Federated learning over physical channels: adaptive algorithms with near-optimal guarantee

Rui Zhang
Department of Electrical Engineering
University at Buffalo
Buffalo, NY 14226 USA
rzhang45@buffalo.edu

Wenlong Mou
Department of Statistical Science
University of Toronto
Toronto, ON Canada
wmou.work@gmail.com

Abstract

In federated learning, communication cost can be significantly reduced by transmitting real-valued gradient information directly through physical channels. However, the bias induced by hardware quantization and large variance due to channel noise create significant challenges for convergence analysis and algorithm design.

In this paper, we propose a new class of pre-coding and post-coding techniques to ensure exact unbiasedness and low variance of the transmitted stochastic gradient. Building upon these techniques, we design adaptive federated stochastic gradient descent (SGD) algorithms that can be implemented over physical channels for both downlink broadcasting and uplink transmission. We establish theoretical guarantees for the proposed algorithms, demonstrating convergence rates that are adaptive to the stochastic gradient noise level from data, without degradation due to channel noise. We also demonstrate the practical effectiveness of our algorithms through simulation studies with deep learning models. Our simulation results with CIFAR-10 and MNIST datasets show test accuracy matching that of the full-precision coded channel, costing only 20% of communication symbols.

1 Introduction

In modern machine learning applications, large datasets are often distributed across multiple heterogeneous worker machines. To jointly solve the optimization problem, the distributed machines need to transmit the information through communication channels. The bottleneck of distributed and federated learning is often the communication cost [1, 2]. The development of efficient federated learning algorithms with low communication cost has been a central topic in the machine learning community for the past decade (see [3–6] and references therein).

Most federated learning literature focuses on network-layer abstraction of communication channels, which allows error-free transmissions of real-valued data [7]. Nevertheless, such a transmission scheme can be costly in practice, requiring error correction codes and high-precision floating-point numbers. On the other hand, first-order stochastic optimization algorithms are known to be resilient to random noises in gradient oracles. In particular, if we want to minimize an objective function F, given a stochastic gradient oracle $\widehat{g}(\theta)$ satisfying the conditions

$$\mathbb{E}\big[\widehat{g}(\theta) \mid \theta\big] = \nabla F(\theta), \quad \text{and} \quad \mathbb{E}\big[\|\widehat{g}(\theta)\|_2^2 \mid \theta\big] < +\infty, \tag{1}$$

various stochastic first-order methods are guaranteed to converge efficiently. This observation motivates study of federated learning over physical communication channels [8, 9]. For example, when transmitting stochastic gradients directly through analogue channels with Additive Gaussian White Noise (AWGN), the stochastic gradient oracle $\widehat{g}(\theta)$ satisfies the condition (1), thereby achieving the same convergence guarantees with significantly reduced communication costs. [10–14].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Despite the inspiring progress, existing federated learning algorithms over physical channels suffer from both practical limitations and theoretical gaps. From the signal processing perspective, existing works overlook hardware constraints in practical systems: due to the quantization steps in the conversion between analogue and digital signals, Eq. (1) is not satisfied in general, and the biases in stochastic oracle may deteriorate the convergence of stochastic optimization algorithms. From the optimization perspective, existing algorithms either require fully coded communication in one of the uplink and the downlink [8, 12], or require the noise level to decay sufficiently fast [13, 15] – under these transmission schemes, the reduction in communication costs are limited. Finally, the noisy communication channel may amplify the variance of stochastic gradient oracles, leading to sub-optimal performance compared to coded channels. These limitations motivate the key question:

> Can we significantly reduce the communication costs of federated learning using practical physical channels for both downlink and uplink, while retaining desirable performance guarantees?

We answer this question in the affirmative by introducing a new class of channels signal processing techniques and adaptive stochastic gradient descent (SGD) algorithms for federated learning over physical channels. Our contribution are threefold:

- To tackle the biases induced by analogue-to-digital conversion (ADC), we introduce a stochastic post-coding procedure that corrects the biases in the quantized signals. The post-coding procedure is adaptive to the noise level in the communication channel, and can be implemented with low computational overhead.
- Under a worker-server architecture, we propose a new class of adaptive federated SGD algorithms that transmit the stochastic gradient information primarily through physical channels, and use the coded channel to synchronize parameters only once in a while. The communication cost is significantly lower than existing algorithms.
- We establish theoretical guarantees for the proposed algorithms, demonstrating near-optimal convergence rates. In particular, we show that the error bounds are adaptive to the stochastic gradient noise level, achieving statistical errors comparable to those of coded channels.

Notations We use $[m]$ to denotes the set $\{1, 2, \ldots, m\}$. For $j \in [d]$, we use $e_j \in \mathbb{R}^d$ to denote an indicator vector with 1 in the j-th coordinate and 0 elsewhere. We use $f \circ g$ to denote the composition of functions f and g, i.e., $(f \circ g)(x) = f(g(x))$. For $p \in [1, +\infty)$, we denote the vector ℓ^p norm by $\|x\|_p := \big(\sum_{j=1}^d |x_j|^p\big)^{1/p}$, and $\|x\|_\infty := \max_{j\in[d]} |x_j|$. Given a pair of vector norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, we use $|\!|\!|A|\!|\!|_{X\to Y} := \sup_{\|x\|_X=1} \|Ax\|_Y$ to denote the induced matrix norm. We use $\langle x, y\rangle$ to denote the standard inner product in $\mathbb{R}^d$. For the iterative algorithms we study we use $\mathcal{F}_k$ to denote the σ-field generated by the first k iterations.

Technical highlights The key technical challenge in designing federated learning algorithms over physical channels lies in constructing stochastic gradient oracles that satisfy the unbiasedness condition Eq. (1) with low variance. We introduce a novel post-coding technique that uses a probability transition kernel to eliminate the bias (see Section 3.1), and a scale-adaptive transformation to adapt the variance (see Section 3.2) to the variance of stochastic gradients from data. Applying these techniques with a refined analysis of stochastic optimization algorithms, we establish near-optimal convergence guarantees for the proposed algorithms (see Section C). In particular, the error bound depends on the stochastic gradient noise at the minimizer, matching the statistically optimal rates for stochastic optimization.

2 Problem setup

We consider a federated optimization problem with m workers machines. Each worker machine $j \in \{1, 2, \cdots, m\}$ is associated with a probability distribution $\mathbb{P}_j$ and a dataset $\mathcal{D}_j = \big(X_i^{(j)}\big)_{i=1}^n$, such that $X_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_j$ for each j. Our goal is to jointly solve the following optimization problem

$$\min_{\theta\in\mathbb{R}^d} F(\theta) := \mathbb{E}_{\mathbb{P}}\big[f(\theta; X)\big] \quad \text{where } \mathbb{P} = \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j. \tag{2}$$

A central server machine is used to aggregate information from different worker machines. In particular, a communication link exists between each worker machine $j \in [m]$ and the server machine. The algorithm can choose to transmit information through either coded or physical channels. In the following subsection, we will discuss these two types of channels.

2.1 Models for physical constraints

In this section, we summarize the physical models for communication channels and hardware devices considered in this paper.

2.1.1 Coded vs. physical transmission channels

In standard coded communication systems, the gradients and model parameters are transmitted as floating numbers through a coded channel. To transmit a real number with a floating number precision of 2^{-b}, we need b bits to encode the information. The information is further modulated as PAM signal. Given a PAM of order 2^ℓ, and an error correction code with overhead rate α, the average number of symbols needed to transmit a real number is $\frac{b}{\ell}(1+\alpha)$. For example, with a 32-bit floating precision number, PAM-4 modulation, and 5.8% forward error correction overhead [16], we need 8.46 symbols on average to transmit a real number.

Physical channels, on the other hand, transmits the real numbers as analogue signals directly. Due to the random noise in the communication channels and the hardware constraints, information cannot be transmitted exactly in such channels. Throughout this paper, we consider a channel with Additive Gaussian White Noise (AWGN). Given an input sequence $(X_1, X_2, \cdots, X_t)$, the channel outputs

$$\mathcal{C}(X_i) := X_i + \varepsilon_i, \quad \text{for } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_c^2). \tag{3}$$

For the rest of this paper, we use $\mathcal{C}$ to denote the random mapping defined by Eq. (3).

In addition to the AWGN model, we also consider the impact of hardware constraints on the communication process.

2.1.2 Conversion between analogue and digital signals

Information is stored as floating point numbers in the memory of digital devices. For transmission, the floating point numbers need to be converted to analogue signals through a digital-to-analogue converter (DAC). On the other hand, the received analogue signals are converted back to digital signals through an analogue-to-digital converter (ADC).

Concretely, let the quantization levels be $z_1 < z_2 < \cdots < z_q$ with $|z_i - z_{i-1}| = \Delta$ for each i, the DAC hardware takes digital representation for one of the levels z_i as input, and the output is the corresponding analogue signal. The ADC hardware takes the received analogue signal x as input and outputs its nearest quantized level, i.e., the ADC component implements a deterministic mapping

$$\mathcal{Q}_{\mathrm{C}}(x) = \arg\min_{z_i} |x - z_i|.$$

Note that the numerical precision of floating point numbers is usually much higher than that of the quantized levels, i.e., we have $|z_i - z_{i+1}| \gg 2^{-b}$. This means that the quantization process can introduce significant errors, especially when the input signal x is not well-aligned with the quantization levels. In order to mitigate the quantization error, we use a randomized algorithmic quantizer $\mathcal{Q}_{\mathrm{D}}$ that maps floating-point data to the quantization levels, before passing through DAC.

$$\mathcal{Q}_{\mathrm{D}}(x) = \begin{cases} z_1 & \text{if } x < z_1, \\ z_q & \text{if } x \geq z_q, \\ z_{i+\iota} & \text{if } x \in [z_i, z_{i+1}), \text{ where } \iota \sim \mathrm{Ber}\Big(\frac{x - z_i}{z_{i+1} - z_i}\Big). \end{cases}$$

The randomized mapping makes the algorithmic quantizer $\mathcal{Q}_{\mathrm{D}}$ unbiased, i.e., for any $x \in [-1, 1]$, we have $\mathbb{E}[\mathcal{Q}_{\mathrm{D}}(x)] = x$. This randomized quantization scheme has been employed in federated learning literature [11, 17].

The output of the algorithmic quantizer is then transmitted through the DAC, the channel, and the ADC, sequentially. The overall mapping from the real-valued data to the received digital signals is

given by $\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}$, where $\mathcal{Q}_{\mathrm{C}}$ and $\mathcal{Q}_{\mathrm{D}}$ represent the ADC and algorithmic quantizer defined above, respectively; the channel $\mathcal{C}$ is defined in Eq. (3).

3 Adaptive SGD algorithms for physical channels

Let us now present the key components of our federated learning algorithms over physical channels. The framework consists of three components: a stochastic post-coding procedure that ensures unbiasedness of the received signals, a scale-adaptive transformation that makes the variance of the stochastic gradient oracle adaptive to the noise level, and periodic synchronization of global model parameters to reduce coded communication rounds.

The communication between workers and the centralized server follows the following protocol: assuming that a two-way link has been established between the server and each worker, they can exchange information through two types of channels:

- A physical channel, where the signals pass through the DAC unit, the channel, and the ADC unit,sequentially. The channel takes a quantized scalar value $x \in \{z_1, z_2, \cdots, z_q\}$ as input, and outputs $\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(x)$ at the receiver side.
- A coded channel, which takes bit sequences as input, and uses error correction codes to guarantee error-free transmissions.

Throughout our algorithms, we assume that the uplink and downlink channels can use both the physical and coded channels. The communication cost is measured by the total number of symbols transmitted through either channels.

3.1 Stochastic post-coding

Although the additive Gaussian noises in physical channels are unbiased, when combined with the nonlinear quantization process, they can lead to biased estimates, which jeopardize the convergence of stochastic optimization algorithms. To address this issue, we introduce a stochastic post-coding procedure to correct the bias, as described in the following section.

For $i \in [q]$, the composition mapping $\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}$ is generally biased, i.e., $\mathbb{E}[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_i)] \neq z_i$. The goal of the stochastic post-coding procedure is to construct a stochastic mapping $\mathcal{H}$, such that

$$\mathbb{E}\big[\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_i)\big] = z_i, \quad \text{for } i \in \{2, 3, \cdots, q-1\}. \tag{4}$$

Note that we only guarantee the unbiasedness of the mapping for the quantization levels in the interior of the quantization grid. Since the output space is constrained in $\{z_1, z_2, \cdots, z_q\}$, the mapping $\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}$ cannot guarantee unbiasedness for the boundary points z_1 and z_q. However, as long as $q \geq 4$, we can still carry information using only the interior points. Let us now describe the post-coding mapping $\mathcal{H}$.

To start with, we use P to indicate the transition probabilities of the mapping $\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}$, i.e., for $i, j \in [q]$

$$P_{i,j} = \mathbb{P}(\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_i) = z_j) = \begin{cases} \Phi\big(\frac{z_j + \Delta/2 - z_i}{\sigma_c}\big) - \Phi\big(\frac{z_j - \Delta/2 - z_i}{\sigma_c}\big), & j \in [2, q-1], \\ \Phi\big(\frac{z_1 + \Delta/2 - z_i}{\sigma_c}\big), & j = 1, \\ 1 - \Phi\big(\frac{z_q - \Delta/2 - z_i}{\sigma_c}\big), & j = q, \end{cases}$$

where Φ is the CDF of the standard normal distribution. We use the transition matrix H to represent the mapping $\mathcal{H}$, i.e., $H_{i,j} = \mathbb{P}\big(\mathcal{H}(z_i) = z_j\big)$. We solve the following linear program to find the matrix H:

$$\min_{H \in \mathbb{R}^{q \times q}, v \in \mathbb{R}} v \qquad \text{such that} \tag{5a}$$

$$H_{i,j} \geq 0, \quad \forall i, j \in [q]; \ \sum_{j=1}^{q} H_{i,j} = 1, \quad \forall i \in [q]; \tag{5b}$$

$$e_j^\top P H z = z_j \quad \forall j \in \{2, 3, \cdots, q-1\}, \tag{5c}$$

$$\sum_{i=1}^{q} (PH)_{j,i} \cdot (z_i - z_j)^2 \leq v \quad \forall j \in \{2, 3, \cdots, q-1\}. \tag{5d}$$

The constraint (5b) ensures that the mapping $\mathcal{H}$ is a valid probability transition matrix, and the constraint (5c) guarantees the unbiasedness property (4). Under the constraint (5d), the objective function (5a) minimizes the worst-case variance of the mapping $\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}$. If the linear program (5) is feasible, with an optimal solution (H^*, v^*). Then the mapping $\mathcal{H}$ satisfies the unbiasedness property (4), and

$$\operatorname{var}\big(\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_i)\big) \leq v^*, \quad \forall i \in \{2, 3, \cdots, q-1\}.$$

The construction of the mapping $\mathcal{H}$ relies on the feasibility of the linear program (5). While feasibility is not guaranteed in general, the following lemma shows that the linear program is feasible when the SNR is sufficiently large.

Lemma 1. *For any $\sigma_c \leq \Delta/2$, the linear program* (5) *is feasible. Furthermore, the optimal value v^* satisfies the bound $v^* \leq 4\Delta^2$.*

See Section D.1.1 for the proof of this lemma. This lemma ensures feasibility of the linear program (5) when the noise level is small enough. Furthermore, it also guarantees that the variance of the post-coding mapping $\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}$ is dominated by the quantization error Δ^2.

Let us briefly discuss the implementation of the post-coding mapping $\mathcal{H}$. Given a pre-specified noise level σ_c and a quantization grid $\{z_1, z_2, \cdots, z_q\}$, the server can solve the linear program (5) offline to obtain the optimal transition matrix H^*. During the online transmission process, given an input quantized signal z_i, the server generates a random index j according to the distribution defined by the i-th row of H^*, and outputs z_j as the post-coded signal. This procedure can be efficiently implemented by DSP hardware using standard techniques such as alias method.

3.2 Scale-adaptive transformation

Another component in our algorithms is a scale-adaptive transformation $(\beta_\omega, \Psi_\omega)$, parametrized by a tuning parameter $\omega > 0$. Given an input scalar x, we define a pair of functions

$$\beta_\omega(x) := \max\Big(0, \lceil \log_2\big(\omega^{-1}\,|x|\,\big)\rceil\Big), \quad \text{and} \quad \Psi_\omega(x) := (1-\Delta)x / \big(2^{\beta_\omega(x)}\omega\big). \tag{6a}$$

In other words, we compare $|x|$ with a binary grid $(2^k\omega)_{k\geq 0}$, and sort it to the level corresponding to the index $\beta_\omega(x)$. We then re-scale the scalar with respect to this grid level to obtain $\Psi_\omega(x)$. Clearly, it is always guaranteed that $|\Psi_\omega(x)| \leq 1-\Delta$, so that the output lies within the interval $[z_2, z_{q-1}]$, applicable to the post-coding scheme described above. The tuning parameter ω is a small positive scalar chosen to address the trade-offs between communication complexity and statistical errors, which will be reflected in the theoretical guarantees.

For notational convenience, we also extend the mapping Ψ to real vectors, in an entry-wise fashion. Concretely, given $x = [x_1, x_2, \cdots, x_d]^\top \in \mathbb{R}^d$, we define $\Psi_\omega(x) := [\Psi_\omega(x_1), \cdots, \Psi_\omega(x_d)]^\top$ and $\beta_\omega(x) := [\beta_\omega(x_1), \cdots, \beta_\omega(x_d)]^\top$. We also define the inverse operation that assembles the information transmitted through two channels into the original real-valued data:

$$A_\omega(\psi, b) := \frac{1}{1-\Delta} 2^b \omega \cdot \psi, \tag{6b}$$

and its vectorized version defined as entry-wise operations. This function assembles the information transmitted through two channels into the original real-valued data.

We need the following technical lemma about the physical channel and the quantization process.

Lemma 2. *Given a deterministic vector $u \in \mathbb{R}^d$, define*

$$\widehat{u} := A_\omega\Big(\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}\big(\Psi_\omega(u)\big), \beta_\omega(u)\Big),$$

we have $\mathbb{E}[\widehat{u}] = u$, and the following inequalities hold

$$\mathbb{E}\big[\|\widehat{u} - u\|_2^2\big] \leq (4v^* + \Delta^2) \cdot \big(4\|u\|_2^2 + \omega^2 d\big).$$

See Section D.1.2 for the proof of this lemma. This lemma ensures that the post-coded physical channel $\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}$ satisfies Eq. (1), thereby facilitating the convergence of stochastic optimization algorithms. It is worth noting that the variance of the post-coded physical channel is adaptive to the transmitted signal u itself.

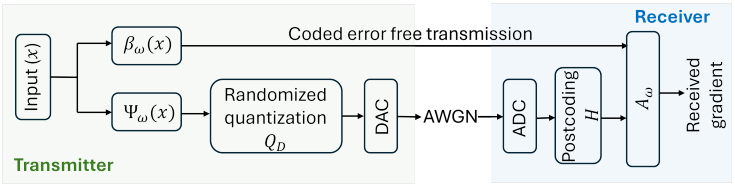


Figure 1. Block diagram of transmission process for physical channels. This transmission scheme applies to both uplink and downlink transmissions in federated learning.

In Fig. 1, we describe the overall transmission process for physical channels, combining the quantization, channel noise, and post-coding steps, as well as the scale-adaptive transformation framework.

The scale-adaptive transformation can also be implemented efficiently using DSP hardware. Since the input x is represented as floating point numbers, we can extract the exponent part of the floating point representation to compute $\beta_\omega(x)$ efficiently. The re-scaling step in $\Psi_\omega(x)$ and $A_\omega(\psi, b)$ can be implemented using bit-shift operations, which are computationally efficient.

3.3 Adaptive SGD over physical channels

Given the data transmission routines established in the previous sections, we are now ready to describe the algorithmic framework for federated learning over physical channels. We work with a worker-server network architecture. In each round, the workers compute the local stochastic gradient and send it to the server, and the server broadcasts the aggregated gradient information. All the data transmissions in this process use the transmission scheme described in Fig. 1, with the scale information transmitted through the coded channel, and the normalized values transmitted through the physical channel.

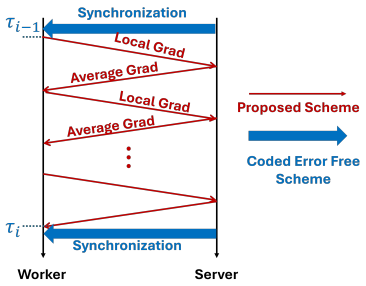


Figure 2. Federated learning algorithm overview

Additionally, we introduce a synchronization step to maintain stability. In particular, given an increasing sequence $\tau_1 < \tau_2 < \cdots < \tau_k < \cdots$ of time steps, the central machine broadcasts the current global model parameters θ_k to all workers at time step τ_k, for each k. Upon receiving the synchronization message, the workers replace their local model parameters by the global ones. The synchronization steps do not need to be frequent. In our theoretical analysis, we will show a bound on the requirement for time intervals between synchronization steps.

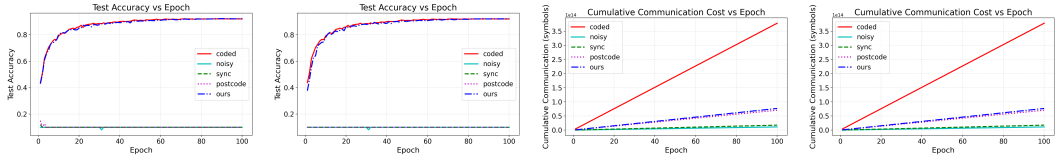
In Fig. 2, we provide an illustration of the adaptive SGD framework. In Algorithms 1 and 2 in Appendix B, we present the detailed procedures for server and workers, respectively.

4 Simulation studies

In this section, we conduct simulation studies to validate the theoretical findings of our paper. We consider a simple federated learning problem of image classification on the CIFAR-10 and MNIST dataset. We compared the performance of 5 different transmission schemes: the fully coded channel; direct use of the noisy channel; the post-coding and scale-adaptive transformation scheme without synchronization; the synchronized channel without post-coding or scale-adaptive transformation; and our proposed channel that incorporates all these techniques. We test the performance of these schemes in terms of test accuracy and communication costs (measured in terms of number of symbols). The details about the simulation setup are provided in Section E.

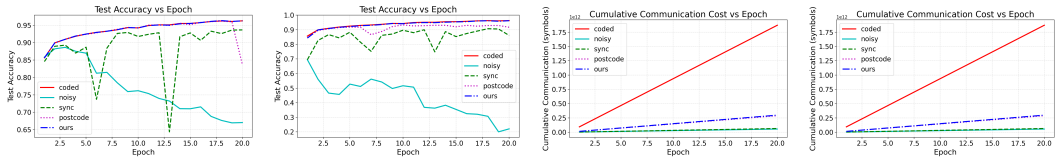
We present the results for CIFAR-10 and MNIST datasets in Fig. 3 and Fig. 4, respectively. In each figure, sub-figures (a) and (b) show the test accuracy over epochs for high and low SNR regimes, respectively, while sub-figures (c) and (d) show the communication cost over epochs for high and low SNR regimes, respectively. The communication cost is measured in terms of the total number of symbols transmitted through the channel, which include both coded and physical transmissions.

From Figs. 3 and 4 (a, b), we consistently observe that the test accuracy of our method matches the performance of coded transmission schemes in both high and low SNR regimes. Indeed, in all our simulations, the test accuracy of our method and the coded transmission results differs by less than



(a) (b) (c) (d)

Figure 3. Simulation results for CIFAR-10 dataset. (a) and (b): Test accuracy over epochs for high and low SNR regimes, respectively. (c) and (d): Communication cost over epochs for high and low SNR regimes, respectively.



(a) (b) (c) (d)

Figure 4. Simulation results for MNIST dataset. (a) and (b): Test accuracy over epochs for high and low SNR regimes, respectively. (c) and (d): Communication cost over epochs for high and low SNR regimes, respectively.

0.1%. In contrast, if we use the noisy physical channel directly, or use the post-coding approach or the synchronization framework alone, the test accuracy drops significantly, and even degrades to random guessing. This difference is especially pronounced in the more challenging CIFAR-10 dataset. On the other hand, the communication cost of our method is consistently lower than that of the coded transmission schemes, significant savings in the number of symbols transmitted ($4\times$ on CIFAR-10 dataset, $5\times$ on MNIST dataset), as shown in Figs. 3 and 4 (c) and (d). The post-coding and scale-adaptive transformation leads to some overhead in communication cost compared to the direct use of the noisy channel, but this is outweighed by the gains in test accuracy.

5 Discussion and conclusion

In this paper, we propose a novel algorithmic framework for communication-efficient distributed learning with quantized gradients over bi-directional noisy channels. We introduce three key technical components:

- A post-coding scheme that ensures unbiased transmitted signal, even with non-linear quantization.
- A scale-adaptive transformation that dynamically adjusts the quantization levels.
- A federated learning framework that stabilizes the training process by synchronizing model parameters periodically.

Our theoretical results demonstrate that, under standard assumptions on the loss function and stochastic gradients, the proposed method achieves convergence rates comparable to those of fully centralized methods, even in the presence of low-resolution ADC/DAC quantization and high channel noise. We also provide empirical evidence supporting our theoretical findings: the proposed scheme achieves the same test accuracy with less than 20% of the communication cost compared to fully coded communications. The simulation results further illustrate that all the technical components are indispensable for achieving the desired performance.

This work opens several avenues for future research. One promising direction is to explore improved schemes to transmit the coded part in scale-adaptive transformation, which could further reduce the communication overhead. Additionally, investigating the impact of different network topologies on the performance of the proposed framework could yield valuable insights. Finally, extending the algorithms to accommodate more realistic communication channel models in wireless and optical systems, as well as more hardware constraints, remains an important challenge.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant ECCS-2512911 and Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2024-05092.

References

[1] Li, M., D. G. Andersen, A. Smola, et al. Communication efficient distributed machine learning with the parameter server. *Advances in neural information processing systems*, 27, 2014.

[2] Li, T., A. K. Sahu, A. Talwalkar, et al. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[3] Zhang, Y., J. C. Duchi, M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

[4] Boyd, S., N. Parikh, E. Chu, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

[5] Chang, T.-H., M. Hong, H.-T. Wai, et al. Distributed learning in the nonconvex world: From batch data to streaming and beyond. *IEEE Signal Processing Magazine*, 37(3):26–38, 2020.

[6] McMahan, B., E. Moore, D. Ramage, et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[7] Arjevani, Y., O. Shamir. Communication complexity of distributed convex learning and optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

[8] Amiri, M. M., D. Gündüz. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. *IEEE Transactions on Signal Processing*, 68:2155–2169, 2020.

[9] Yang, K., T. Jiang, Y. Shi, et al. Federated learning via over-the-air computation. *IEEE Transactions on Wireless Communications*, 19(3):2022–2035, 2020.

[10] Amiri, M. M., D. Gündüz. Federated learning over wireless fading channels. *IEEE Transactions on Wireless Communications*, 19(5):3546–3557, 2020.

[11] Amiri, M. M., D. Gunduz, S. R. Kulkarni, et al. Federated learning with quantized global model updates. *arXiv preprint arXiv:2006.10672*, 2020.

[12] Amiri, M. M., D. Gündüz, S. R. Kulkarni, et al. Convergence of federated learning over a noisy downlink. *IEEE Transactions on Wireless Communications*, 21(3):1422–1437, 2021.

[13] Wei, X., C. Shen. Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1253–1268, 2022.

[14] Xiao, B., X. Yu, W. Ni, et al. Over-the-air federated learning: Status quo, open challenges, and future directions. *Fundamental Research*, 2024.

[15] Upadhyay, A., A. Hashemi. Noisy communication of information in federated learning: An improved convergence analysis. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pages 666–669. IEEE, 2023.

[16] IEEE Standard for Ethernet, 2018. Revision of IEEE Std 802.3-2015.

[17] Youn, Y., Z. Hu, J. Ziani, et al. Randomized quantization is all you need for differential privacy in federated learning. *arXiv preprint arXiv:2306.11913*, 2023.

[18] Shah, S. M., L. Su, V. K. N. Lau. Robust federated learning over noisy fading channels. *IEEE Internet of Things Journal*, 10(9):7993–8013, 2022.

[19] Tegin, B., T. M. Duman. Blind federated learning at the wireless edge with low-resolution ADC and DAC. *IEEE Transactions on Wireless Communications*, 20(12):7786–7798, 2021.

[20] Li, J., Z. Chen, K. F. E. Chong, et al. Robust federated learning over the air: Combating heavy-tailed noise with median anchored clipping. In *2025 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 89–96. IEEE, 2025.

[21] Zhang, N., M. Tao, J. Wang, et al. Coded over-the-air computation for model aggregation in federated learning. *IEEE Communications Letters*, 27(1):160–164, 2022.

[22] Ang, F., L. Chen, N. Zhao, et al. Robust federated learning with noisy communication. *IEEE Transactions on Communications*, 68(6):3452–3464, 2020.

[23] Guo, W., R. Li, C. Huang, et al. Joint device selection and power control for wireless federated learning. *IEEE Journal on Selected Areas in Communications*, 40(8):2395–2410, 2022.

[24] Yang, H., P. Qiu, J. Liu, et al. Over-the-air federated learning with joint adaptive computation and power control. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1259–1264. IEEE, 2022.

[25] Yao, J., W. Xu, Z. Yang, et al. Wireless federated learning over resource-constrained networks: Digital versus analog transmissions. *IEEE Transactions on Wireless Communications*, 23(10):14020–14036, 2024.

[26] Wannamaker, R. A., S. P. Lipshitz, J. Vanderkooy, et al. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing*, 48(2):499–516, 2002.

[27] Aysal, T. C., M. J. Coates, M. G. Rabbat. Distributed average consensus with dithered quantization. *IEEE Transactions on Signal Processing*, 56(10):4905–4918, 2008.

[28] Hasırcıoğlu, B., D. Gündüz. Communication efficient private federated learning using dithering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7575–7579. IEEE, 2024.

[29] Köster, U., T. Webb, X. Wang, et al. Flexpoint: An adaptive numerical format for efficient training of deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[30] Gupta, S., A. Agrawal, K. Gopalakrishnan, et al. Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR, 2015.

[31] Stich, S. U. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[32] Karimireddy, S. P., S. Kale, M. Mohri, et al. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[33] Moulines, E., F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[34] Ilandarideva, S., A. Juditsky, G. Lan, et al. Accelerated stochastic approximation with state-dependent noise. *Mathematical Programming*, pages 1–42, 2024.

[35] Proakis, J. G., M. Salehi. *Digital communications*, vol. 4. McGraw-hill New York, 2001.

[36] Agrell, E., M. Secondini. Information-theoretic tools for optical communications engineers. In *2018 IEEE Photonics Conference (IPC)*, pages 1–5. IEEE, 2018.

A Additional related works

Recently, federated learning over noisy channels has gained significant research attention. In addition to aforementioned works, several works have explored the interplay between communication and learning in this context. Extending the basic AWGN channel model, several practical communication scenarios are considered [10, 18–20], including the Gaussian noise and quantization process. Note that the stochastic noise becomes biased after applying the non-linear quantization mapping, creating obstacles for convergence of gradient-based algorithms. To our knowledge, our work is the first to construct an exact unbiased gradient oracle under this setting.

The performance of federated learning relies on estimation error for the gradient information on the receiver side. Advanced signal processing techniques have been employed to reduce the estimation error. [19] studies OFDM channels and shows that the noise can be mitigated by increasing number of receiver antennae. [21] combines channel coding techniques with the noisy transmission to reduce the error. [22] used regularization to improve the robustness of federated learning under noisy channels. Additionally, a recent line of research [23–25] studied resource allocation strategies for wireless federated learning, focusing on optimizing the trade-off between communication efficiency and learning performance.

Section 3 will present three main techniques we introduce to ensure near-optimal performance. Let us discuss connection of these techniques with the existing literature.

- The post-coding procedure is related to dithering [26], which injects random perturbations before quantization to reduce error; similar ideas appear in distributed computation [27] and federated learning [28]. Compared to existing dithering techniques, our post-coding procedure uses an additional randomization step after quantization to eliminate the bias exactly. To our knowledge, this is the first unbiased scheme that can be implemented over physical channels with both AWGN noise and ADC/DAC hardware constraints.
- The scale-adaptive transformation, which separates scale and normalized value, is akin to techniques for efficient low-precision training [29, 30]. We use this technique to establish adaptivity to noise level in stochastic gradient oracles, which is novel in the context of federated learning over physical channels.
- Periodic synchronization of global model parameters at the server is a classical strategy in distributed optimization [31, 32]. We use this strategy to reduce coded communication rounds. It is possible to further reduce the number of communication rounds (coded or physical) by using local updates at worker machines, which we leave for future work.

B Detailed description of the algorithms

Algorithm 1 Adaptive SGD over physical channels: worker side

Require: Initial point θ_0, where $\theta_0^{(j)} = \theta_0$ for $j \in [m]$.
for $k = 1, 2, \cdots, n$ **do**
Sample a local data $X_k^{(j)}$, and compute

$$g_k^{(j)} \sim \mathcal{Q}_{\mathrm{D}}\big(\Psi_\omega\big(\nabla f(\theta_{k-1}^{(j)}, X_k^{(j)})\big)\big)$$
$$\beta_k^{(j)} := \beta_\omega\big(\nabla f(\theta_{k-1}^{(j)}, X_k^{(j)})\big),$$

Transmit the real vector $g_k^{(j)}$ through the physical channel, and transmit the discrete vector $\beta_k^{(j)}$ to server through the coded channel.
Receive a real vector $\widehat{h}_k^{(j)} \sim \mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(h_k)$ through post-coded physical channel, and a discrete vector β_k through coded channel; update local parameter

$$\theta_k^{(j)} = \theta_{k-1}^{(j)} - \eta_k A_\omega\big(\widehat{h}_k^{(j)}, \beta_k\big).$$

if $k \in \{\tau_1, \tau_2, \cdots\}$ **then**
Receive θ_k from the server through the coded channel, and let $\theta_k^{(j)} = \theta_k$.
end if
end for

C Theoretical guarantees

We present the theoretical guarantees for the adaptive SGD algorithm over physical channels. We will first present the results under strongly convex settings, and then move on to non-convex settings.

C.1 Technical assumptions

We make the following technical assumptions in our analysis.

Assumption 1. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, population-level loss function F satisfies*

$$F(\theta_1) - F(\theta_2) \leq \langle \nabla F(\theta_1),\, \theta_1 - \theta_2 \rangle + \frac{L}{2}\|\theta_1 - \theta_2\|_2^2, \tag{7a}$$

$$F(\theta_1) - F(\theta_2) \geq \langle \nabla F(\theta_1),\, \theta_1 - \theta_2 \rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2. \tag{7b}$$

Algorithm 2 Adaptive SGD over physical channels: server side

for $k = 1, 2, \cdots, n$ **do**
  for each machine $j \in [m]$ **do**
    Receive the transmitted data $\widehat{g}_k^{(j)} \sim \mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(g_k^{(j)})$ through post-coded physical channel, and $\beta_k^{(j)}$ through coded channel.
  end for
  Aggregate the received information and update local parameter

$$u_k = \frac{1}{m} \sum_{j=1}^{m} A_\omega\big(\widehat{g}_k^{(j)}, \beta_k^{(j)}\big), \quad \text{and} \quad \theta_k = \theta_{k-1} - \eta u_k.$$

  Send $h_k = \mathcal{Q}_{\mathrm{D}}(\Psi_\omega(u_k))$ through physical channels, and $\beta_k = \beta_\omega(u_k)$ through coded channels.
end for
if $k \in \{\tau_1, \tau_2, \cdots\}$ **then**
  Send θ_k to each workers through the coded channel.
end if

This assumption is standard in convex optimization literature. Note that we only require strong convexity and smoothness to hold for the population-level loss function F.

Assumption 2. *The stochastic gradient oracle satisfies the moment bound for any $\theta \in \mathbb{R}^d$*

$$\mathbb{E}_{\mathbb{P}_j}\big[\|\nabla f(\theta, X)\|_2^2\big] \le \sigma_{*,j}^2 + \ell^2\big(F(\theta) - F(\theta^*)\big).$$

We also define the average noise level $\sigma_*^2 := \frac{1}{m}\sum_{j=1}^m \sigma_{*,j}^2$.

The noise level σ_*^2 captures the average uncertainty in the gradient estimates across different workers, which governs the optimal statistical error for machine learning problems. This assumption is known as "state-dependent noise" condition in stochastic optimization literature [33, 34]. It is more general than the standard bounded variance assumption, which requires $\mathbb{E}_{\mathbb{P}_j}[\|\nabla f(\theta, X)\|_2^2] \le \sigma_j^2$ for any θ. The state-dependent noise condition is satisfied in many statistical learning problems, including generalized linear models.

C.2 Results under strongly convex settings

Under assumptions in Section C.1, we have the following convergence bounds for the adaptive SGD algorithm over physical channels, with last-iterate and average-iterate guarantees.

To establish the theoretical results, we require the stepsize schedule $(\eta_k)_{k\ge1}$ to satisfy

$$\eta_k \le (1 + \eta_{k+1}\mu/8)\eta_{k+1}, \qquad \text{and} \qquad \eta_k \le \frac{c_0}{\ell^2 + L}, \tag{8a}$$

for some universal constant $c_0 > 0$.

Given the stepsize schedule, we need the synchronization times to satisfy the bounds for $i = 1, 2, \cdots$

$$T(\tau_i) - T(\tau_{i-1}) \le \frac{1}{2L}, \quad \text{where } T(k) = \sum_{t=1}^{k} \eta_t. \tag{8b}$$

Under above setup, we can establish the theoretical results under strongly convex setup.

Theorem 1. *Under Assumptions 1 and 2, given the stepsize sequence and synchronization times described above, for any $n \ge 1$, we have*

$$\mathbb{E}\big[\|\theta_n - \theta^*\|_2^2\big] \le e^{-\frac{\mu}{2}T(n)}\|\theta_0 - \theta^*\|_2^2 + \frac{c\eta_n}{\mu}\Big(\frac{\sigma_*^2}{m} + (v^* + \Delta^2)\omega^2 d\Big).$$

See Section D.2 for the proof of Theorem 1.

A few remarks are in order. First, the bound in Theorem 1 is comparable to standard results for SGD in the centralized setting, where the convergence rate takes the form

$$\mathbb{E}\big[\|\theta_n - \theta^*\|_2^2\big] \le e^{-\frac{\mu}{2}T(n)}\|\theta_0 - \theta^*\|_2^2 + c\eta_n \frac{\sigma_*^2}{\mu m}.$$

Compared to this bound, the additional term $\frac{c\eta_n}{\mu}(v^* + \Delta^2)\omega^2 d$ in Theorem 1 accounts for the distributed nature of the optimization problem and the variability in the gradient estimates across different workers. This term can be made small by choosing a small ω, which, however, increases the communication cost. This reflects the trade-off between communication cost and statistical accuracy in distributed optimization. Furthermore, the additional term also depends on the quantity $v^* + \Delta^2$, which reflects the impact of the SNR and hardware constraints. Note that we only need to transmit $\beta_\omega(u)$ through the coded channel, which requires only $O(d \log\log(1/\omega))$ bits. Therefore, we can choose a small ω without incurring significant communication overhead.

By taking the stepsize choice $\eta_k \asymp \frac{1}{\ell^2 + L + \mu k}$, for $n \gtrsim \frac{\ell^2 + L}{\mu}$, we have $T(n) \gtrsim \frac{1}{\mu}\log n$, and first term in the bound is negligible. This leads to a sample complexity bound of

$$\widetilde{O}\Big(\frac{\ell^2 + L}{\mu}\Big) + \widetilde{O}\Big(\frac{1}{\mu^2\varepsilon^2}\Big\{\frac{\sigma_*^2}{m} + (v^* + \Delta^2)\omega^2 d\Big\}\Big)$$

to achieve $\mathbb{E}\big[\|\theta_n - \theta^*\|_2^2\big] \le \varepsilon^2$, where the $\widetilde{O}(\cdot)$ notation hides logarithmic factors. The first term is the optimization error, and the second term is the statistical error. The statistical error matches the minimax optimal rate $\frac{\sigma_*^2}{\mu m}$ when the term $(v^* + \Delta^2)\omega^2 d$ is small enough. Compared to most existing results on federated learning, it is worth noting that the statistical error depends only on the noise level σ_*^2 at the minimizer θ^*, rather than a uniform bound on the stochastic gradient variance over the entire parameter space. In many applications, the noise at the minimizer can be significantly smaller, or even near-zero, leading to improved statistical accuracy.

Finally, we note that the synchronization requirement (8b) is not very stringent. For example, if we take the stepsize choice $\eta_k \asymp \frac{1}{\ell^2 + L + \mu k}$, then we have $T(k) \asymp \frac{1}{\mu}\log k$, and the synchronization requirement (8b) is satisfied as long as $\tau_i/\tau_{i-1} \le c$ for some constant $c > 1$. In other words, the synchronization times can be chosen to be geometrically increasing. On the other hand, if we choose a fixed stepsize $\eta_k = \eta > 0$, then the synchronization requirement (8b) requires $\tau_i - \tau_{i-1} \le \frac{1}{2L\eta}$, i.e., the synchronization steps need to be performed at a constant frequency that is inverse proportional to the stepsize.

C.3 Results under non-convex settings

Similarly, we can also establish results for finding stationary points of a non-convex function.

Assumption 3. *The stochastic gradient oracle satisfies the moment bound*

$$\mathbb{E}_{\mathbb{P}_j}\big[\|\nabla f(\theta, X)\|_2^2\big] \le \sigma_{*,j}^2 + \lambda\|\nabla F(\theta)\|_2^2, \quad \textit{for any } \theta \in \mathbb{R}^d,$$

This assumption is an extension of the state-dependent noise variance assumption 2 to the non-convex setting, which, once again, allows the stochastic gradient variance to be state-dependent and unbounded.

Aligned with standard practice in non-convex optimization literature, we measure the quality of a solution θ by the squared gradient norm $\|\nabla F(\theta)\|_2^2$. To present the result, we need to introduce a random variable R, which takes values in $\{0, 1, \cdots, n-1\}$, with probabilities proportional to the stepsizes,

$$\mathbb{P}(R = k) = \frac{\eta_{k+1}}{\sum_{t=1}^n \eta_t}.$$

Now we can state the main result for the non-convex setting.

Theorem 2. *Under Eq.* (7a) *and Assumption 3, when the stepsizes satisfy* $\eta_k \le \frac{c_0}{L(1+\lambda)}$*, and the synchronization times satisfy Eq.* (8b)*, we have*

$$\mathbb{E}\big[\|\nabla F(\theta_R)\|_2^2\big] \le \frac{F(\theta_0) - F_{\min} + cL\Big\{\frac{\sigma_*^2}{m} + (v^* + \Delta^2)\omega^2 d\Big\} \cdot \sum_{k=1}^n \eta_k^2}{\sum_{k=1}^n \eta_k},$$

where $c_0, c > 0$ are universal constants, and $F_{\min} = \inf_{\theta \in \mathbb{R}^d} F(\theta)$.

See Section D.3 for the proof of this theorem. A few remarks are in order. First, similar to Theorem 1, the bound in Theorem 2 is comparable to standard results for SGD in the centralized setting, with an additional term $L(v^* + \Delta^2)\omega^2 d$ accounting for communication channels and quantization hardwares. By taking the stepsize $\eta_k \asymp 1/\sqrt{n}$, we can achieve an $O(\varepsilon^{-4})$ sample complexity bound to achieve $\mathbb{E}[\|\nabla F(\theta_R)\|_2^2] < \varepsilon^2$, which is standard in the analysis of SGD in non-convex settings. Note that the synchronization requirement (8b) is the same as that in Theorem 1. Under the $1/\sqrt{n}$ stepsize choice, this requirement becomes $\tau_i - \tau_{i-1} \asymp \sqrt{n}$. In other words, we only need $O(\sqrt{n})$ broadcasting steps in the total n iterations.

D Proofs

We present the proofs of the main results in this section.

D.1 Proofs about the transmission channels

In this section, we collect the proofs of the technical lemmas about the transmission channels, the quantization process, and the post-coding procedure.

D.1.1 Proof of Lemma 1

We show feasibility and boundedness of the linear program (5) by direct construction. Given a vector $\zeta \in \mathbb{R}^{q-2}$, we define the $q \times q$ square matrix $H(\zeta)$ as follows (for simplicity, we index the elements of ζ from 2 to $q-1$).

$$H(\zeta) := \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1-\zeta_2}{3} & \frac{1}{3} & \frac{1+\zeta_2}{3} & 0 & \cdots & 0 \\ 0 & \frac{1-\zeta_3}{3} & \frac{1}{3} & \frac{1+\zeta_3}{3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & \frac{1-\zeta_{q-1}}{3} & \frac{1}{3} & \frac{1+\zeta_{q-1}}{3} \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Clearly, the matrix $H(\zeta)$ satisfies the constraint (5b) whenever $\|\zeta\|_\infty \leq 1$. Now we study the constraint (5c). First, since the quantization levels $(z_1, z_2, \cdots, z_q)$ are equi-spaced, for each $j \in \{2, 3, \cdots, q-2\}$, we note that

$$\begin{aligned} e_j^\top P H(0) z &= P_{j,1} z_1 + P_{j,q} z_q + \sum_{i=2}^{q-1} P_{j,i} \cdot \frac{1}{3}(z_i + z_{i-1} + z_{i+1}) \\ &= \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big], \end{aligned}$$

and consequently

$$\begin{aligned} e_j^\top P H(\zeta) z - z_j &= e_j^\top P\big(H(\zeta) - H(0)\big) z + \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big] - z_j \\ &= \sum_{i=2}^{q-1} P_{j,i} \frac{1}{3} \zeta_i (z_{i+1} - z_{i-1}) + \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big] - z_j \\ &= \frac{2\Delta}{3} \cdot \sum_{i=2}^{q-1} P_{j,i} \zeta_i + \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big] - z_j. \end{aligned}$$

Define the matrix $P^* \in \mathbb{R}^{(q-2)\times(q-2)}$ as the restriction of the matrix P to the rows and columns corresponding to the indices $\{2, 3, \cdots, q-1\}$. We claim that the matrix P^* is invertible, satisfying the bound

$$\|\!|(P^*)^{-1}|\!\|_{\infty \to \infty} \leq 3, \quad \text{when } \sigma_c \leq \frac{\Delta}{2}. \tag{9}$$

We prove this result at the end of this section. Taking this operator norm bound as given, we define the vector ζ^* as

$$\zeta^* = \frac{3}{2\Delta} \cdot (P^*)^{-1} \cdot \big(z_j - \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big]\big)_{j=2}^{q-1}.$$

To establish feasibility, we need to show that $\|\zeta^*\|_\infty \leq 1$. By Eq. (9), we have

$$\begin{aligned}\|\zeta^*\|_\infty &\leq \frac{3}{2\Delta} \cdot |\!|\!|(P^*)^{-1}|\!|\!|_{\infty\to\infty} \cdot \max_{j=2,3,\cdots,q-1} \big|z_j - \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big]\big| \\ &\leq \frac{9}{2\Delta} \max_{j=2,3,\cdots,q-1} \big|z_j - \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big]\big|.\end{aligned}$$

For the channel and ADC unit, we note that for any fixed $y \in [-1, 1]$, by symmetry of the normal density, we have

$$\begin{aligned}|\mathbb{E}[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(y)] - y| &= \left| \int_{-\infty}^{\infty} \big(\mathcal{Q}_{\mathrm{C}}(y+z) - y\big) \cdot \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-z^2}{2\sigma_c^2}} dz \right| \\ &= \left| \int_{0}^{\infty} \big(\mathcal{Q}_{\mathrm{C}}(y+z) + \mathcal{Q}_{\mathrm{C}}(y-z) - 2y\big) \cdot \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-z^2}{2\sigma_c^2}} dz \right| \\ &\leq 2\int_{1-|y|}^{+\infty} \frac{z}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-z^2}{2\sigma_c^2}} dz = \sqrt{\frac{2\sigma_c^2}{\pi}} \exp\Big(\frac{-(1-|y|)^2}{2\sigma_c^2}\Big).\end{aligned}$$

When $\sigma_c \leq \Delta/2$, this implies that

$$\max_{j=2,3,\cdots,q-1} \big|z_j - \mathbb{E}\big[\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big]\big| \leq \sigma_c \sqrt{\frac{2}{\pi}} \exp\Big(\frac{-\Delta^2}{2\sigma_c^2}\Big) \leq e^{-2}\sqrt{\frac{2}{\pi}}\sigma_c \leq \frac{1}{9}\Delta.$$

Consequently, we have $\|\zeta^*\|_\infty \leq 1/2 < 1$. So the matrix $H(\zeta^*)$ satisfies the constraints (5b) and (5c) simultaneously, and the linear program (5) is feasible. To derive the optimal value bound, we note that the variance under $H(\zeta^*)$ satisfies

$$\begin{aligned}\mathrm{var}\big(\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big) &\leq 2\mathbb{E}\big[\,|\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j) - \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)|^2\,\big] + 2\mathbb{E}\big[\,|\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j) - z_j|^2\,\big] \\ &\leq 2\Delta^2 + 2\mathbb{E}\big[\,|\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j) - z_j|^2\,\big],\end{aligned}$$

where we use Young's inequality and almost-sure boundedness of the random mapping $\mathcal{H}$. To bound the second term, we define an auxiliary function $\mathcal{Q}'_{\mathrm{C}}(x) := \arg\min_{i\Delta : i\in\mathbb{Z}} |x - i\Delta|$. For any $x \in [-1, 1]$, we clearly have

$$\mathbb{E}\big[\,|\mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(x) - x|^2\,\big] \leq \mathbb{E}\big[\,|\mathcal{Q}'_{\mathrm{C}} \circ \mathcal{C}(x) - x|^2\,\big].$$

and we have

$$\mathbb{E}\big[\,|\mathcal{Q}'_{\mathrm{C}} \circ \mathcal{C}(x) - x|^2\,\big] \leq 2\mathbb{E}\big[\,|\mathcal{Q}'_{\mathrm{C}} \circ \mathcal{C}(x) - \mathcal{C}(x)|^2\,\big] + 2\mathbb{E}\big[\,|\mathcal{C}(x) - x|^2\,\big] \leq 2 \cdot \frac{\Delta^2}{4} + 2\sigma_c^2.$$

Putting them together, we noting that $\sigma_c \leq \Delta/2$, we have

$$\mathrm{var}\big(\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_j)\big) \leq 4\Delta^2,$$

which completes the proof of Lemma 1.

Proof of Eq. (9) For notational consistency, we index the elements of $(q-2)$-dimensional vectors from 2 to $q-1$. For any vector pair $x, y \in \mathbb{R}^{q-2}$ satisfying $x = P^*y$, we note that

$$|x_i| = \left| \sum_{j=2}^{q-1} P_{i,j} y_j \right| \geq P_{i,i}|y_i| - \sum_{j\neq i} P_{i,j}|y_j| \geq P_{i,i}|y_i| - \|y\|_\infty \sum_{j\neq i} P_{i,j}.$$

Taking i_0 to be the index with the largest $|y_i|$, we have

$$\|x\|_\infty \geq |x_{i_0}| \geq P_{i_0,i_0}|y_{i_0}| - \|y\|_\infty \sum_{j\neq i_0} P_{i_0,j} \geq \|y\|_\infty \cdot \Big\{P_{i_0,i_0} - \sum_{j\neq i_0} P_{i_0,j}\Big\} = \|y\|_\infty\big(2P_{i_0,i_0} - 1\big).$$

By definition, we have

$$P_{i_0,i_0} = \Phi\big(\frac{\Delta}{2\sigma_c}\big) - \Phi\big(-\frac{\Delta}{2\sigma_c}\big) \geq \frac{2}{3}, \quad \text{when } \sigma_c \leq \frac{\Delta}{2}.$$

Under this condition, we have that $\|P^* y\|_\infty \geq \frac{1}{3}\|y\|_\infty$, for any $y \in \mathbb{R}^{q-2}$. This implies that the matrix P^* is invertible, and the operator norm $\|\!|(P^*)^{-1}|\!\|_{\infty\to\infty} \leq 3$.

D.1.2 Proof of Lemma 2

We first prove unbiasedness. For each scalar x satisfying $|x| \leq 1 - \Delta$, suppose that $x \in [z_i, z_{i+1})$ for some $i \in \{2, 3, \cdots, q-2\}$. By definition, we have

$$\mathbb{E}[\mathcal{Q}_{\mathrm{D}}(x)] = z_i \mathbb{P}\big(\mathcal{Q}_{\mathrm{D}}(x) = z_i\big) + z_{i+1}\mathbb{P}\big(\mathcal{Q}_{\mathrm{D}}(x) = z_{i+1}\big) = x.$$

Furthermore, note that $\mathcal{Q}_{\mathrm{D}}(x) \in \{z_i, z_{i+1}\}$ for $i \in \{2, 3, \cdots, q-2\}$. So we have $\mathcal{Q}_{\mathrm{D}}(x) \in \{z_2, \cdots, z_{q-1}\}$ almost surely. By the linear program construction (5), we have

$$\mathbb{E}\big[\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}(x) \mid \mathcal{Q}_{\mathrm{D}}(x)\big] = \mathcal{Q}_{\mathrm{D}}(x),$$

and consequently, $\mathbb{E}[\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}(x)] = x$. By construction, each coordinate of $\Psi_\omega(u)$ is bounded by $1 - \Delta$. Applying the above argument to each coordinate, we have $\mathbb{E}[\widehat{u}] = u$.

Now we turn to bound the variance. For each coordinate $i \in [d]$, we note that

$$\mathbb{E}\big[\,|\widehat{u}_i - u_i|^2\,\big] = \frac{2^{2\beta_\omega(u_i)}\omega^2}{(1-\Delta)^2}\mathbb{E}\Big[\,\big|\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}\big(\Psi_\omega(u_i)\big) - \Psi_\omega(u_i)\big|^2\,\Big] \tag{10}$$

By construction, we have that

$$2^{\beta_\omega(u_i)}\omega \leq \max\big(2|u_i|, \omega\big).$$

For the variance terms, we note that

$$\mathbb{E}\big[\,|\mathcal{Q}_{\mathrm{D}}(\Psi_\omega(u_i)) - \Psi_\omega(u_i)|^2\,\big] \leq \sup_{|x|\leq 1-\Delta} \mathrm{var}(\mathcal{Q}_{\mathrm{D}}(x)) \leq \frac{\Delta^2}{4}.$$

$$\mathbb{E}\big[\,\big|\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}\big(\Psi_\omega(u_i)\big) - \mathcal{Q}_{\mathrm{D}}\big(\Psi_\omega(u_i)\big)\big|^2\,\big] \leq \max_{i\in\{2,\cdots,q-2\}} \mathbb{E}\big[\,|\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C}(z_i) - z_i|^2\,\big] \leq v^*.$$

Substituting these bounds into Eq. (10), we have

$$\mathbb{E}\big[\,|\widehat{u}_i - u_i|^2\,\big] \leq 2^{2\beta_\omega(u_i)}\frac{\omega^2}{(1-\Delta)^2}\Big\{v^* + \frac{\Delta^2}{4}\Big\} \leq \big(4v^* + \Delta^2\big) \cdot \big(4|u_i|^2 + \omega^2\big).$$

Aggregating the bounds for all the d coordinates, we conclude that

$$\mathbb{E}\big[\|\widehat{u} - u\|_2^2\big] \leq (4v^* + \Delta^2) \cdot \big(\omega^2 d + 4\|u\|_2^2\big).$$

D.2 Proof of Theorem 1

We start with the one-step error decomposition

$$\mathbb{E}\big[\|\theta_k - \theta^*\|_2^2\big] = \mathbb{E}\big[\|\theta_{k-1} - \theta^*\|_2^2\big] - 2\eta_k\mathbb{E}\big[\langle\theta_{k-1} - \theta^*,\, u_k\rangle\big] + \eta_k^2\mathbb{E}\big[\|u_k\|_2^2\big]. \tag{11}$$

To simplify the notation, we define the average disagreement between the local models and the global model as

$$D_k := \frac{1}{m}\sum_{j=1}^m \mathbb{E}\big[\|\theta_{k-1}^{(j)} - \theta_{k-1}\|_2^2\big]. \tag{12a}$$

We also define the average optimality gap and gradient norm across all workers as

$$G_k := \frac{1}{m}\sum_{j=1}^m \mathbb{E}\big[F(\theta_{k-1}^{(j)}) - F(\theta^*)\big], \tag{12b}$$

$$H_k := \frac{1}{m}\sum_{j=1}^m \mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big]. \tag{12c}$$

The proof crucially relies on bounds on the bias and variance of the aggregated stochastic gradient u_k, which are summarized in the following lemmas.

Lemma 3. *Under the setup of Theorem 1, we have*

$$\mathbb{E}\big[\langle \theta_{k-1} - \theta^*,\, u_k\rangle\big] \geq \frac{\mu}{4}\mathbb{E}\big[\|\theta_{k-1} - \theta^*\|_2^2\big] + \frac{H_{k-1}}{2} + \frac{G_{k-1}}{8L} - 3LD_{k-1}.$$

See Section D.2.1 for the proof of this lemma.

Lemma 4. *Under the setup of Theorem 1, we have*

$$\mathbb{E}[\|u_k\|_2^2] \leq c\frac{\sigma_*^2}{m} + c\frac{v^* + \Delta^2}{m}\omega^2 d + cG_{k-1} + c\frac{\ell^2}{m}H_{k-1}.$$

where c is a universal constant.

See Section D.2.2 for the proof of this lemma.

Lemma 5. *Under the setup of Theorem 1, for $k \in [\tau_{i-1}, \tau_i)$, we have*

$$D_k \leq c_1(v^* + \Delta^2) \cdot \sum_{t=\tau_{i-1}+1}^{k} \eta_t^2 \Big\{ \frac{\sigma_*^2}{m} + \omega^2 d + G_{t-1} + \frac{\ell^2}{m}H_{t-1}\Big\},$$

where $c_1 > 0$ is a universal constant.

See Section D.2.3 for the proof of Lemma 5.

With these lemmas given, let us prove Theorem 1. Substituting Lemma 3 and Lemma 4 into Eq. (11), we have

$$\begin{aligned}\mathbb{E}\big[\|\theta_k - \theta^*\|_2^2\big] \leq \big(1 - \tfrac{\mu\eta_k}{2}\big)\mathbb{E}\big[\|\theta_{k-1} - \theta^*\|_2^2\big] + \tfrac{c\eta_k^2}{m}\big(\sigma_*^2 + (v^* + \Delta^2)\omega^2 d\big) \\ + \big\{c\eta_k^2 - \tfrac{\eta_k}{4L}\big\}G_{k-1} + 6L\eta_k D_{k-1} + \big\{c\tfrac{\ell^2}{m}\eta_k^2 - \tfrac{\eta_k}{2}\big\}H_{k-1}.\end{aligned}$$

By taking stepsize satisfying the stability condition $\eta_k \leq \frac{c_0}{L+\ell^2}$ for any $k \geq 1$, above inequality can be simplified to

$$\mathbb{E}\big[\|\theta_k - \theta^*\|_2^2\big] \leq e^{-\mu\eta_k/2}\mathbb{E}\big[\|\theta_{k-1} - \theta^*\|_2^2\big] + c\eta_k^2\frac{\sigma_*^2 + (v^* + \Delta^2)\omega^2 d}{m} + 6L\eta_k D_{k-1} - \frac{\eta_k}{8L}G_{k-1} - \frac{\eta_k}{4}H_{k-1}.$$

Define the accumulated time steps $T(k) := \sum_{t=1}^{k}\eta_t$ for any $k \geq 1$, we unroll the recursion from $k = \tau_{i-1}$ to $k = \tau_i$, and obtain the bound

$$\begin{aligned}&\mathbb{E}\big[\|\theta_{\tau_i} - \theta^*\|_2^2\big] \\ &\leq e^{-\frac{\mu}{2}(T(\tau_i) - T(\tau_{i-1}))}\mathbb{E}\big[\|\theta_{\tau_{i-1}} - \theta^*\|_2^2\big] + \sum_{k=\tau_{i-1}+1}^{\tau_i} e^{-\frac{\mu}{2}(T(\tau_i) - T(k))}\frac{c\eta_k^2}{m}\big(\sigma_*^2 + (v^* + \Delta^2)\omega^2 d\big) \\ &\quad - \frac{1}{4}\sum_{k=\tau_{i-1}+1}^{\tau_i} e^{-\frac{\mu}{2}(T(\tau_i) - T(k))}\eta_k\big\{\frac{G_{k-1}}{2L} + H_{k-1}\big\} + 6L\sum_{k=\tau_{i-1}+1}^{\tau_i} e^{-\frac{\mu}{2}(T(\tau_i) - T(k))}\eta_k D_{k-1}.\end{aligned}$$

When the synchronization times satisfy $T(\tau_i) - T(\tau_{i-1}) \leq 2/\mu$, above bound can be simplified as

$$\begin{aligned}\mathbb{E}\big[\|\theta_{\tau_i} - \theta^*\|_2^2\big] \leq e^{-\frac{\mu}{2}(T(\tau_i) - T(\tau_{i-1}))}\mathbb{E}\big[\|\theta_{\tau_{i-1}} - \theta^*\|_2^2\big] + \sum_{k=\tau_{i-1}+1}^{\tau_i}\frac{c\eta_k^2}{m}\big(\sigma_*^2 + (v^* + \Delta^2)\omega^2 d\big) \\ + \sum_{k=\tau_{i-1}+1}^{\tau_i}\eta_k\Big\{6LD_{k-1} - \frac{G_{k-1}}{8eL} - \frac{H_{k-1}}{4e}\Big\}. \qquad (13)\end{aligned}$$

Invoking Lemma 5, we note that

$$\begin{aligned}&(v^* + \Delta^2)^{-1}\sum_{k=\tau_{i-1}+1}^{\tau_i}\eta_k D_{k-1} \leq c_1\sum_{k=\tau_{i-1}+1}^{\tau_i}\eta_k\sum_{t=\tau_{i-1}+1}^{k}\eta_t^2\Big\{\frac{\sigma_*^2}{m} + \omega^2 d + G_{t-1} + \frac{\ell^2}{m}H_{t-1}\Big\} \\ &\leq c_1\big(T(\tau_i) - T(\tau_{i-1})\big)\cdot\sum_{t=\tau_{i-1}+1}^{\tau_i}\eta_t^2\Big\{\frac{\sigma_*^2}{m} + \omega^2 d + G_{t-1} + \frac{\ell^2 H_{t-1}}{m}\Big\}.\end{aligned}$$

When the synchronization times satisfy $T(\tau_i) - T(\tau_{i-1}) < \frac{1}{2L}$, and the stepsizes satisfy $\eta_t \leq \frac{1}{24ec_1(L+\ell^2)}$ for every t, we have

$$\sum_{k=\tau_{i-1}+1}^{\tau_i} 6L\eta_k D_{k-1} \leq \sum_{k=\tau_{i-1}+1}^{\tau_i} \eta_k\Big\{\frac{G_{k-1}}{8eL} + \frac{H_{k-1}}{4e}\Big\} + 3c_1(v^* + \Delta^2) \sum_{k=\tau_{i-1}+1}^{\tau_i} \eta_k^2\big(\frac{\sigma_*^2}{m} + \omega^2 d\big)$$

Substituting them to Eq. (13) we have the the recursive bound

$$\mathbb{E}\big[\|\theta_{\tau_i} - \theta^*\|_2^2\big] \leq e^{-\frac{\mu}{2}(T(\tau_i)-T(\tau_{i-1}))}\mathbb{E}\big[\|\theta_{\tau_{i-1}} - \theta^*\|_2^2\big] + c\Big(\frac{\sigma_*^2}{m} + (v^*+\Delta^2)\omega^2 d\Big) \sum_{k=\tau_{i-1}+1}^{\tau_i} \eta_k^2.$$

Solving the recursion, we have

$$\mathbb{E}\big[\|\theta_{\tau_r} - \theta^*\|_2^2\big] \leq e^{-\frac{\mu}{2}T(\tau_r)}\|\theta_0 - \theta^*\|_2^2 + ce\Big(\frac{\sigma_*^2}{m} + (v^* + \Delta^2)\omega^2 d\Big) \sum_{k=1}^{\tau_r} e^{-\frac{\mu}{2}(T(\tau_r)-T(k))}\eta_k^2.$$

For the summation term, we claim that

$$\sum_{k=1}^{n} e^{-\frac{\mu}{2}(T(n)-T(k))}\eta_k^2 \leq 8\eta_n, \tag{14}$$

whenever $(1 + \mu\eta_k/8)\eta_k \geq \eta_{k-1}$ for every $k \geq 2$. Substituting this bound into the above inequality completes the proof of Theorem 1. It remains to prove the claim (14).

Proof of Eq. (14) Define $A_n := \sum_{k=1}^n e^{-\frac{\mu}{2}(T(n)-T(k))}\eta_k^2$. We prove the result by induction. For $n = 1$, clearly we have $A_1 = \eta_1^2 \leq 8\eta_1$. Assuming that $A_{n-1} \leq 8\eta_{n-1}$, for the case of A_n, we note the recursive formula

$$A_n = e^{-\mu\eta_n/2}A_{n-1} + \eta_n^2 \leq \big(1 - \mu\eta_n/4\big)A_{n-1} + \eta_n^2.$$

By induction hypothesis, we have $A_{n-1} \leq 8\eta_{n-1}/\mu$. Substituting into the recursive bound, we have $A_n \leq 8(1 - \mu\eta_n/4)\eta_{n-1}/\mu + \eta_n^2 \leq 8(1 - \mu\eta_n/4)\eta_{n-1}/\mu + \eta_n^2 \leq 8\frac{1-\mu\eta_n/4}{1+\mu\eta_n/8}\eta_n/\mu + \eta_n^2 \leq 8\eta_n/\mu$, which completes the induction proof.

D.2.1 Proof of Lemma 3

We start by noting the following basic inequalities for strongly convex and smooth functions. For any $\theta \in \mathbb{R}^d$, we have

$$\langle \nabla F(\theta),\, \theta - \theta^* \rangle \geq \frac{\mu L}{\mu + L}\|\theta - \theta^*\|_2^2 + \frac{1}{\mu + L}\|\nabla F(\theta)\|_2^2, \tag{15a}$$

$$\langle \nabla F(\theta),\, \theta - \theta^* \rangle \geq F(\theta) - F(\theta^*) + \frac{\mu}{2}\|\theta - \theta^*\|_2^2. \tag{15b}$$

Applying Lemma 2 to the uplink transmission process in the k-th iteration, we can compute the conditional expectation.

$$\mathbb{E}\big[u_k \mid (g_k^{(j)}, \beta_k^{(j)})_{j=1}^m\big] = \frac{1}{m}\sum_{j=1}^m \nabla f(\theta_{k-1}^{(j)}, X_k^{(j)}).$$

Now we further take expectations conditionally on $\mathcal{F}_{k-1}$ to obtain $\mathbb{E}\big[u_k \mid \mathcal{F}_{k-1}\big] = \frac{1}{m}\sum_{j=1}^m \nabla F(\theta_{k-1}^{(j)})$, and therefore

$$\mathbb{E}\big[\langle \theta_{k-1} - \theta^*,\, u_k\rangle\big] = \frac{1}{m}\sum_{j=1}^m \mathbb{E}\Big[\langle \theta_{k-1} - \theta^*,\, \nabla F(\theta_{k-1}^{(j)})\rangle\Big].$$

For each $j \in [m]$, by applying Eq. (15a) and Eq. (15b), we have

$$\mathbb{E}\big[\langle \theta_{k-1}^{(j)} - \theta^*,\, \nabla F(\theta_{k-1}^{(j)})\rangle\big] \geq \frac{\mu\mathbb{E}[\|\theta_{k-1}^{(j)} - \theta^*\|_2^2]}{2} + \frac{\mathbb{E}[F(\theta_{k-1}^{(j)}) - F(\theta^*)]}{2} + \frac{\mathbb{E}[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2]}{4L}.$$

By Cauchy–Schwarz inequality and Young's inequality, we also note that

$$\Big|\mathbb{E}\big[\langle\theta_{k-1}^{(j)}-\theta^*,\,\nabla F(\theta_{k-1}^{(j)})\rangle\big]-\mathbb{E}\big[\langle\theta_{k-1}-\theta^*,\,\nabla F(\theta_{k-1}^{(j)})\rangle\big]\Big|$$
$$\leq\sqrt{\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta_{k-1}\|_2^2\big]}\cdot\sqrt{\mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big]}\leq\frac{1}{8L}\mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big]+2L\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta_{k-1}\|_2^2\big].$$

Consequently, we can bound the inner product from below.

$$\mathbb{E}\big[\langle\theta_{k-1}-\theta^*,\,u_k\rangle\big]\geq\frac{\mu}{2m}\sum_{j=1}^m\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta^*\|_2^2\big]+\frac{1}{2}H_{k-1}+\frac{1}{8L}G_{k-1}-2LD_{k-1}.\tag{16}$$

Finally, for the first term on the right-hand-side, we note that

$$\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta^*\|_2^2\big]\geq\frac{1}{2}\mathbb{E}\big[\|\theta_{k-1}-\theta^*\|_2^2\big]-\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta_{k-1}\|_2^2\big],$$

for each $j\in[m]$. Averaging over m workers, we have that

$$\frac{\mu}{2m}\sum_{j=1}^m\mathbb{E}\big[\|\theta_{k-1}^{(j)}-\theta^*\|_2^2\big]\geq\frac{\mu}{4}\mathbb{E}\big[\|\theta_{k-1}-\theta^*\|_2^2\big]-\frac{\mu}{2}D_{k-1}.$$

Substituting to Eq. (16) completes the proof of Lemma 3.

D.2.2 Proof of Lemma 4

By definition, we have $u_k=\frac{1}{m}\sum_{j=1}^m A_\omega(\widehat{g}_k^{(j)},\beta_k^{(j)})$, where each $\widehat{g}_k^{(j)}$ is obtained by applying the composed stochastic transformation $\mathcal{H}\circ\mathcal{Q}_{\mathrm{C}}\circ\mathcal{C}\circ\mathcal{Q}_{\mathrm{D}}$ independently to the quantized signal $\Psi_\omega(\nabla f(\theta_{k-1}^{(j)},X_k^{(j)}))$. By Lemma 2, we have

$$\operatorname{var}\Big(A_\omega(\widehat{g}_k^{(j)},\beta_k^{(j)})\mid g_k^{(j)},\beta_k^{(j)}\Big)\leq(4v^*+\Delta^2)\big(4\|\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2+\omega^2 d\big).$$

Since the transmission between server and each workers are independent, we have

$$\operatorname{var}\big(u_k\mid(g_k^{(j)},\beta_k^{(j)})_{j\in[m]}\big)\leq\frac{4v^*+\Delta^2}{m}\Big\{\omega^2 d+\frac{4}{m}\sum_{j=1}^m\|\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2\Big\}.$$

Lemma 2 also implies that

$$\mathbb{E}\big[u_k\mid(g_k^{(j)},\beta_k^{(j)})_{j\in[m]}\big]=\frac{1}{m}\sum_{j=1}^m\nabla f(\theta_{k-1}^{(j)},X_k^{(j)}).$$

We can then bound the total second moment as

$$\begin{aligned}\mathbb{E}\big[\|u_k\|_2^2\big]&\leq\mathbb{E}\Big[\operatorname{var}\big(u_k\mid(g_k^{(j)},\beta_k^{(j)})_{j\in[m]}\big)\Big]+\mathbb{E}\Big\{\|\mathbb{E}\big[u_k\mid(g_k^{(j)},\beta_k^{(j)})_{j\in[m]}\big]\|_2\Big\}^2\\&\leq\frac{4v^*+\Delta^2}{m}\Big\{\omega^2 d+\frac{4}{m}\sum_{j=1}^m\mathbb{E}\big[\|\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2\big]\Big\}+\mathbb{E}\Big[\|\frac{1}{m}\sum_{j=1}^m\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2\Big].\end{aligned}$$

Conditionally on the filtration $\mathcal{F}_{k-1}$, we can invoke Assumption 2 to obtain the bounds

$$\mathbb{E}\big[\|\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2\mid\mathcal{F}_{k-1}\big]\leq\|\nabla F(\theta_{k-1}^{(j)})\|_2^2+\sigma_{*,j}^2+\ell^2\big(F(\theta_{k-1}^{(j)})-F(\theta^*)\big),$$

and since the data points are independently sampled at each worker, we have

$$\begin{aligned}\mathbb{E}\Big[\|\frac{1}{m}\sum_{j=1}^m\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\|_2^2\mid\mathcal{F}_{k-1}\Big]&=\|\frac{1}{m}\sum_{j=1}^m\nabla F(\theta_{k-1}^{(j)})\|_2^2+\sum_{j=1}^m\frac{\operatorname{var}\big(\nabla f(\theta_{k-1}^{(j)},X_k^{(j)})\mid\mathcal{F}_{k-1}\big)}{m^2}\\&\leq\frac{1}{m}\sum_{j=1}^m\|\nabla F(\theta_{k-1}^{(j)})\|_2^2+\sum_{j=1}^m\frac{\sigma_{*,j}^2+\ell^2\big(F(\theta_{k-1}^{(j)})-F(\theta^*)\big)}{m^2}.\end{aligned}$$

Substituting these bounds into the variance bound, we have

$$\mathbb{E}\big[\|u_k\|_2^2\big]\leq\frac{4v^*+\Delta^2}{m}\omega^2 d+\frac{1+16v^*+4\Delta^2}{m}\big(\sigma_*^2+\ell^2 H_{k-1}\big)+\Big\{1+\frac{16v^*+4\Delta^2}{m}\Big\}G_{k-1},$$

which completes the proof of Lemma 4.

D.2.3 Proof of Lemma 5

Define the vector

$$\widehat{u}_k^{(j)} := A_\omega\big(\widehat{h}_k^{(j)}, \beta_k\big) = A_\omega\big(\mathcal{H} \circ \mathcal{Q}_{\mathrm{C}} \circ \mathcal{C} \circ \mathcal{Q}_{\mathrm{D}}(\Psi_\omega(u_k)), \beta_\omega(u_k)\big).$$

The recursive update formulae can be written as

$$\theta_k = \theta_{k-1} - \eta_k u_k, \quad \text{and} \quad \theta_k^{(j)} = \theta_{k-1}^{(j)} - \eta_k \widehat{u}_k^{(j)}, \quad \text{for } j \in [m].$$

For each $j \in [m]$, we have the error expansion.

$$\mathbb{E}\big[\|\theta_k - \theta_k^{(j)}\|_2^2\big] = \mathbb{E}\big[\|\theta_{k-1} - \theta_{k-1}^{(j)}\|_2^2\big] + \eta_k^2 \mathbb{E}\big[\|u_k - \widehat{u}_k^{(j)}\|_2^2\big] - 2\eta_k \mathbb{E}\big[\langle \theta_{k-1} - \theta_{k-1}^{(j)}, u_k - \widehat{u}_k^{(j)} \rangle\big].$$

By Lemma 2, we have

$$\mathbb{E}\big[\widehat{u}_k^{(j)} \mid u_k\big] = u_k.$$

Note that the transmission between server and workers in k-th round is independent of $\mathcal{F}_{k-1}$. So we have

$$\mathbb{E}\big[\langle \theta_{k-1} - \theta_{k-1}^{(j)}, u_k - \widehat{u}_k^{(j)} \rangle\big] = 0.$$

On the other hand, by Lemma 2, we have the variance bound

$$\mathbb{E}\big[\|u_k - \widehat{u}_k^{(j)}\|_2^2 \mid u_k\big] \le (4v^* + \Delta^2)\big(4\|u_k\|_2^2 + \omega^2 d\big).$$

Putting them together, we have the recursion

$$D_k \le D_{k-1} + \eta_k^2 \mathbb{E}\big[\|u_k - \widehat{u}_k^{(j)}\|_2^2\big] \le D_{k-1} + \eta_k^2 (4v^* + \Delta^2)\big(4\mathbb{E}\big[\|u_k\|_2^2\big] + \omega^2 d\big). \tag{17}$$

On the other hand, when $k \in \{\tau_1, \tau_2, \cdots\}$, we have $\theta_k^{(j)} = \theta_k$ for all $j \in [m]$, and so $D_k = 0$.

For $k \in [\tau_{i-1}, \tau_i)$, unrolling the recursion (17) from k down to τ_{i-1}, we have

$$D_k \le (4v^* + \Delta^2) \sum_{t=\tau_{i-1}+1}^{k} \eta_t^2 \big(4\mathbb{E}\big[\|u_t\|_2^2\big] + \omega^2 d\big).$$

Substituting the bounds in Lemma 4, we have

$$D_k \le c_1(v^* + \Delta^2) \cdot \sum_{t=\tau_{i-1}+1}^{k} \eta_t^2 \Big\{ \frac{\sigma_*^2}{m} + \omega^2 d + cG_{k-1} + c\frac{\ell^2}{m} H_{k-1} \Big\},$$

which completes the proof of Lemma 5.

D.3 Proof of Theorem 2

By smoothness of the function F, we have

$$\mathbb{E}\big[F(\theta_k)\big] \le \mathbb{E}\big[F(\theta_{k-1})\big] - \eta_k \mathbb{E}\big[\langle \nabla F(\theta_{k-1}), u_k \rangle\big] + \frac{\eta_k^2}{2} L \mathbb{E}[\|u_k\|_2^2]. \tag{18}$$

We define the discrepancy term D_k and average gradient norm G_k according to Eq. (12) in the proof of Theorem 1.

We use the following lemma to control the cross term

Lemma 6. *Under the setup of Theorem 2, for each k, we have*

$$\mathbb{E}\big[\langle \nabla F(\theta_{k-1}), u_k \rangle\big] \ge \tfrac{1}{4}\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] + \tfrac{G_{k-1}}{4} - \tfrac{L^2 D_{k-1}}{2}.$$

See Section D.3.1 for the proof of this lemma.

Following the arguments in Lemma 4 using Assumption 3, it is easy to see that

$$\mathbb{E}[\|u_k\|_2^2] \le c\frac{\sigma_*^2}{m} + c\frac{v^* + \Delta^2}{m}\omega^2 d + c(1+\lambda)G_{k-1}. \tag{19}$$

Applying this bound to the arguments in Lemma 5, it is easy to show that

$$D_k \le c_1(v^* + \Delta^2) \sum_{t=\tau_{i-1}+1}^{k} \eta_t^2 \Big\{ \frac{\sigma_*^2}{m} + \omega^2 d + (1+\lambda) G_{k-1} \Big\}, \tag{20}$$

for each $k \in [\tau_{i-1}, \tau_i)$.

Now we substitute Lemma 6 and Eq. (19) into Eq. (18). Telescoping the summation from time 0 to time n, we note that

$$\begin{aligned} &\frac{1}{4}\sum_{k=1}^{n} \eta_k \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] + \frac{1}{4}\sum_{k=1}^{n} \eta_k G_{k-1} - \frac{L^2}{2}\sum_{k=1}^{n} \eta_k D_{k-1} \\ &\le \mathbb{E}[F(\theta_0)] - \mathbb{E}[F(\theta_n)] + \sum_{k=1}^{n} \eta_k^2 L \Big\{ c\frac{\sigma_*^2}{m} + c\frac{v^* + \Delta^2}{m}\omega^2 d + c(1+\lambda) G_{k-1} \Big\}. \end{aligned} \tag{21}$$

By Eq. (20), if the synchronization times satisfy $T(\tau_i) - T(\tau_{i-1}) \le 1/(2L)$ for each $i = 1, 2, \cdots$, we have

$$\sum_{k=1}^{n} \eta_k D_{k-1} \le \tfrac{c_1(v^*+\Delta^2)}{2L} \sum_{t=1}^{n} \eta_t^2 \Big\{ \frac{\sigma_*^2}{m} + \omega^2 d + (1+\lambda) G_{k-1} \Big\}$$

Substituting back to Eq. (21), when the stepsize satisfies $\eta_k \le \frac{c_0}{L(1+\lambda)}$, we have

$$\frac{1}{4}\sum_{k=1}^{n} \eta_k \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] \le \mathbb{E}[F(\theta_0)] - \mathbb{E}[F(\theta_n)] + cL\Big\{ \frac{\sigma_*^2}{m} + (v^* + \Delta^2)\omega^2 d \Big\} \sum_{k=1}^{n} \eta_k^2.$$

By the definition of the random variable R, we have

$$\sum_{k=1}^{n} \eta_k \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] = \mathbb{E}\big[\|\nabla F(\theta_R)\|_2^2\big] \cdot \sum_{k=1}^{n} \eta_k.$$

Substituting back to the telescope formula completes the proof.

D.3.1 Proof of Lemma 6

By Lemma 2, we have $\mathbb{E}\big[u_k \mid \mathcal{F}_{k-1}\big] = \frac{1}{m}\sum_{j=1}^{m} \nabla F(\theta_{k-1}^{(j)})$, and consequently

$$\begin{aligned} \mathbb{E}\big[\langle \nabla F(\theta_{k-1}),\, u_k \rangle\big] &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\big[\langle \nabla F(\theta_{k-1}),\, \nabla F(\theta_{k-1}^{(j)}) \rangle\big] \\ &\ge \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] - \frac{1}{m}\sum_{j=1}^{m} \sqrt{\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big]\mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)}) - \nabla F(\theta_{k-1})\|_2^2\big]} \\ &\ge \frac{1}{2}\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|_2^2\big] - \frac{L^2}{2} D_{k-1}, \end{aligned}$$

where in the last inequality, we use the smoothness of the function F and the Cauchy–Schwarz inequality.

On the other hand, we note that

$$\begin{aligned} \mathbb{E}\big[\langle \nabla F(\theta_{k-1}),\, u_k \rangle\big] &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\big[\langle \nabla F(\theta_{k-1}),\, \nabla F(\theta_{k-1}^{(j)}) \rangle\big] \\ &\ge \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big] - \frac{1}{m}\sum_{j=1}^{m} \sqrt{\mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big]\mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)}) - \nabla F(\theta_{k-1})\|_2^2\big]} \\ &\ge \frac{1}{2m}\sum_{j=1}^{m} \mathbb{E}\big[\|\nabla F(\theta_{k-1}^{(j)})\|_2^2\big] - \frac{L^2}{2} D_{k-1}. \end{aligned}$$

Combining the two bounds completes the proof of this lemma.

E Details of the simulation studies

The training data is distributed across $m = 10$ worker machines, and each worker has the data for each label. We use ResNet-18 architecture for the CIFAR-10 dataset, and a simple 4-layer convolutional neural network for the MNIST dataset.

We use the cross-entropy loss function for training, with stepsize $\eta = 0.01$, and the batch size is set to 64. We compare 5 different transmission schemes:

- **Coded**: In this scheme, all the information is transmitted through the coded channel, using 32-bit floating point precision.
- **Noisy**: In this scheme, the information is transmitted directly through the physical channel described in Section 2.1, which includes the DAC unit, the AWGN channel, and the ADC unit.
- **Postcode:** In this scheme, we apply the post-coding and scale-adaptive transformation techniques in Section 3.1 and Section 3.2, to transmit each parameter in an unbiased manner. A simple distributed SGD algorithm is used for training, without the synchronization step described in Algorithms 1 and 2.
- **Sync:** In this scheme, we run the distributed optimization framework in Algorithms 1 and 2, with the global model parameters being synchronized across all workers at communication rounds $\tau_1, \tau_2, \cdots$. The transmission is performed over the noisy channel described in Section 2.1, without the post-coding and scale-adaptive transformation techniques. In our simulation studies, the synchronization time frequency is set to be 30 communication rounds for CIFAR-10 dataset, and 100 communication rounds for MNIST dataset.
- **Ours:** This scheme corresponds to the full algorithms described in Algorithms 1 and 2, incorporating post-coding, scale-adaptive transformation, and synchronization.

We test these 5 different methods under simulated communication channels. To ensure a fair comparison, we require the average signal power to be the same for coded and noisy channels. We consider two different SNR regimes in our experiments:

- High SNR regime: we let $\sigma_c = 0.05$ in the communication channel, and the quantization levels are set to $q = 16$ (i.e., 4 bits). For coded communication, we consider a PAM-8 modulation with Gray mapping.[1] In this case, the SNR is approximately 19.5dB, and the pre-FEC BER is about 1.04×10^{-3} (see e.g. [35]). Following industry standards for FEC overhead [16, 36], we assume an FEC overhead of 5.8%.
- Low SNR regime: we let $\sigma_c = 0.2$ in the communication channel, and the quantization levels are set to $q = 8$ (i.e., 3 bits). In this case, the SNR is approximately 5.5dB. For coded communication, we consider a BPSK modulation, leading to a pre-FEC BER of 3.86×10^{-3} (see e.g. [35]). Following industry standards for FEC overhead [16, 36], we assume an FEC overhead of 5.8%.

We run the algorithms for a total of 100 epochs on CIFAR-10 dataset, and 20 epochs on MNIST dataset, and we report the test accuracy and communication overhead for each transmission scheme.

[1] In many communication systems, QAM modulations are used. In such a case, we use the real and imaginary parts to encode two different PAM symbols, so the number of symbols is halved for both coded and physical channels, and the comparison remains valid.