. H. Abed, "Distributed optimization over weighted directed graphs using row stochastic matrix," in *2016 American Control Conference (ACC)*, pp. 7165–7170, 2016.

[50] D. T. A. Nguyen, S. Wang, D. T. Nguyen, A. Nedich, and H. V. Poor, "Decentralized federated learning with gradient tracking over time-varying directed networks," 2024.

[51] E. Jeong and M. Kountouris, "Personalized decentralized federated learning with knowledge distillation," in *ICC 2023 - IEEE International Conference on Communications*, pp. 1982–1987, 2023.

[52] Y. Liu, Y. Shi, Q. Li, B. Wu, X. Wang, and L. Shen, "Decentralized directed collaboration for personalized federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23168–23178, June 2024.

[53] A. Nabli and E. Oyallon, "DADAO: Decoupled accelerated decentralized asynchronous optimization," in *International Conference on Machine Learning*, pp. 25604–25626, PMLR, 2023.

[54] A. Nabli, E. Belilovsky, and E. Oyallon, "$\mathbf{A}^2\mathbf{CiD}^2$: Accelerating asynchronous communication in decentralized deep learning," in *Advances in Neural Information Processing Systems* (A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 47451–47474, Curran Associates, Inc., 2023.

[55] E. Belilovsky, M. Eickenberg, and E. Oyallon, "Decoupled greedy learning of CNNs," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 736–745, PMLR, 13–18 Jul 2020.

[56] J. Jiang, W. Zhang, J. GU, and W. Zhu, "Asynchronous decentralized online learning," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 20185–20196, Curran Associates, Inc., 2021.

[57] J. Postel, "Rfc0793: Transmission control protocol," 1981.

[58] J. F. C. Kingman, *Poisson processes*, vol. 3. Clarendon Press, 1992.

[59] A. Hashemi, A. Acharya, R. Das, H. Vikalo, S. Sanghavi, and I. Dhillon, "On the benefits of multiple gossip steps in communication-constrained decentralized optimization," 2020.

[60] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: an extension of mnist to handwritten letters," 2017.

[61] R. Cattral and F. Oppacher, "Poker Hand." UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C5KW38.

[62] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5136–5151, 2021.

[63] H. Xie, M. Xia, P. Wu, S. Wang, and K. Huang, "Decentralized federated learning with asynchronous parameter sharing for large-scale iot networks," *IEEE Internet of Things Journal*, pp. 1–1, 2024.

[64] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487, PMLR, 2019.

[65] M. Karnaugh, "The map method for synthesis of combinational logic circuits," *Transactions of the American Institute of Electrical Engineers,*

Part I: Communication and Electronics, vol. 72, no. 5, pp. 593–599, 1953.

APPENDIX

A. Preliminaries

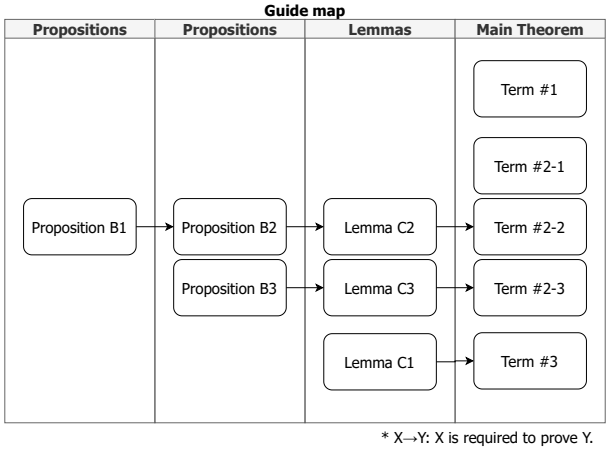


Fig. A.9. A metaphoric map that guides the correlation of each proposition and lemma for proving the main theorem.

Before the proof of Theorem 1, it is essential to verify (i) how many communication events and (ii) how many local gradient updates occur during P. In PPP, communication events occur $\lambda_i P$ times on average during P, which indicates the expectation of broadcasting frequency of node i. To find the bound for $\mathbf{x}_{t_0+P} - \mathbf{x}_{t_0}$, we need to specify how many reception events happen in a random node i during the elapsed time of P. For simplicity, we write $\int_P$ to indicate $\int_{t_0}^{t_0+P}$. The reference model of node i update during P is

$$\begin{aligned}\mathbf{x}_{t_0+P}^{(i)} - \mathbf{x}_{t_0}^{(i)} &= \int_P \sum_j \Pr[i \in \mathcal{N}_t(j)]\Delta_t^{(j)}\,\mathrm{d}t \\ &= \int_P \sum_j q_t^{ji}\Delta_t^{(j)}\,\mathrm{d}t \\ &= \gamma \int_P \sum_j q_t^{ji} \sum_{b=0}^{B-1} \mathbf{g}_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\,,\end{aligned}$$

which is heterogeneous across nodes. A floored notation $\lfloor t\rfloor$ indicates the latest moment no later than time t that user j computes $\Delta^{(j)}$.

A superscripted or subscripted $\star$ on some variables is analogous to a "don't-care" (DC) term in digital logic [65]. For instance, $q_\star$ is the same as any q_i, where i can be any user index in $\mathcal{U}$ without loss of generality.

B. Propositions

Proposition B.1. *If Assumption 5 is satisfied, a decentralized learning network with $N \geq 4$ clients satisfies*

$$\sum_{j\neq i} q_t^{ji}\left\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \leq \frac{9N\zeta^2}{4}$$

for all $i, j \in \mathcal{U}$ and $t \in [0, T)$.

Proof.

$$\begin{aligned}&\sum_{j\neq i} q_t^{ji}\left\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \\ &= \sum_{j\neq i} q_t^{ji}\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_t^{(i)}) + \nabla f(\mathbf{x}_t)\|^2 \\ &\overset{(a)}{\leq} \left(1+\frac{1}{\sqrt{N}}\right)\sum_{j\neq i} q_t^{ji}\left\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f(\mathbf{x}_t)\right\|^2 \\ &\quad + (\sqrt{N}+1)\sum_{j\neq i} q_t^{ji}\left\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\mathbf{x}_t)\right\|^2 \\ &\overset{(b)}{\leq} \left(1+\frac{1}{\sqrt{N}}\right)\sum_{j\neq i} q_t^{ji}\left\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f(\mathbf{x}_t)\right\|^2 \\ &\quad + (\sqrt{N}+1)\zeta^2 \\ &\overset{(c)}{\leq} (N+2\sqrt{N}+1)\zeta^2 \quad \overset{(d)}{\leq} \frac{9N\zeta^2}{4},\end{aligned}$$

where (a) is due to Young's inequality; (b) comes from Assumption 5; (c) takes the fact that $q_\star^\star \leq 1$ for any user nodes; (d) is always true for $N \geq 4$ since $\frac{5N}{4} - 2\sqrt{N} - 1 \geq 0$ for any $\sqrt{N} \geq 2$, which satisfies the given condition about N. □

Remark. This proposition appears in the proof of Proposition B.2.

Proposition B.2. *(Upper bound for superpositioned model deviations.) Let $h_j(b) = \mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}$ denote the difference between a local model calculated by feeding each batch with an index b and the local reference model. For all $i, j \in \mathcal{U}$, when $\gamma \leq \frac{1}{8BL}$, we have*

$$\begin{aligned}&\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\right] \\ &\leq \frac{2}{5}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\right\|^2\right] \\ &\quad + \frac{9N\zeta^2}{10L^2} + \frac{16B\gamma^2\sigma^2}{5} + \frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\,.\end{aligned}$$

Proof. We rephrase the $b+1^{\text{th}}$ term, $h_j(b+1) = \mathbf{y}_{t,b+1}^{(j)} - \mathbf{x}_t^{(j)}$, as follows.

$$\begin{aligned}&\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{y}_{t,b+1}^{(j)} - \mathbf{x}_t^{(j)}\right\|^2\right] \\ &= \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma g_j(\mathbf{y}_{t,b}^{(j)})\right\|^2\right] \\ &\overset{(a)}{=} \sum_{j\neq i} q_t^{ji}\left\|\mathbb{E}_{\cdot|\mathcal{Q}}\left[\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma g_j(\mathbf{y}_{t,b}^{(j)})\right]\right\|^2 \\ &\quad + \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma g_j(\mathbf{y}_{t,b}^{(j)}) \\ &\qquad\qquad - \mathbb{E}_{\cdot|\mathcal{Q}}\left[\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma g_j(\mathbf{y}_{t,b}^{(j)})\right]\Big\|^2\Big] \\ &= \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma\nabla f_j(\mathbf{y}_{t,b}^{(j)})\right\|^2\right] \\ &\quad + \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\left[\left\|\gamma\big(g_j(\mathbf{y}_{t,b}^{(j)}) - \nabla f_j(\mathbf{y}_{t,b}^{(j)})\big)\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} - \gamma\nabla f_j(\mathbf{y}_{t,b}^{(j)})\Big\|^2\Big] + \gamma^2\sigma^2 \\
&= \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)} + \gamma\nabla f_j(\mathbf{x}_t^{(j)}) \\
&\qquad\qquad - \gamma\nabla f_j(\mathbf{y}_{t,b}^{(j)}) - \gamma\nabla f_j(\mathbf{x}_t^{(j)}) - \gamma\nabla f_i(\mathbf{x}_t^{(i)}) \\
&\qquad\qquad + \gamma\nabla f_i(\mathbf{x}_t^{(i)}) - \gamma\nabla f_i(\mathbf{x}_{t_0}^{(i)}) + \gamma\nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big\|^2\Big] \\
&\quad + \gamma^2\sigma^2 \\
&\overset{(c)}{\leq} \Big(1 + \frac{1}{2B-1}\Big)\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big] \\
&\quad + 2B\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\nabla f_j(\mathbf{y}_{t,b}^{(j)}) - \nabla f_j(\mathbf{x}_t^{(j)}) \\
&\qquad\qquad + \nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f_i(\mathbf{x}_t^{(i)}) \\
&\qquad\qquad + \nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\mathbf{x}_{t_0}^{(i)}) \\
&\qquad\qquad + \nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big\|^2\Big] + \gamma^2\sigma^2 \\
&\leq \Big(1 + \frac{1}{2B-1}\Big)\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big] \\
&\quad + 8B\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\nabla f_j(\mathbf{y}_{t,b}^{(j)}) - \nabla f_j(\mathbf{x}_t^{(j)})\Big\|^2\Big] \\
&\quad + 8B\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_j(\mathbf{x}_t^{(j)}) - \nabla f_i(\mathbf{x}_t^{(i)})\big\|^2\big] \\
&\quad + 8B\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \\
&\quad + 8B\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] + \gamma^2\sigma^2 \\
&\overset{(d)}{\leq} \Big(1 + \frac{1}{2B-1}\Big)\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big] \\
&\quad + 8BL^2\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big] + 18BN\gamma^2\zeta^2 \\
&\quad + 8BL^2\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&\quad + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] + \gamma^2\sigma^2 \\
&= \big(1 + 8BL^2\gamma^2 + \frac{1}{2B-1}\big)\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] \\
&\quad + 18BN\gamma^2\zeta^2 + \gamma^2\sigma^2 \\
&\quad + 8BL^2\gamma^2\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\|^2] \\
&\quad + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&\overset{(e)}{\leq} \left(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\right)\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] \\
&\quad + \frac{9N\zeta^2}{32BL^2} + \gamma^2\sigma^2 + \frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&\quad + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \qquad \text{(B.9)}
\end{aligned}
$$

where (a) is from the definition of variance; (b) is derived from the definition of σ in Assumption 4; (c) follows from Young's inequality; and (d) uses L-smoothness on the second and the fourth term, and applies Proposition B.1 on the third term. Afterwards, the first two terms are integrated; (e) is derived from the fact that $\gamma^2 \leq \frac{1}{64B^2L^2}$ and that

$$
\begin{aligned}
8BL^2\gamma^2 + \frac{1}{2B-1} &\leq \frac{1}{8B} + \frac{1}{2B-1} \leq \frac{1}{8B-4} + \frac{1}{2B-1} \\
&= \frac{5}{8\big(B-\frac{1}{2}\big)}.
\end{aligned}
$$

Let $H(b)$ indicate $\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big]$. From the last line of inequality B.9, we have

$$
\begin{aligned}
H(b+1) \leq{}& \Big(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\Big)H(b) \\
&+ \frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&+ 18BN\gamma^2\zeta^2 + \gamma^2\sigma^2 \\
&+ 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\ . \qquad \text{(B.10)}
\end{aligned}
$$

Since $\mathbf{y}_{t,0}^{(j)} = \mathbf{x}_t^{(j)}$ for all $t,\ j$ based on Algorithm 1,

$$
H(0) = \sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,0}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] = 0.
$$

Recurring inequality B.10 from $H(0)$, we can get

$$
\begin{aligned}
H(b) \leq{}& \Big(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\Big)^b H(0) + \sum_{b'=0}^{b-1}\Big(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\Big)^{b'} \\
&\cdot \Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&+ \frac{9N\zeta^2}{32BL^2} + \gamma^2\sigma^2 + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big) \\
\leq{}& \Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&+ \frac{9N\zeta^2}{32BL^2} + \gamma^2\sigma^2 + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big) \\
&\cdot \sum_{b=0}^{B-1}\left(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\right)^b \\
={}& \left[\left(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\right)^B - 1\right]\cdot\frac{8\big(B-\frac{1}{2}\big)}{5} \\
&\cdot \Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&+ \frac{9N\zeta^2}{32BL^2} + \gamma^2\sigma^2 + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big) \\
={}& \left[\left(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\right)^{B-\frac{1}{2}}\left(1 + \frac{5}{8\big(B-\frac{1}{2}\big)}\right)^{\frac{1}{2}} - 1\right] \\
&\cdot \frac{8\big(B-\frac{1}{2}\big)}{5}\cdot\Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] \\
&+ \frac{9N\zeta^2}{32BL^2} + \gamma^2\sigma^2 + 8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big) \\
\overset{(a)}{\leq}{}& \left[e^{\frac{5}{8}}\cdot\frac{3}{2} - 1\right]\cdot\frac{8\big(B-\frac{1}{2}\big)}{5}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)}-\mathbf{x}_{t_0}^{(i)}\big\|^2\big]+\frac{9N\zeta^2}{32BL^2}\\
&+\gamma^2\sigma^2+8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big)\\
&\overset{(b)}{\leq}\frac{16}{5}B\Big(\frac{1}{8B}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)}-\mathbf{x}_{t_0}^{(i)}\big\|^2\big]+\frac{9N\zeta^2}{32BL^2}\\
&+\gamma^2\sigma^2+8B\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big)\\
&=\frac{2}{5}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)}-\mathbf{x}_{t_0}^{(i)}\big\|^2\big]+\frac{9N\zeta^2}{10L^2}\\
&+\frac{16B\gamma^2\sigma^2}{5}+\frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]
\end{aligned}
$$

where (a) comes from $(1+x)^{1/x}\leq e$ and $B\geq 1$, which results in $\Big(1+\frac{5}{8(B-(1/2))}\Big)^{1/2}\leq\big(1+\frac{5}{4}\big)^{1/2}=\frac{3}{2}$; (b) is due to $\frac{3}{2}e^{\frac{5}{8}}-1\leq 2$ and $B-\frac{1}{2}\leq B$. □

Remark. This proposition appears in the proof of Lemma C.3, which is used at the first term of inequality D.18.

Proposition B.3. *(Upper bound for the local reference model change) When Assumption 2 holds and* $\gamma\leq\min(\frac{1}{8BL},\frac{1}{8BLN\Psi})$*, we have*

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\|\mathbf{y}_{t,b}^{(j)}-\mathbf{x}_t^{(j)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&\quad+\frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]+\frac{3\sigma^2}{16L^2}\ ,
\end{aligned}
$$

for all $i,j\in\mathcal{U}$ *and for* $t\in[t_0,t_0+P)$.

Proof. Here, we use $\int_\tau$ to replace $\int_{\tau=t_0}^{t}$ for simplicity of writing. We can rephrase the left side of the inequality as below by bringing Appendix A.

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&=\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1} g_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&=\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\\
&+\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\big[g_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\big]\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&\overset{(i)}{\leq}\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\big[g_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\big]\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&\overset{(a)}{\leq}\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\psi_j(t_0,t)\\
&\quad\int_\tau\Big\|\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\big[g_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\big]\Big\|^2\,\mathrm{d}\tau\Bigg]\\
&\leq\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)N\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\psi_j(t_0,t)\int_\tau\sum_{n\neq j} q_\tau^{nj}\\
&\quad\Big\|\sum_{b=0}^{B-1}\big[g_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\big]\Big\|^2\,\mathrm{d}\tau\Bigg]\\
&\leq\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)BN\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\psi_j(t_0,t)\int_\tau\sum_{n\neq j} q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\big\|g_n(\mathbf{y}_{\tau,\star}^{(n)})-\nabla f_n(\mathbf{y}_{\tau,\star}^{(n)})\big\|^2\,\mathrm{d}\tau\Bigg]\\
&\overset{(b)}{\leq}\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\,\mathrm{d}\tau\Big\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&\leq\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\Big[\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})\\
&\quad-\nabla f_n(\mathbf{x}_\tau^{(n)})+\nabla f_n(\mathbf{x}_\tau^{(n)})-\nabla f_j(\mathbf{x}_\tau^{(j)})+\nabla f_j(\mathbf{x}_\tau^{(j)})\\
&\quad-\nabla f_j(\mathbf{x}_{t_0}^{(j)})+\nabla f_j(\mathbf{x}_{t_0}^{(j)})-\nabla f_i(\mathbf{x}_{t_0}^{(i)})\\
&\quad+\nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big]\,\mathrm{d}\tau\Big\|^2\Bigg]+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&=\mu_1\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i} q_t^{ji}\cdot\\
&\quad\Big\|\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\Big[\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{x}_\tau^{(n)})\Big]\,\mathrm{d}\tau\\
&\quad+\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\Big[\nabla f_n(\mathbf{x}_\tau^{(n)})-\nabla f_j(\mathbf{x}_\tau^{(j)})\Big]\,\mathrm{d}\tau\\
&\quad+\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\Big[\nabla f_j(\mathbf{x}_\tau^{(j)})-\nabla f_j(\mathbf{x}_{t_0}^{(j)})\Big]\,\mathrm{d}\tau\\
&\quad+\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\Big[\nabla f_j(\mathbf{x}_{t_0}^{(j)})-\nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big]\,\mathrm{d}\tau\\
&\quad+\gamma\int_\tau\sum_{n\neq j} q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_i(\mathbf{x}_{t_0}^{(i)})\,\mathrm{d}\tau\Big\|^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&\overset{(ii,c)}{\leq} 4\mu_1\mu_2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\Big[\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{x}_\tau^{(n)})\Big]\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+4\mu_1\mu_2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\Big[\nabla f_n(\mathbf{x}_\tau^{(n)})-\nabla f_j(\mathbf{x}_\tau^{(j)})\Big]\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+4\mu_1\mu_2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\Big[\nabla f_j(\mathbf{x}_\tau^{(j)})-\nabla f_j(\mathbf{x}_{t_0}^{(j)})\Big]\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+4\mu_1\mu_2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\Big[\nabla f_j(\mathbf{x}_{t_0}^{(j)})-\nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big]\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)\\
&\quad\cdot\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\sum_{b=0}^{B-1}\nabla f_i(\mathbf{x}_{t_0}^{(i)})\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&\leq 4\mu_1\mu_2BN\gamma^2\psi_j(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\sum_{b=0}^{B-1}\Big\|\nabla f_n(\mathbf{y}_{\tau,b}^{(n)})-\nabla f_n(\mathbf{x}_\tau^{(n)})\Big\|^2\,\mathrm{d}\tau\Bigg]\\
&+4\mu_1\mu_2B^2\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\Big\|\psi_j(t_0,t)\sum_{n\neq j}q_\star^{nj}\\
&\quad\big[\nabla f_n(\mathbf{x}_\star^{(n)})-\nabla f_j(\mathbf{x}_\star^{(j)})\big]\Big\|^2\Big]\\
&+4\mu_1\mu_2B^2\gamma^2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Bigg\|\int_\tau\sum_{n\neq j}q_\tau^{nj}\\
&\quad\Big[\nabla f_j(\mathbf{x}_\tau^{(j)})-\nabla f_j(\mathbf{x}_{t_0}^{(j)})\Big]\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+4\mu_1\mu_2\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\Big\|\nabla f_j(\mathbf{x}_{t_0}^{(j)})-\nabla f_i(\mathbf{x}_{t_0}^{(i)})\Big\|^2\\
&\quad\cdot\Bigg\|\gamma\int_\tau\sum_{n\neq j}q_\tau^{nj}\sum_{b=0}^{B-1}1\,\mathrm{d}\tau\Bigg\|^2\Bigg]\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)B^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&\overset{(d)}{\leq} 4\mu_1\mu_2BL^2N\gamma^2\psi_j(t_0,t)\\
&\quad\cdot\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\int_\tau\sum_{n\neq j}q_\tau^{nj}\sum_{b=0}^{B-1}\big\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\big\|^2\,\mathrm{d}\tau\Big]\\
&+4\mu_1\mu_2B^2\gamma^2\psi_j^2(t_0,t)\cdot\frac{2N\zeta^2}{N-4}\\
&+4\mu_1\mu_2B^2L^2\gamma^2\\
&\quad\cdot\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\Big\|\int_\tau\sum_{n\neq j}q_\tau^{nj}[\mathbf{x}_\tau^{(j)}-\mathbf{x}_{t_0}^{(j)}]\,\mathrm{d}\tau\Big\|^2\Big]\\
&+\frac{8\mu_1\mu_2B^2N\gamma^2\psi_j^2(t_0,t)\zeta^2}{N-4}\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)B^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\\
&\overset{(e)}{\leq} 4\mu_1\mu_2BL^2N\gamma^2\psi_j(t_0,t)\\
&\quad\cdot\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\int_\tau\sum_{n\neq j}q_\tau^{nj}\sum_{b=0}^{B-1}\big\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\big\|^2\,\mathrm{d}\tau\Big]\\
&+4\mu_1\mu_2B^2L^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\big\|\mathbf{x}_{\tau_{\max}}^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&+\frac{16\mu_1\mu_2B^2N\gamma^2\psi_j^2(t_0,t)\zeta^2}{N-4}\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)B^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2\ ,
\end{aligned}
\tag{B.11}
$$

where two coefficients larger than one, denoted by μ_1 and μ_2, are introduced in (i) and (ii), respectively. In inequality B.11, (a) uses $\big\|\int_{\tau=t_0}^t\sum_{n\neq j}q_\tau^{nj}\mathbf{z}_\tau\,\mathrm{d}\tau\big\|^2 \leq \psi_j(t_0,t)\int_{\tau=t_0}^t\|\sum_{n\neq j}q_\tau^{nj}\mathbf{z}_\tau\|^2\,\mathrm{d}\tau$ and $\big\|\sum_{m=1}^M\mathbf{z}_m\big\|^2 \leq M\sum_{m=1}^M\|\mathbf{z}_m\|^2$ for any vector $\mathbf{z}_\star\in\mathbb{R}^d$.[1] (b) comes from the definition of σ^2 in Assumption 4. In (c), Jensen's inequality is applied once again. (d) takes L-smoothness on the first and the second term, while Lemma 1 is applied on the third term. In (e), the third term of inequality B.11 already contains the current lemma. Here, we introduce an index of the instant $\tau_{\max}\in[t_0,t)$ that satisfies $\tau_{\max}=\operatorname{argmax}_\tau\|\mathbf{x}_\tau^{(j)}-\mathbf{x}_{t_0}^{(j)}\|^2$.

The first term of inequality B.11, which includes Lemma B.2, can be rephrased as follows:

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\int_\tau\sum_{n\neq j}q_\tau^{nj}\sum_{b=0}^{B-1}\big\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\big\|^2\,\mathrm{d}\tau\Bigg]\\
&\leq B\mathbb{E}_{\cdot|\mathcal{Q}}\Bigg[\sum_{j\neq i}q_t^{ji}\int_\tau\sum_{n\neq j}q_\tau^{nj}\big\|\mathbf{y}_{\tau,\star}^{(n)}-\mathbf{x}_\tau^{(n)}\big\|^2\,\mathrm{d}\tau\Bigg]\\
&\leq B\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i}q_t^{ji}\psi_j(t_0,t)\sum_{n\neq j}q_\star^{nj}\big\|\mathbf{y}_{\star,\star}^{(n)}-\mathbf{x}_\star^{(n)}\big\|^2\Big]
\end{aligned}
$$

[1]This results in $\big\|\sum_{j\in\mathcal{U}}q_\star^{ji}\mathbf{z}_j\big\|^2 \leq N\sum_{j\in\mathcal{U}}q_\star^{ji}\|\mathbf{z}_j\|^2$ and $\|\sum_{b=0}^{B-1}\mathbf{z}_b\|^2\leq B\sum_{b=0}^{B-1}\|\mathbf{z}_b\|^2$.

$$\leq B\psi_j(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big] \tag{B.12}$$

We continue rephrasing the primary inequality B.11:

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq 4\mu_1\mu_2B^2L^2N\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]\\
&+4\mu_1\mu_2B^2L^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_{\tau_{\max}}^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&+\frac{16\mu_1\mu_2B^2N\gamma^2\psi_j^2(t_0,t)\zeta^2}{N-4}\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)B^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\psi_j^2(t_0,t)\sigma^2.
\end{aligned}$$

After rearranging the inequality in order to integrate those terms including $\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_\star^{ji}\big\|\mathbf{x}_\star^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]$, we have

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&-4\mu_1\mu_2B^2L^2\gamma^2\psi_j^2(t_0,t)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_{\tau_{\max}}^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq(1-4\mu_1\mu_2B^2L^2\gamma^2\Psi^2)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq(1-4\mu_1\mu_2B^2L^2N\gamma^2\Psi^2)\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big].
\end{aligned}$$

Hence, we can rephrase the inequality as

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq\frac{1}{1-4\mu_1\mu_2B^2L^2N\gamma^2\Psi^2}\\
&\cdot\Bigg[4\mu_1\mu_2B^2L^2N\gamma^2\Psi^2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]\\
&+\frac{16\mu_1\mu_2B^2N\gamma^2\zeta^2\Psi^2}{N-4}\\
&+\mu_1\Big(1+\frac{1}{\mu_2-1}\Big)B^2\gamma^2\Psi^2\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\Big(1+\frac{1}{\mu_1-1}\Big)B^2N^2\gamma^2\sigma^2\Psi^2\Bigg]\\
&\overset{(i)}{\leq}3\cdot\Bigg[\frac{2}{3}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]+\frac{8\zeta^2}{3L^2(N-4)}\\
&+\frac{B\gamma\Psi}{6LN(1-BL\gamma\Psi)}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\frac{B^2N^2\gamma^2\sigma^2\Psi^2}{1-6BLN\gamma\Psi}\Bigg]\\
&=2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&+\frac{B\gamma\Psi}{2LN(1-BL\gamma\Psi)}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\frac{3B^2N^2\gamma^2\sigma^2\Psi^2}{1-6BLN\gamma\Psi}\ .
\end{aligned}\tag{B.13}$$

where (i) $\mu_1=\frac{1}{6BLN\gamma\Psi}$ and $c=\frac{1}{BL\gamma\Psi}$ are applied. Additionally, if $\Psi>0$, $\gamma\leq\frac{1}{8BLN\Psi}$ is the tighter upper bound than $\gamma\leq\frac{1}{8BL}$. With this remark, the upper bound in inequality B.13 can be simplified even more as

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)}-\mathbf{x}_{t_0}^{(j)}\big\|^2\Big]\\
&\leq 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&+\frac{\frac{1}{8LN}}{2LN(1-\frac{1}{8N})}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\\
&+\frac{3B^2N^2\sigma^2\Psi^2\cdot\frac{1}{64B^2L^2N^2\Psi^2}}{1-\frac{6BLN\Psi}{8BLN\Psi}}\\
&\leq 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&+\frac{1}{2L^2N(8N-1)}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]+\frac{3\sigma^2}{16L^2}\\
&\leq 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{n\neq j} q_\tau^{nj}\|\mathbf{y}_{\tau,b}^{(n)}-\mathbf{x}_\tau^{(n)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&+\frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]+\frac{3\sigma^2}{16L^2}\\
&\overset{(a)}{\leq} 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\|\mathbf{y}_{\tau,b}^{(j)}-\mathbf{x}_\tau^{(j)}\|^2\Big]+\frac{8\zeta^2}{L^2(N-4)}\\
&+\frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]+\frac{3\sigma^2}{16L^2}\ ,
\end{aligned}\tag{B.14}$$

where (a) is satisfied without loss of generality. □

Remark. This proposition appears in Lemma C.4, which is then used at the second term of inequality D.18.

C. Lemmas

In collaborative learning, local computations often occur more frequently than communication. This frequency difference avoids duplicating transmissions, which can happen in the reverse scenario.

Lemma C.2. *For all $n\in\mathcal{U}$, there are no fewer $\mathbf{y}_{t,b}^{(n)}$ than $\mathbf{y}_{\lfloor t\rfloor,b}^{(n)}$ within any given range of time $\{t|t\in[t_0,t_0+P)\}$. In other words, it also satisfies that*

$$\Big\|\int_P\sum_{b=0}^{B-1}\mathbf{g}_n(\mathbf{y}_{\lfloor t\rfloor,b}^{(n)})\,\mathrm{d}t\Big\|^2\leq\Big\|\int_P\sum_{b=0}^{B-1}\mathbf{g}_n(\mathbf{y}_{t,b}^{(n)})\,\mathrm{d}t\Big\|^2. \tag{C.15}$$

Proof. In Fig. (C.10), the value of $\Delta_{t_4}^{(j)}$ can differ from $\Delta_{t_3}^{(j)}$ if another node transmits a message to node j, thereby affecting the value of $\mathbf{x}^{(j)}$. To facilitate our analysis, we assume that each user creates a backup of the non-transmitted local updates for the upcoming transmission event. Returning to the scenario depicted in Fig. (C.10), based on this assumption, user j sends

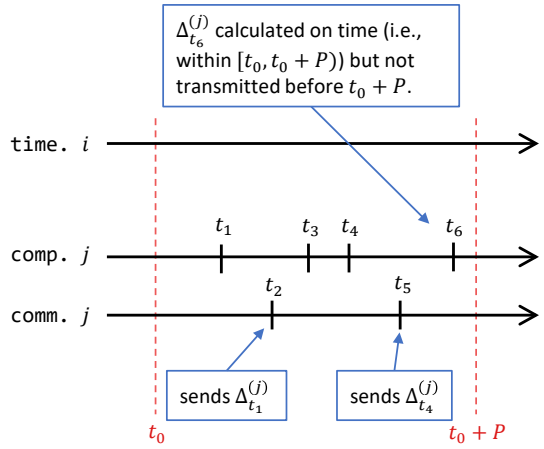


Fig. C.10. The latest local update of j is unable to be transmitted within the given range $[t_0, t_0+P)$ because of the independence between computation timestamps and communication (transmission) timestamps.

both $\Delta_{t_3}^{(j)}$ and $\Delta_{t_4}^{(j)}$ to user i at the earliest transmission event time, which is t_5. □

Lemma C.3. *(Upper bound for superpositioned model deviations.) For all $i, j \in \mathcal{U}$, when $\gamma \leq \min(\frac{1}{8BL}, \frac{1}{8BLN\Psi})$, we have*

$$\begin{aligned}
&\sum_{j\neq i} q_t^{ji} \mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] \\
&\leq \frac{16\zeta^2}{L^2(N-4)} + \frac{3\sigma^2}{8L^2} + \frac{9N\zeta^2}{2L^2} + 16B\gamma^2\sigma^2 \\
&+ \Big(\frac{1}{8L^2N} + 128B^2\gamma^2\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \ .
\end{aligned}$$

Proof. Proposition B.2 and Proposition B.3 can be interpreted as a system of linear inequalities. Applying (a) Proposition B.2 and (b) Proposition B.3 respectively, we get

$$\begin{aligned}
&\sum_{j\neq i} q_t^{ji} \mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] \\
&\overset{(a)}{\leq} \frac{2}{5}\sum_{j\neq i} q_t^{ji} \mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] + \frac{9N\zeta^2}{10L^2} + \frac{16B\gamma^2\sigma^2}{5} \\
&+ \frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&\overset{(b)}{\leq} \frac{2}{5}\Big(2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\Big] + \frac{8\zeta^2}{L^2(N-4)} \\
&\quad + \frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] + \frac{3\sigma^2}{16L^2}\Big) \\
&+ \frac{9N\zeta^2}{10L^2} + \frac{16B\gamma^2\sigma^2}{5} + \frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&= \frac{4}{5}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\Big] + \frac{16\zeta^2}{5L^2(N-4)} \\
&+ \frac{1}{40L^2N}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] + \frac{3\sigma^2}{40L^2} + \frac{9N\zeta^2}{10L^2} \\
&+ \frac{16B\gamma^2\sigma^2}{5} + \frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \ ,
\end{aligned}$$

and therefore,

$$\begin{aligned}
&\frac{1}{5}\sum_{j\neq i} q_t^{ji} \mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\big] \\
&\leq \frac{16\zeta^2}{5L^2(N-4)} + \frac{3\sigma^2}{40L^2} + \frac{9N\zeta^2}{10L^2} + \frac{16B\gamma^2\sigma^2}{5} \\
&+ \Big(\frac{1}{40L^2N} + \frac{128B^2\gamma^2}{5}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \ .
\end{aligned}$$

□

Lemma C.4. *(Upper bound for the local reference model change) When Assumption 2 holds true during $[t_0, t)$ for all users (i.e., when the number of events during the given period $[t_0, t)$ is finite) and $\gamma \leq \min(\frac{1}{8BL}, \frac{1}{8BLN\Psi})$, we have*

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)} - \mathbf{x}_{t_0}^{(j)}\big\|^2\Big] \\
&\leq \frac{9N\zeta^2}{L^2} + 32B\gamma^2\sigma^2 + \frac{40\zeta^2}{L^2(N-4)} + \frac{15\sigma^2}{16L^2} \\
&\quad + \Big(256B^2\gamma^2 + \frac{5}{16L^2N^2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \ .
\end{aligned}$$

Proof. Approaching in the same fashion as in Lemma C.3, we have

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)} - \mathbf{x}_{t_0}^{(j)}\big\|^2\Big] \\
&\overset{(a)}{\leq} 2\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\|\mathbf{y}_{t,b}^{(j)} - \mathbf{x}_t^{(j)}\|^2\Big] + \frac{8\zeta^2}{L^2(N-4)} \\
&+ \frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] + \frac{3\sigma^2}{16L^2} \\
&\overset{(b)}{\leq} 2\Big(\frac{2}{5}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] + \frac{9N\zeta^2}{10L^2} \\
&\quad + \frac{16B\gamma^2\sigma^2}{5} + \frac{128B^2\gamma^2}{5}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2]\Big) \\
&+ \frac{8\zeta^2}{L^2(N-4)} + \frac{1}{16L^2N^2}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] + \frac{3\sigma^2}{16L^2} \\
&= \frac{4}{5}\sum_{j\neq i} q_t^{ji}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\mathbf{x}_t^{(i)} - \mathbf{x}_{t_0}^{(i)}\big\|^2\big] + \frac{9N\zeta^2}{5L^2} + \frac{32B\gamma^2\sigma^2}{5} \\
&+ \frac{8\zeta^2}{L^2(N-4)} + \frac{3\sigma^2}{16L^2} \\
&+ \Big(\frac{256B^2\gamma^2}{5} + \frac{1}{16L^2N^2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \ ,
\end{aligned}$$

where (a) uses Proposition B.3, and (b) comes from Lemma C.3. □

D. Proof of Theorem 1

The proof of Theorem 1 is based on the proof provided in [25].

Beginning with rephrasing the L-smoothness between $f_i(\mathbf{x}_{t_0+P}^{(i)})$ and $f_i(\mathbf{x}_{t_0}^{(i)})$, we have

$$\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q},t_0}[f_i(\mathbf{x}_{t_0+P}^{(i)})] \\
&\leq f_i(\mathbf{x}_{t_0}^{(i)}) + \mathbb{E}_{\cdot|\mathcal{Q},t_0}[\langle\nabla f_i(\mathbf{x}_{t_0}^{(i)}),\ \mathbf{x}_{t_0+P}^{(i)} - \mathbf{x}_{t_0}^{(i)}\rangle]
\end{aligned}$$

$$
\begin{aligned}
&+\frac{L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\big\|\mathbf{x}^{(i)}_{t_0+P}-\mathbf{x}^{(i)}_{t_0}\big\|^2\Big]\\
&\le f_i(\mathbf{x}^{(i)}_{t_0})-\gamma\Big\langle\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\\
&\qquad\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big]\Big\rangle\\
&+\frac{\gamma^2L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\Big]\\
&=f_i(\mathbf{x}^{(i)}_{t_0})-\gamma\Big\langle\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\\
&\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbb{E}\big[\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})|\mathcal{Q},\mathbf{y}^{(j)}_{\lfloor t\rfloor,b},\mathbf{x}^{(i)}_{t_0}\big]\,\mathrm{d}t\Big]\Big\rangle\\
&+\frac{\gamma^2L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\Big]\\
&=f_i(\mathbf{x}^{(i)}_{t_0})-\gamma\Big\langle\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\\
&\qquad\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big]\Big\rangle\\
&+\frac{\gamma^2L}{2}\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\Big]
\end{aligned}
$$

Taking expectation on both sides over $\mathbf{x}^{(i)}_{t_0}$, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0+P})]\\
&\le\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})]-\gamma\mathbb{E}_{\cdot|\mathcal{Q},t_0}\Big[\Big\langle\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\\
&\qquad\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\rangle\Big]\\
&+\frac{\gamma^2L}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\Big]
\end{aligned}
\tag{D.16}
$$

Here, we reintroduce a finite variable from Definition 1, $\Psi\in\mathbb{R}^+$, to indicate the maximum total number of all message exchanging events during the period $[t_0,t_0+P)$. We set an assumption that $\Psi\ge 3$ for any time elapse $[t_0,t_0+P)$ in which t_0 is multiple to P.

Considering the second term in the inequality D.16,

$$
\begin{aligned}
&-\Big\langle\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\rangle\\
&=-\frac{1}{B\Psi}\Big\langle B\Psi\nabla f_i(\mathbf{x}^{(i)}_{t_0}),\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\Big\rangle\\
&=\frac{1}{2B\Psi}\Big\|B\Psi\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\\
&\quad-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&\quad-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\\
&=\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\big[\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\big]\,\mathrm{d}t\Big\|^2\\
&\quad-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&\quad-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\\
&=\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}[\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})]\,\mathrm{d}t\Big\|^2\\
&\quad-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&\quad-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2
\end{aligned}
$$

In order to deal with two variables controlled by different agents, two terms are added and subtracted for further proof: the local model gradient calculated by j and its local reference model, respectively.

$$
\begin{aligned}
&=\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\Big[\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\nabla f_j(\mathbf{x}^{(j)}_{t_0})\\
&\quad+\nabla f_j(\mathbf{x}^{(j)}_{t_0})-\nabla f_j\big(\mathbf{x}^{(j)}_t\big)+\nabla f_j\big(\mathbf{x}^{(j)}_t\big)\\
&\quad-\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\Big]\,\mathrm{d}t\Big\|^2-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\\
&\le\frac{3}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\big[\nabla f_j\big(\mathbf{x}^{(j)}_t\big)-\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\big]\,\mathrm{d}t\Big\|^2\\
&\quad+\frac{3}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\big[\nabla f_j(\mathbf{x}^{(j)}_{t_0})-\nabla f_j\big(\mathbf{x}^{(j)}_t\big)\big]\,\mathrm{d}t\Big\|^2\\
&\quad+\frac{3}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}[\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\nabla f_j(\mathbf{x}^{(j)}_{t_0})]\,\mathrm{d}t\Big\|^2\\
&\quad-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&\quad-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2\\
&\overset{(a)}{\le}\frac{3}{2B\Psi}\Big\|B\Psi\sum_{j\neq i}q^{ji}_t\big[\nabla f_j\big(\mathbf{x}^{(j)}_t\big)-\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\big]\Big\|^2\\
&\quad+\frac{3}{2B\Psi}\Big\|B\Psi\sum_{j\neq i}q^{ji}_t\big[\nabla f_j(\mathbf{x}^{(j)}_{t_0})-\nabla f_j\big(\mathbf{x}^{(j)}_t\big)\big]\Big\|^2\\
&\quad+\frac{3}{2B\Psi}\Big\|B\Psi\sum_{j\neq i}q^{ji}_t[\nabla f_i(\mathbf{x}^{(i)}_{t_0})-\nabla f_j(\mathbf{x}^{(j)}_{t_0})]\Big\|^2\\
&\quad-\frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2\\
&\quad-\frac{1}{2B\Psi}\Big\|\int_P\sum_{j\neq i}q^{ji}_t\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \frac{3BN\Psi}{2}\sum_{j\neq i} q_t^{ji}\Big\|\nabla f_j\big(\mathbf{x}_t^{(j)}\big) - \nabla f_j\big(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)}\big)\Big\|^2 \\
&+ \frac{3BN\Psi}{2}\sum_{j\neq i} q_t^{ji}\Big\|\nabla f_j\big(\mathbf{x}_{t_0}^{(j)}\big) - \nabla f_j\big(\mathbf{x}_t^{(j)}\big)\Big\|^2 \\
&+ \frac{3B\Psi}{2}\Big\|\sum_{j\neq i} q_t^{ji}\big[\nabla f_i(\mathbf{x}_{t_0}^{(i)}) - \nabla f_j(\mathbf{x}_{t_0}^{(j)})\big]\Big\|^2 \\
&- \frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2 \\
&- \frac{1}{2B\Psi}\Big\|\int_P \sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2 \\
&\overset{(c)}{\leq} \frac{3BL^2N\Psi}{2}\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)} - \mathbf{y}_{\lfloor t\rfloor,b}^{(j)}\big\|^2 \\
&+ \frac{3BL^2N\Psi}{2}\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_{t_0}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2 \\
&+ \frac{3BN\Psi\zeta^2}{N-4} - \frac{B\Psi}{2}\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2 \\
&- \frac{1}{2B\Psi}\Big\|\int_P \sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2 \qquad \text{(D.17)}
\end{aligned}
$$

where (a) reflects Jensen's inequality on the first three terms, pulling the terms out of the L2 norms; (b) is valid because $\|\sum_{j=1}^N q(j)\mathbf{z}(j)\|^2 \leq N\sum_{j=1}^N q(j)\|\mathbf{z}(j)\|^2$ for all $q_\star \in [0,1]$; (c) L-smoothness on the first two terms and Lemma 1 on the third term.

Hence, the expectation can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[-\Big\langle\nabla f_i(\mathbf{x}_{t_0}^{(i)}), \int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\rangle\Big] \\
&\leq \frac{3BL^2N\Psi}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_t^{(j)} - \mathbf{y}_{\lfloor t\rfloor,b}^{(j)}\big\|^2\Big] \\
&+ \frac{3BL^2N\Psi}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{j\neq i} q_t^{ji}\big\|\mathbf{x}_{t_0}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\Big] + \frac{3BN\Psi\zeta^2}{N-4} \\
&- \frac{B\Psi}{2}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&- \frac{1}{2B\Psi}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&\overset{(a)}{\leq} \frac{3BL^2N\Psi}{2}\Bigg[\frac{16\zeta^2}{L^2(N-4)} + \frac{3\sigma^2}{8L^2} + \frac{9N\zeta^2}{2L^2} + 16B\gamma^2\sigma^2 \\
&\qquad + \Big(\frac{1}{8L^2N} + 128B^2\gamma^2\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\Bigg] \\
&+ \frac{3BL^2N\Psi}{2}\Bigg[\frac{9N\zeta^2}{L^2} + 32B\gamma^2\sigma^2 + \frac{40\zeta^2}{L^2(N-4)} + \frac{15\sigma^2}{16L^2} \\
&\qquad + \Big(256B^2\gamma^2 + \frac{5}{16L^2N^2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\Bigg] \\
&+ \frac{3BN\Psi\zeta^2}{N-4} - \frac{B\Psi}{2}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&- \frac{1}{2B\Psi}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&= \frac{3BL^2N\Psi}{2}\Bigg[\frac{56\zeta^2}{L^2(N-4)} + \frac{21\sigma^2}{16L^2} + \frac{27N\zeta^2}{2L^2} + 48B\gamma^2\sigma^2 \\
&\qquad + \Big(\frac{7}{16L^2N} + 384B^2\gamma^2\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big]\Bigg] \\
&+ \frac{3BN\Psi\zeta^2}{N-4} - \frac{B\Psi}{2}\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&- \frac{1}{2B\Psi}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&= \frac{87BN\zeta^2\Psi}{N-4} + \frac{63BN\sigma^2\Psi}{32} + \frac{81BN^2\zeta^2\Psi}{4} \\
&+ 72B^2L^2N\gamma^2\sigma^2\Psi \\
&+ B\Psi\Big(\frac{21}{32N} + 576B^2L^2N\gamma^2 - \frac{1}{2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}\big[\big\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\big\|^2\big] \\
&- \frac{1}{2B\Psi}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \qquad \text{(D.18)}
\end{aligned}
$$

where (a) uses Lemma C.3 and Lemma C.4 on the first two terms, respectively.

Considering the third term in the inequality D.16,

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\mathbf{g}_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&= \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&+ \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}[\mathbf{g}_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)}) - \nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})]\,\mathrm{d}t\Big\|^2\Big] \\
&= \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] \\
&+ \int_P\sum_{j\neq i}(q_t^{ji})^2\sum_{b=0}^{B-1}\mathbb{E}_{\cdot|\mathcal{Q}}\big[\|\mathbf{g}_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)}) - \nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\|^2\big]\,\mathrm{d}t \\
&\overset{(a)}{\leq} \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big] + B\rho^2\sigma^2
\end{aligned} \qquad \text{(D.19)}
$$

where (a) is derived from the definition of σ in Assumption 4 and ρ.

Plugging D.18 and D.19, the inequality D.16 is rephrased as:

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}_{t_0+P}^{(i)})] \\
&\leq \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}_{t_0}^{(i)})] + \gamma\Bigg[\frac{87BN\zeta^2\Psi}{N-4} + \frac{63BN\sigma^2\Psi}{32} \\
&\quad + \frac{81BN^2\zeta^2\Psi}{4} + 72B^2L^2N\gamma^2\sigma^2\Psi \\
&\quad + B\Psi\Big(\frac{21}{32N} + 576B^2L^2N\gamma^2 - \frac{1}{2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}_{t_0}^{(i)})\|^2] \\
&\quad - \frac{1}{2B\Psi}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P\sum_{j\neq i} q_t^{ji}\sum_{b=0}^{B-1}\nabla f_j(\mathbf{y}_{\lfloor t\rfloor,b}^{(j)})\,\mathrm{d}t\Big\|^2\Big]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\gamma^2 L}{2}\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P \sum_{j\neq i} q_t^{ji} \sum_{b=0}^{B-1} \nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\,\Big\|^2\Big] \\
&+\frac{BL\gamma^2\rho^2\sigma^2}{2} \\
&\leq \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] + \frac{87BN\gamma\zeta^2\Psi}{N-4} + \frac{63BN\gamma\sigma^2\Psi}{32} \\
&+\frac{81BN^2\gamma\zeta^2\Psi}{4} + 72B^2L^2N\gamma^3\sigma^2\Psi \\
&+B\gamma\Psi\Big(\frac{21}{32N} + 576B^2L^2N\gamma^2 - \frac{1}{2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2] \\
&+\Big(\frac{\gamma^2 L}{2} - \frac{\gamma}{2B\Psi}\Big) \\
&\quad\cdot \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\int_P \sum_{j\neq i} q_t^{ji} \sum_{b=0}^{B-1} \nabla f_j(\mathbf{y}^{(j)}_{\lfloor t\rfloor,b})\,\mathrm{d}t\,\Big\|^2\Big] \\
&+\frac{BL\gamma^2\rho^2\sigma^2}{2} \\
&\overset{(a)}{\leq} \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] + \frac{87BN\gamma\zeta^2\Psi}{N-4} + \frac{63BN\gamma\sigma^2\Psi}{32} \\
&+\frac{81BN^2\gamma\zeta^2\Psi}{4} + 72B^2L^2N\gamma^3\sigma^2\Psi \\
&+B\gamma\Psi\Big(\frac{21}{32N} + 576B^2L^2N\gamma^2 - \frac{1}{2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2] \\
&+\frac{BL\gamma^2\rho^2\sigma^2}{2} \\
&\overset{(b)}{\leq} \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] + \frac{87BN\gamma\zeta^2\Psi}{N-4} + \frac{63BN\gamma\sigma^2\Psi}{32} \\
&+\frac{81BN^2\gamma\zeta^2\Psi}{4} + 72B^2L^2N\gamma^3\sigma^2\Psi \\
&+B\gamma\Psi\Big(\frac{21}{32N} + \frac{9}{N\Psi^2} - \frac{1}{2}\Big)\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2] \\
&+\frac{BL\gamma^2\rho^2\sigma^2}{2},
\end{aligned}
\tag{D.20}
$$

where (a) negates the term including $\nabla f_j(\mathbf{y}^{(j)}_{t,b})$ because $\frac{\gamma^2 L}{2} - \frac{\gamma}{2B\Psi} < 0$ based on the upper bound of γ; (b) bounds the coefficient of the term including $\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2]$ to simplify the further analysis.

After rearrangement, we have

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\|^2] \\
&\leq \frac{\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] - \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0+P})]}{B\gamma\Psi\Big(\frac{1}{2} - \frac{21}{32N} - \frac{9}{N\Psi^2}\Big)} \\
&\quad+\frac{1}{\frac{1}{2} - \frac{21}{32N} - \frac{9}{N\Psi^2}}\Bigg(\frac{87N\zeta^2}{N-4} + \frac{63N\sigma^2}{32} + \frac{81N^2\zeta^2}{4} \\
&\qquad+72BL^2N\gamma^2\sigma^2 + \frac{L\gamma\rho^2\sigma^2}{2\Psi}\Bigg) \\
&\overset{(a)}{\leq} \frac{128\big(\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] - \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0+P})]\big)}{11B\gamma\Psi} \\
&\quad+\frac{128}{11}\Bigg(\frac{87N\zeta^2}{N-4} + \frac{63N\sigma^2}{32} + \frac{81N^2\zeta^2}{4} \\
&\qquad+72BL^2N\gamma^2\sigma^2 + \frac{L\gamma\rho^2\sigma^2}{2\Psi}\Bigg) \\
&= \frac{128\big(\mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0})] - \mathbb{E}_{\cdot|\mathcal{Q}}[f_i(\mathbf{x}^{(i)}_{t_0+P})]\big)}{11B\gamma\Psi} + \frac{11136N\zeta^2}{11(N-4)} \\
&\quad+\frac{252N\sigma^2}{11} + \frac{2592N^2\zeta^2}{11} \\
&\quad+9216BL^2N\gamma^2\sigma^2 + \frac{64L\gamma\rho^2\sigma^2}{11\Psi}\ ,
\end{aligned}
\tag{D.21}
$$

where (a) makes the denominator smaller than the derived upper bound of inequality D.20 by using $N > 4$ and $\Psi \geq 3$, resulting in

$$
\frac{1}{\frac{1}{2} - \frac{21}{32N} - \frac{9}{N\Psi^2}} \leq \frac{1}{\frac{1}{2} - \frac{21}{32\cdot 4} - \frac{9}{4\cdot 3^2}} = \frac{128}{11}.
$$

Finally, the minimum value of $\mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f(\mathbf{x}_t)\|^2]$ over time t can be found as:

$$
\begin{aligned}
&\min_t \mathbb{E}_{\cdot|\mathcal{Q}}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&= \min_t \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}^{(i)}_t)\Big\|^2\Big] \\
&\leq \min_{t_0\in\{0,P,\cdots,(\lfloor\frac{T}{P}\rfloor-1)P\}} \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}^{(i)}_{t_0})\Big\|^2\Big] \\
&\overset{(a)}{\leq} \min_{t_0\in\{0,P,\cdots,(\lfloor\frac{T}{P}\rfloor-1)P\}} \frac{1}{N}\cdot\mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{i=1}^N \big\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\big\|^2\Big] \\
&\leq \frac{1}{N\lfloor\frac{T}{P}\rfloor}\cdot\sum_{t_0=0,P,2P,\cdots,(\lfloor\frac{T}{P}\rfloor-1)P} \mathbb{E}_{\cdot|\mathcal{Q}}\Big[\sum_{i=1}^N \big\|\nabla f_i(\mathbf{x}^{(i)}_{t_0})\big\|^2\Big] \\
&\overset{(b)}{\leq} \frac{1}{N\lfloor\frac{T}{P}\rfloor}\sum_{i=1}^N\Bigg[\frac{128\big(f_i(\mathbf{x}^{(i)}_0) - f_i^*\big)}{11B\gamma\Psi}\Bigg] + \frac{11136\zeta^2}{11(N-4)} \\
&\quad+\frac{252\sigma^2}{11} + \frac{2592N\zeta^2}{11} + 9216BL^2\gamma^2\sigma^2 + \frac{64L\gamma\rho^2\sigma^2}{11N\Psi} \\
&\overset{(c)}{\leq} \frac{128}{11B\gamma\Psi}(f(\mathbf{x}_0) - f^*) + \frac{11136\zeta^2}{11(N-4)} + \frac{252\sigma^2}{11} \\
&\quad+\frac{2592N\zeta^2}{11} + 9216BL^2\gamma^2\sigma^2 + \frac{64L\gamma\rho^2\sigma^2}{11N\Psi} \\
&= \mathcal{O}\Big(\frac{\mathcal{F}}{B\gamma\Psi} + \frac{\zeta^2}{N-4} + \sigma^2 + N\zeta^2 \\
&\qquad+BL^2\gamma^2\sigma^2 + \frac{L\gamma\rho^2\sigma^2}{N\Psi}\Big)
\end{aligned}
$$

where (a) is due to Jensen's inequality; the first term of (b) is an implantation of inequality D.21 whereas the other terms are independent on t_0; (c) takes that $P \leq T$ and the definition of $f(\mathbf{x}_\star)$.

E. Pseudo algorithm of DRACO

In this section, we provide a pseudo-algorithm and a flowchart for the intuitive reproduction of source code. The flowchart in Fig. E.11 includes only the transmission/reception procedure of DRACO, corresponding to lines 19-35 (excluding periodic unification parts) of Algorithm 2.

Algorithm 2 Pseudo algorithm of Alg. (1).

INITIALIZE {Generate `ListEvents`(i)}
for $i = 1, \cdots, N$ **do**
  Generate $t \sim exp(\lambda_i)$
  Append $[t, i]$ to `ListEventsGrad`(i)
  for event in `ListEventsGrad`(i) **do**
    Generate $t \sim exp(\lambda_{ij})$ or $t \leftarrow$transmission delay
    for $j \in \mathcal{N}(i)$ **do**
      Append $[t, j]$ to `ListEventsComm`(i)
    end for
  end for
  `ListEvents`(i) $\leftarrow$ `ListEventsGrad`(i) + `ListEventsComm`(i)
end for{Generate `ListEvents` over all clients}
for $i = 1, \cdots, N$ **do**
  Stack `ListEvents`(i) on `ListEvents`
end for
Sort `ListEvents` by t in ascending order.
Add the event indices k in front of each element.
$K \leftarrow |$`ListEvents`$|$
for $k = 1, \cdots, K$ **do**
  $(i, j) \leftarrow$ `ListEvents`$(k, 0)$, `ListEvents`$(k, 1)$
  if $i == j$ **then**
    for $b = 0, \cdots, B-1$ **do**
      $\mathbf{y}_{b+1}^{(i)} \leftarrow \mathbf{y}_b^{(i)} - \gamma g_i(\mathbf{y}_b^{(i)})$ {local batch training}
    end for
    $\Delta_k^{(i)} \leftarrow \mathbf{y}_{k,B}^{(i)} - \mathbf{x}_k^{(i)}$
  else
    for $j \in \mathcal{U} \setminus \{i\}$ **do**
      if `event_code`==“unification” **then**
        $\mathbf{x}^{(j)} \leftarrow \tilde{\mathbf{x}}^{(hub)}$
      else
        $\mathbf{x}^{(j)} \leftarrow \mathbf{x}^{(j)} + q_k^{ij} \tilde{\Delta}^{(i)}$ {aggregation}
      end if
    end for
  end if
end for
if $t \equiv 0 \pmod{P}$ and $t > 0$ and i is the hub **then**
  $\mathbf{x}^{(hub)} \leftarrow \mathbf{x}^{(i)}$
end if

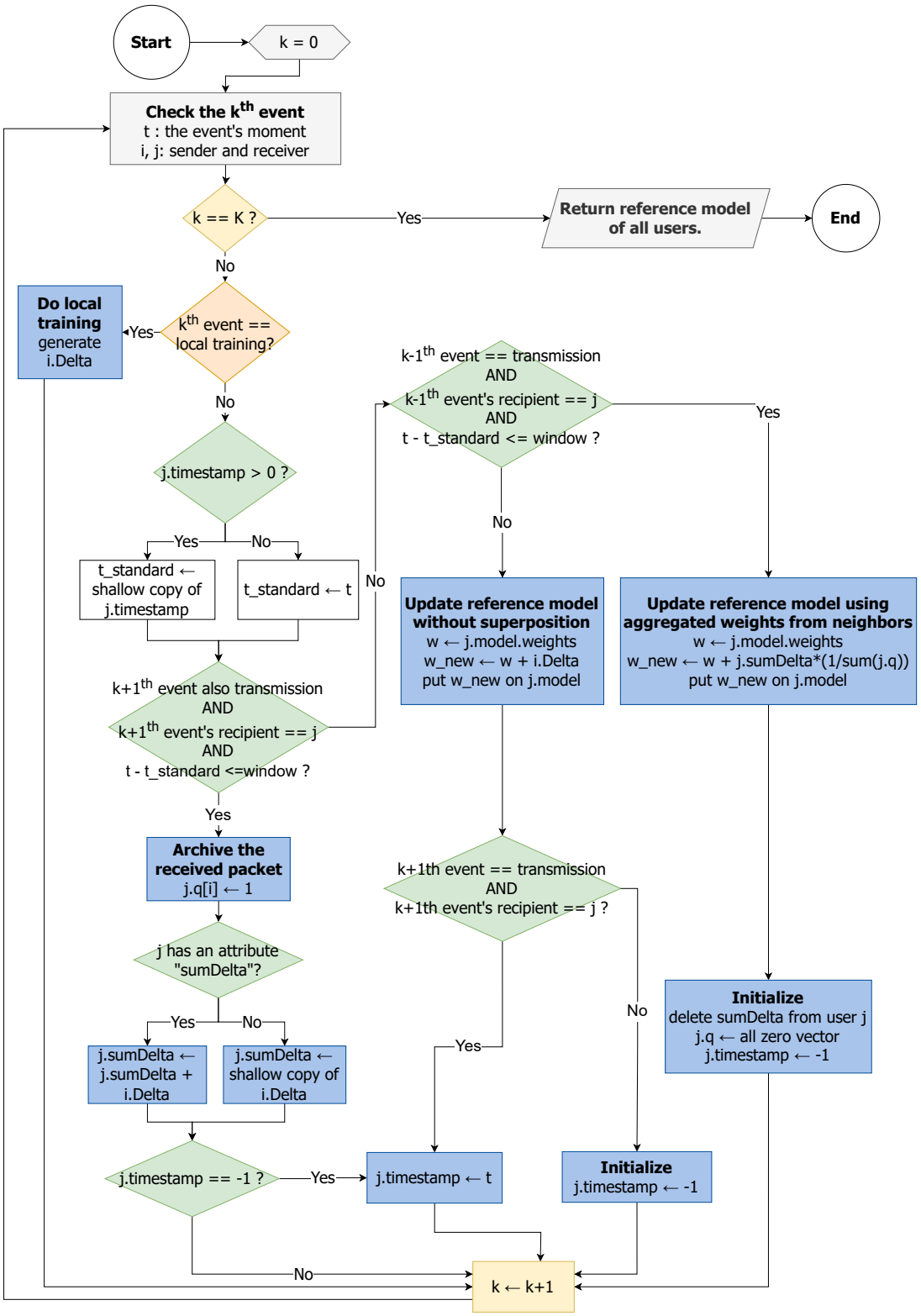


Fig. E.11. Flowchart of DRACO after initialization