BEYOND PARAMETER CONSENSUS: COMMUNICATION-EFFICIENT PERSONALIZED FEDERATED ADMM WITH ONE-BIT CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative model training without sharing raw data, but repeated exchanges of high-dimensional model parameters incur substantial communication costs. Under heterogeneous data distributions, reducing these costs while maintaining effective cross-client collaboration and client-specific adaptation remains challenging. To address this challenge, we propose Federated ADMM with One-bit Consensus (FICO), a communication-efficient personalized federated learning framework. FICO relaxes exact parameter consensus to one-bit consensus over low-dimensional model representations, preserving client-specific variations while coordinating clients through a shared one-bit code. We develop a proximal ADMM algorithm that alternates between personalized model optimization and shared consensus code updates. Each client transmits one-bit signs of local cost differences together with a single amplitude scalar, allowing the server to select the shared code using compact messages while full-precision model variables remain local. Theoretical analysis establishes convergence to stationarity under suitable regularity and error conditions. Experimental results demonstrate that FICO achieves a more favorable trade-off between communication cost and personalized model accuracy than the evaluated baselines.

1 INTRODUCTION

Federated learning (FL) enables collaborative model training without collecting raw data centrally (Li et al., 2020a). Many existing methods pursue a shared-model objective that admits a *parameter-consensus* formulation: each client maintains a local parameter copy, and these copies are required to agree on a common solution. FedAvg (McMahan et al., 2017) approaches this objective through repeated model aggregation, while FedProx (Li et al., 2020b) and SCAFFOLD (Karimireddy et al., 2020) improve optimization under data heterogeneity through proximal regularization and client-drift correction, respectively. Federated methods based on the alternating direction method of multipliers (ADMM) (Boyd et al., 2011), such as FedADMM-InSa (Song et al., 2025), make the consensus structure explicit through equality constraints between local parameter copies and a shared variable. Although parameter consensus provides a direct mechanism for collaboration, a common model must accommodate potentially conflicting client objectives, and exchanging high-dimensional parameters or updates incurs substantial communication costs (Stich et al., 2018). These limitations motivate a fundamental question: *what must clients agree on to collaborate effectively while preserving personalization and communication efficiency?*

Personalized federated learning (PFL) relaxes parameter sharing to accommodate client-specific objectives (Tan et al., 2023). Regularization-based methods, including pFedMe (Dinh et al., 2020) and Ditto (Li et al., 2021a), encourage personalized models to remain close to a common model across the full parameter space while allowing their parameters to differ. FLAME (Zhu et al., 2024) applies ADMM to a formulation based on the Moreau envelope to jointly train personalized and global models. Other methods adopt *partial parameter consensus* by sharing selected model components and keeping the remaining components private (Arivazhagan et al., 2019; Li et al., 2021b). For example, FedRep (Collins et al., 2021) learns a shared representation network with client-specific prediction heads, while FedAPM (Zhu et al., 2025) uses ADMM to coordinate shared components alongside private parameter updates. These approaches provide flexibility through full-parameter regulariza-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion or partial parameter sharing. Nevertheless, these methods still coordinate clients through model parameter values, and repeatedly exchanging these parameters can be costly even when only part of the model is shared.

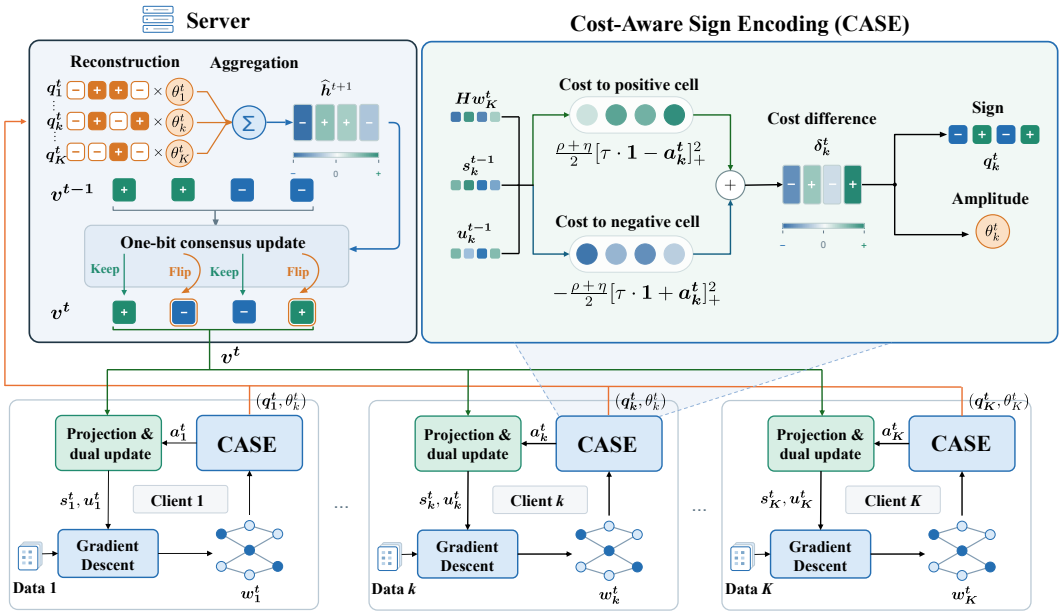

Figure 1: An overview of the proposed FICO framework. At each round, each client updates its personalized model and computes cost differences between different quantization cells. Then each client transmits the one-bit signs of the differences with one amplitude scalar to the server. The server reconstructs and aggregates the approximate cost differences to update the shared one-bit consensus code, which is broadcast to all clients. Each client then performs projection and dual updates for the next round.

Communication compression addresses this cost by reducing the information transmitted in each exchange (Wen et al., 2017; Karimireddy et al., 2019). QSGD (Alistarh et al., 2017) quantizes gradients, while signSGD (Bernstein et al., 2018) communicates their signs. In federated training, FedPAQ (Reisizadeh et al., 2020) combines periodic averaging with quantized model updates, and z-SignFedAvg (Tang et al., 2024) integrates stochastic sign compression with multiple local updates. These methods reduce message precision, but compression alone does not change the underlying shared-model objective. More closely related to our work, pFed1BS (Cheng et al., 2026) moves coordination to one-bit model sketches: the server constructs a shared one-bit consensus code through weighted majority voting, and clients use a sign-alignment regularizer for personalized training. This enables bidirectional compression while retaining distinct client models. Its code selection minimizes disagreement with the uploaded sketches, which convey model signs rather than explicit optimization-cost differences between candidate codes. This motivates incorporating both the direction and strength of local code preferences into low-bit coordination.

We propose *Federated ADMM with One-bit Consensus* (FICO) to coordinate personalized models through a shared quantization cell in a low-dimensional projection space as shown in Fig. 1. Unlike full or partial parameter consensus, FICO requires agreement on a shared one-bit consensus code while allowing distinct projection magnitudes and model parameters, preserving flexibility for local adaptation. Beyond sign voting over model sketches, FICO derives local cost differences between candidate signs from an ADMM subproblem and uses their compressed values to select the shared code. Each client uploads one sign bit per coordinate and one amplitude scalar, and the server broadcasts the shared code. This connects code selection to local optimization costs while enabling bidirectional low-bit communication and keeping full-precision models local.

Our main contributions are as follows:

- **From parameter consensus to one-bit consensus.** We formulate personalized federated learning through shared quantization-cell constraints. A jointly selected one-bit consensus code coordinates low-dimensional model projections while allowing continuous representations and model parameters to differ across clients, providing an explicit mechanism for collaboration with local flexibility.
- **Cost-driven consensus with one-bit communication.** We derive a proximal ADMM algorithm that reduces shared-code selection to candidate-sign cost comparisons and admits closed-form auxiliary projections. Its communication protocol conveys compressed cost differences using one bit per coordinate plus one amplitude scalar in each client's uplink, and broadcasts one bit per coordinate in the downlink.
- **Stationarity guarantees and accuracy–communication benefits.** Under the stated assumptions, we bound the average expected squared Karush–Kuhn–Tucker (KKT) residual for continuous variables within the selected cells by an $O(1/T)$ term after T communication rounds plus a stochastic-gradient error term. Under stronger summability conditions on the errors, we further establish vanishing residuals and stationarity of accumulation points. Experiments under heterogeneous client data demonstrate a favorable accuracy–communication trade-off relative to the evaluated personalized and communication-efficient baselines.

2 Problem Formulation

We consider a federated learning system with K clients. Each client $k \in \{1, \dots, K\}$ maintains a personalized model $\boldsymbol{w}_k \in \mathbb{R}^d$ and holds a local dataset $\mathcal{D}_k = \{\xi_{k,i}\}_{i=1}^{N_k}$, where d is the number of model parameters, $\xi_{k,i}$ is the i-th training sample, and $N_k = |\mathcal{D}_k|$ is the local sample count. The local objective of client k is to minimize the expected loss over its local data distribution $\mathcal{P}_k$:

$$f_k(\boldsymbol{w}_k) = \mathbb{E}_{\xi_k \sim \mathcal{P}_k}\left[\hat{f}_k(\boldsymbol{w}_k; \xi_k)\right]. \tag{1}$$

Here ξ_k is a random sample drawn from $\mathcal{P}_k$, $\mathbb{E}$ denotes expectation over the indicated randomness, and $\hat{f}_k(\boldsymbol{w}_k; \xi_k)$ is the sample-wise loss at $\boldsymbol{w}_k$. Throughout, $\{\boldsymbol{z}_k\}_{k=1}^K$, abbreviated as $\{\boldsymbol{z}_k\}_k$, denotes a family indexed by clients. For vectors of equal dimension, we arrange this family as the columns of a matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_K]$, whose k-th column corresponds to client k. The personalized federated learning objective is formulated as

$$\min_{\{\boldsymbol{w}_k\}_{k=1}^K} \sum_{k=1}^{K} p_k f_k(\boldsymbol{w}_k), \tag{2}$$

where $p_k = \frac{N_k}{\sum_{j=1}^K N_j}$ denotes the aggregation weight of client k.

In conventional federated ADMM formulations, a shared variable $\boldsymbol{z}$ is introduced to coordinate distributed models through *parameter consensus*:

$$\boldsymbol{w}_k = \boldsymbol{z}, \qquad k = 1, \dots, K. \tag{3}$$

Although parameter consensus enables effective collaboration among clients, forcing all local models to converge to an identical solution may suppress client-specific characteristics and limit personalization under heterogeneous data distributions.

To preserve model personalization while maintaining cross-client coordination, we relax parameter consensus to a weaker *one-bit consensus*. Specifically, let $\boldsymbol{H} \in \mathbb{R}^{m \times d}$ denote a shared transformation operator, where m is the representation dimension. We apply $\mathrm{sign}(\cdot)$ elementwise, with $\mathrm{sign}(v) = +1$ for $v \geq 0$ and -1 otherwise, so its output lies in $\{-1, +1\}^m$. Instead of requiring identical model parameters, one-bit consensus only requires the transformed representations of different clients to share the same one-bit code:

$$\mathrm{sign}(\boldsymbol{H}\boldsymbol{w}_k) = \mathrm{sign}(\boldsymbol{H}\boldsymbol{z}), \qquad k = 1, \dots, K. \tag{4}$$

This constraint allows different clients to maintain distinct full-precision models while preserving a common low-dimensional coordination signal. Since the coordination among clients depends only

on the one-bit representation of $\boldsymbol{H}\boldsymbol{z}$, rather than the full-precision variable $\boldsymbol{z}$ itself, we introduce the shared one-bit consensus variable

$$\boldsymbol{v} \triangleq \operatorname{sign}(\boldsymbol{H}\boldsymbol{z}) \in \{-1, +1\}^m. \tag{5}$$

The one-bit consensus constraint can then be equivalently rewritten as

$$\operatorname{sign}(\boldsymbol{H}\boldsymbol{w}_k) = \boldsymbol{v}, \qquad k = 1, \ldots, K. \tag{6}$$

To obtain an optimization formulation suitable for ADMM, we introduce an auxiliary transformed variable $\boldsymbol{s}_k \in \mathbb{R}^m$ satisfying

$$\boldsymbol{H}\boldsymbol{w}_k = \boldsymbol{s}_k, \qquad k = 1, \ldots, K. \tag{7}$$

For a given one-bit consensus code $\boldsymbol{v}$, we impose a positive margin $\tau > 0$ and define the closed convex consensus region as

$$\mathcal{C}_{\boldsymbol{v}}^{\tau} = \{\boldsymbol{s} \in \mathbb{R}^m : \boldsymbol{v} \odot \boldsymbol{s} \succeq \tau \mathbf{1}\}, \tag{8}$$

where $\odot$ denotes the elementwise product, $\succeq$ means coordinatewise inequality, and $\mathbf{1} \in \mathbb{R}^m$ is the all-ones vector. We use $\mathbf{0}$ for a zero vector or matrix of the required dimension. Membership in this region implies $\operatorname{sign}(\boldsymbol{s}) = \boldsymbol{v}$ and keeps each coordinate at least τ away from zero. We therefore strengthen sign agreement by requiring

$$\boldsymbol{s}_k \in \mathcal{C}_{\boldsymbol{v}}^{\tau}, \qquad k = 1, \ldots, K. \tag{9}$$

Accordingly, the personalized federated learning problem with one-bit consensus is formulated as

$$\begin{aligned} \min_{\{\boldsymbol{w}_k, \boldsymbol{s}_k\}_{k=1}^K, \boldsymbol{v}} \quad & \sum_{k=1}^{K} p_k f_k(\boldsymbol{w}_k) \\ \text{s.t.} \quad & \boldsymbol{H}\boldsymbol{w}_k = \boldsymbol{s}_k \in \mathcal{C}_{\boldsymbol{v}}^{\tau} \quad (k = 1, \ldots, K), \qquad \boldsymbol{v} \in \{-1, +1\}^m. \end{aligned} \tag{10}$$

The constraint $\boldsymbol{H}\boldsymbol{w}_k = \boldsymbol{s}_k$ connects each personalized model with its transformed representation, while $\mathcal{C}_{\boldsymbol{v}}^{\tau}$ coordinates different clients through the shared one-bit consensus code $\boldsymbol{v}$. Based on this formulation, we next develop a federated ADMM algorithm with one-bit consensus communication.

3 Proposed Method

In this section, we develop the proposed FICO algorithm for solving the one-bit consensus problem in (10). The key idea is to update the shared one-bit consensus code according to the local cost differences induced by different binary choices and exchange these optimization-related quantities through one-bit communication.

3.1 ADMM Subproblems and One-Bit Consensus Update

Using this column convention, let $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{d \times K}$, $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K] \in \mathbb{R}^{m \times K}$, and $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K] \in \mathbb{R}^{m \times K}$ collect the personalized models, auxiliary variables, and scaled dual variables, respectively, where $\boldsymbol{u}_k \in \mathbb{R}^m$ denotes the scaled dual variable associated with the constraint $\boldsymbol{H}\boldsymbol{w}_k = \boldsymbol{s}_k$. The scaled augmented Lagrangian of (10) is

$$\mathcal{L}_{\rho}(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{v}, \boldsymbol{U}) = \sum_{k=1}^{K} p_k \left(f_k(\boldsymbol{w}_k) + \iota_{\mathcal{C}_{\boldsymbol{v}}^{\tau}}(\boldsymbol{s}_k) + \frac{\rho}{2} \|\boldsymbol{H}\boldsymbol{w}_k - \boldsymbol{s}_k + \boldsymbol{u}_k\|_2^2 \right). \tag{11}$$

Here $\rho > 0$ is the penalty parameter, and the extended-valued indicator $\iota_{\mathcal{C}_{\boldsymbol{v}}^{\tau}}(\boldsymbol{s})$ is zero for $\boldsymbol{s} \in \mathcal{C}_{\boldsymbol{v}}^{\tau}$ and $+\infty$ otherwise.

The superscript t denotes the communication-round index, with $t = 0$ denoting initialization. At each round, FICO alternately updates the personalized models, the shared one-bit code, and the auxiliary variables. Given $\boldsymbol{s}_k^t$ and $\boldsymbol{u}_k^t$, client k updates its personalized model by solving

$$\boldsymbol{w}_k^{t+1} \approx \arg\min_{\boldsymbol{w}} f_k(\boldsymbol{w}) + \frac{\rho}{2} \|\boldsymbol{H}\boldsymbol{w} - \boldsymbol{s}_k^t + \boldsymbol{u}_k^t\|_2^2 + \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}_k^t\|_2^2, \tag{12}$$

where $\beta \geq 0$ is a proximal parameter. Each client approximately solves this subproblem by mini-batch SGD; the finite local updates are specified in Assumption A3.

After the local models have been updated, $\boldsymbol{S}$ and $\boldsymbol{v}$ remain coupled through the constraint $\boldsymbol{s}_k \in \mathcal{C}_{\boldsymbol{v}}^{\tau}$. Therefore, we consider the following subproblem:

$$\min_{\substack{\boldsymbol{v}\in\{-1,+1\}^m \\ \boldsymbol{s}_k\in\mathcal{C}_{\boldsymbol{v}}^{\tau},\ k=1,\dots,K}} \frac{\rho}{2}\sum_{k=1}^{K} p_k \|\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{u}_k^t - \boldsymbol{s}_k\|_2^2 + \frac{\eta}{2}\sum_{k=1}^{K} p_k \|\boldsymbol{s}_k - \boldsymbol{s}_k^t\|_2^2 + \chi d_{\mathrm{H}}(\boldsymbol{v}, \boldsymbol{v}^t), \tag{13}$$

where $\eta \geq 0$ is the auxiliary-variable proximal parameter, and $\chi \geq 0$ penalizes code changes measured by the Hamming distance $d_{\mathrm{H}}(\boldsymbol{v}, \boldsymbol{v}^t) = |\{j : v_j \neq v_j^t\}|$, where $|\cdot|$ denotes set cardinality.

For a fixed consensus code $\boldsymbol{v}$, the optimization of $\boldsymbol{s}_k$ can be separated among clients. By completing the square, the update of $\boldsymbol{s}_k$ can be rewritten as

$$\begin{aligned} \boldsymbol{s}_k^{t+1} &= \arg\min_{\boldsymbol{s}_k\in\mathcal{C}_{\boldsymbol{v}}^{\tau}} \frac{\rho}{2}\|\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{u}_k^t - \boldsymbol{s}_k\|_2^2 + \frac{\eta}{2}\|\boldsymbol{s}_k - \boldsymbol{s}_k^t\|_2^2 \\ &= \arg\min_{\boldsymbol{s}_k\in\mathcal{C}_{\boldsymbol{v}}^{\tau}} \frac{\rho+\eta}{2}\|\boldsymbol{s}_k - \boldsymbol{a}_k^{t+1}\|_2^2 + \mathrm{const}, \end{aligned} \tag{14}$$

where const is independent of the optimization variable $\boldsymbol{s}_k$, and the projection center $\boldsymbol{a}_k^{t+1} \in \mathbb{R}^m$ is given by

$$\boldsymbol{a}_k^{t+1} = \frac{\rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{u}_k^t) + \eta\boldsymbol{s}_k^t}{\rho+\eta}. \tag{15}$$

For a fixed one-bit consensus code $\boldsymbol{v}$, (14) is the Euclidean projection $\boldsymbol{s}_k^{t+1} = \Pi_{\mathcal{C}_{\boldsymbol{v}}^{\tau}}(\boldsymbol{a}_k^{t+1})$, defined for $\boldsymbol{x} \in \mathbb{R}^m$ by

$$\Pi_{\mathcal{C}_{\boldsymbol{v}}^{\tau}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{s}\in\mathcal{C}_{\boldsymbol{v}}^{\tau}} \|\boldsymbol{s} - \boldsymbol{x}\|_2^2. \tag{16}$$

Thus, $\boldsymbol{s}_k^{t+1}$ is the point in the selected consensus region closest to $\boldsymbol{a}_k^{t+1}$.

The above result indicates that the remaining optimization is to determine the shared consensus code $\boldsymbol{v}$. By eliminating the auxiliary variables through their optimal solutions, we obtain an equivalent binary optimization problem as follows:

$$\min_{\boldsymbol{v}\in\{-1,+1\}^m} \frac{1}{2}\langle \boldsymbol{h}^{t+1}, \boldsymbol{v}\rangle + \chi d_{\mathrm{H}}(\boldsymbol{v}, \boldsymbol{v}^t), \tag{17}$$

where $\boldsymbol{h}^{t+1} = \sum_{k=1}^{K} p_k \boldsymbol{\delta}_k^{t+1}$, and the j-th element of $\boldsymbol{\delta}_k^{t+1}$ is defined as

$$\delta_{k,j}^{t+1} = \frac{\rho+\eta}{2}\left([\tau - a_{k,j}^{t+1}]_+^2 - [\tau + a_{k,j}^{t+1}]_+^2\right). \tag{18}$$

Here $[v]_+ = \max\{v, 0\}$, $j \in \{1, \dots, m\}$ indexes representation coordinates, and $a_{k,j}^{t+1}$ is the j-th component of $\boldsymbol{a}_k^{t+1}$; we use analogous component notation for other vectors.

Particularly, the following lemma formalizes the equivalence between the original code update and the binary optimization problem, which is proved in Appendix D.1.

Lemma 1. *After eliminating the auxiliary variables $\{\boldsymbol{s}_k\}_{k=1}^K$, the joint optimization problem in* (13) *is equivalently reduced to the binary optimization problem in* (17)*.*

Problem (17) separates over coordinates. For coordinate j, retaining v_j^t has cost $h_j^{t+1}v_j^t/2$, whereas switching to $-v_j^t$ has cost $-h_j^{t+1}v_j^t/2+\chi$. The decrease obtained by switching is therefore $v_j^t h_j^{t+1} - \chi$. Resolving ties by retaining the previous sign gives the exact minimizer

$$v_j^{t+1,\star} = \begin{cases} -v_j^t, & v_j^t h_j^{t+1} > \chi, \\ v_j^t, & \text{otherwise}, \end{cases} \qquad j = 1, \dots, m. \tag{19}$$

3.2 One-Bit Communication of Consensus Information

Although only the cost difference vectors $\boldsymbol{\delta}_k^{t+1}$ are required for updating the shared consensus code, directly transmitting the full-precision vectors introduces considerable communication overhead. FICO therefore compresses $\boldsymbol{\delta}_k^{t+1}$ into one-bit messages while preserving the magnitude information through an additional scalar amplitude. Specifically, client k transmits the sign vector

$$\boldsymbol{q}_k^{t+1} = \mathrm{sign}(\boldsymbol{\delta}_k^{t+1}) \in \{-1, +1\}^m, \tag{20}$$

together with the nonnegative amplitude coefficient θ_k^{t+1}:

$$\theta_k^{t+1} = \arg\min_{\theta \geq 0} \|\boldsymbol{\delta}_k^{t+1} - \theta \boldsymbol{q}_k^{t+1}\|_2^2 = \frac{1}{m}\|\boldsymbol{\delta}_k^{t+1}\|_1. \tag{21}$$

The server reconstructs an approximate local cost difference vector as

$$\widehat{\boldsymbol{\delta}}_k^{t+1} = \theta_k^{t+1} \boldsymbol{q}_k^{t+1}, \tag{22}$$

and aggregates the compressed information as

$$\widehat{\boldsymbol{h}}^{t+1} = \sum_{k=1}^{K} p_k \widehat{\boldsymbol{\delta}}_k^{t+1}. \tag{23}$$

The server replaces $\boldsymbol{h}^{t+1}$ with $\widehat{\boldsymbol{h}}^{t+1}$ in (17) to obtain the compressed surrogate for code selection. We introduce a fixed tolerance $\epsilon_{\text{tie}} \geq 0$ for nearly tied decisions, measured in units of this surrogate objective. A coordinate changes only when the estimated decrease $v_j^t \widehat{h}_j^{t+1} - \chi$ exceeds ϵ_{tie}, giving the update

$$v_j^{t+1} = \begin{cases} -v_j^t, & v_j^t \widehat{h}_j^{t+1} > \chi + \epsilon_{\text{tie}}, \\ v_j^t, & \text{otherwise}, \end{cases} \qquad j = 1, \ldots, m. \tag{24}$$

When $\epsilon_{\text{tie}} = 0$, this is an exact minimizer of the compressed surrogate, with the old sign retained at equality. For $\epsilon_{\text{tie}} > 0$, an estimated improvement no greater than the tolerance leaves the old sign unchanged; equivalently, the rule exactly minimizes the compressed surrogate with switching penalty $\chi + \epsilon_{\text{tie}}$.

After receiving the updated consensus code $\boldsymbol{v}^{t+1}$, each client updates the auxiliary variable by projecting $\boldsymbol{a}_k^{t+1}$ onto the corresponding one-bit consensus region:

$$\boldsymbol{s}_k^{t+1} = \Pi_{\mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}}(\boldsymbol{a}_k^{t+1}). \tag{25}$$

The dual variable is then updated as

$$\boldsymbol{u}_k^{t+1} = \boldsymbol{u}_k^t + \boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}. \tag{26}$$

Through this design, FICO only communicates the one-bit representation of the consensus-related cost differences and one scalar amplitude, while the personalized models and full-precision auxiliary variables remain locally maintained.

To quantify this communication cost, let b_θ denote the number of bits used to represent the scalar amplitude. Each client uploads m sign bits and one b_θ-bit amplitude, and receives an m-bit consensus code. With all K clients participating and each broadcast counted once per recipient, the per-round uplink, downlink, and total communication costs in bits are

$$\begin{aligned} C_{\text{up}} &= K(m + b_\theta), \qquad C_{\text{down}} = Km, \\ C_{\text{round}} &= C_{\text{up}} + C_{\text{down}} = K(2m + b_\theta). \end{aligned} \tag{27}$$

For a 32-bit amplitude, the total cost is $K(2m + 32)$ bits, equivalent to $K(2m + 32)/(8 \times 1024^2)$ MB. The fixed sketching operator is generated locally from a shared seed and need not be transmitted each round.

3.3 Algorithm

FICO runs for T communication rounds, each consisting of three stages: (i) local personalized model optimization, (ii) server-side one-bit consensus code selection, and (iii) client-side projection and dual updates. We use a local SGD step size $\gamma > 0$, and let $N_{k,t}$ denote the number of local steps performed by client k in round t. The complete procedure is presented in Algorithm 1 in Appendix A.

4 THEORETICAL ANALYSIS

We establish a convergence bound for FICO using a KKT residual that measures feasibility and first-order stationarity of the continuous variables within the selected quantization cells.

Let $\boldsymbol{\lambda}_k^t = \rho \boldsymbol{u}_k^t$ and $\boldsymbol{\Lambda}^t = [\boldsymbol{\lambda}_1^t, \ldots, \boldsymbol{\lambda}_K^t] = \rho \boldsymbol{U}^t \in \mathbb{R}^{m \times K}$. For a fixed one-bit consensus code $\boldsymbol{v}$, the unaugmented Lagrangian is

$$\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\Lambda}) = \sum_k p_k \Big[f_k(\boldsymbol{w}_k) + \iota_{\mathcal{C}_{\boldsymbol{v}}^\tau}(\boldsymbol{s}_k) + \langle \boldsymbol{\lambda}_k, \boldsymbol{H}\boldsymbol{w}_k - \boldsymbol{s}_k \rangle \Big]. \tag{28}$$

For this fixed code, a KKT point $(\boldsymbol{W}^\star, \boldsymbol{S}^\star, \boldsymbol{\Lambda}^\star)$ satisfies

$$\boldsymbol{H}\boldsymbol{w}_k^\star = \boldsymbol{s}_k^\star \in \mathcal{C}_{\boldsymbol{v}}^\tau, \qquad \nabla f_k(\boldsymbol{w}_k^\star) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^\star = \boldsymbol{0}, \qquad \boldsymbol{\lambda}_k^\star \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}}^\tau}(\boldsymbol{s}_k^\star), \quad \forall k, \tag{29}$$

where $\mathcal{N}_{\mathcal{C}}(\boldsymbol{s})$ denotes the convex normal cone to $\mathcal{C}$ at $\boldsymbol{s} \in \mathcal{C}$. Initialization and projection ensure $\boldsymbol{s}_k^t \in \mathcal{C}_{\boldsymbol{v}^t}^\tau$ at every round. With $\operatorname{dist}(\boldsymbol{x}, \mathcal{C}) = \inf_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$ denoting Euclidean distance to a nonempty set, we define the nonnegative selected-cell KKT residual r_t through

$$r_t^2 = \sum_k p_k \Big[\|\boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t\|_2^2 + \|\nabla f_k(\boldsymbol{w}_k^t) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^t\|_2^2 + \operatorname{dist}^2\big(\boldsymbol{\lambda}_k^t, \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}^t}^\tau}(\boldsymbol{s}_k^t)\big) \Big]. \tag{30}$$

All assumptions for the result below are collected in Appendix B. Here b is the mini-batch size, $N_{\max}$ bounds the number of local steps, and $\bar{\sigma}^2 = \sum_k p_k \sigma_k^2$ is the weighted gradient-error parameter defined in Assumption A5.

Theorem 1 (Expected stationarity with fixed mini-batches)**.** *Under Assumptions A1–A6, there exist finite $c_0, c_1 \geq 0$, independent of T for the fixed run, such that*

$$\min_{1 \leq t \leq T} \mathbb{E}[r_t^2] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r_t^2] \leq \frac{c_0}{T} + \frac{c_1 N_{\max} \bar{\sigma}^2}{b}, \qquad T \geq 1. \tag{31}$$

Theorem 1 gives an $O(T^{-1}) + O(N_{\max}\bar{\sigma}^2/b)$ upper bound on the average expected squared KKT residual within the selected cells. The first term decreases with the number of communication rounds, while the second accounts for stochastic gradient errors. With a fixed batch size, this bound does not imply exact stationarity under persistent gradient noise. The constants are independent of T for the fixed run, but are not asserted to be uniform across batch sizes. The proof is provided in Appendix C.1. Under summable gradient-error budgets, Theorem 2 in Appendix C.2 further shows that, almost surely, $r_t \to 0$, the shared code changes finitely many times, and every accumulation point satisfies the KKT conditions for the final cell.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

5.1.1 DATASETS

We evaluate FICO on five datasets: MNIST (LeCun et al., 1998), Fashion-MNIST (FMNIST) (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR-10, and CIFAR-100. Each dataset is partitioned among 20 clients using a Dirichlet distribution (Hsu et al., 2019) with concentration parameters α_{Dir} = 0.1 and α_{Dir} = 0.5 to induce different levels of label distribution heterogeneity. All methods use identical data partitions and are evaluated on the clients' local test sets.

5.1.2 COMPARED METHODS

We compare FICO with eight representative baselines: FedAvg (McMahan et al., 2017), FedProto (Tan et al., 2022), FLAME (Zhu et al., 2024), FedAPM (Zhu et al., 2025), FedADMM-InSa (Song et al., 2025), PP-QADMM (Elgabli, 2025), z-SignFedAvg (Tang et al., 2024), and pFed1BS (Cheng et al., 2026). These baselines cover conventional federated learning, personalized collaboration, ADMM-based federated optimization, and communication compression. Among them, pFed1BS is particularly relevant to our work, as it also supports personalized training through bidirectional low-bit communication.

Table 1: Maximum test accuracy (%) under the Dirichlet non-IID setting with $\alpha_{\text{Dir}} = 0.1$.

Method	MNIST	FMNIST	CIFAR-10	SVHN	CIFAR-100
Full / Partial parameter consensus					
FedAvg	98.50 ± 0.10	96.41 ± 0.23	89.93 ± 2.74	93.65 ± 1.11	63.43 ± 1.65
FLAME	98.76 ± 0.19	97.23 ± 0.44	90.41 ± 3.25	95.11 ± 1.22	64.27 ± 1.32
FedAPM	99.03 ± 0.11	97.29 ± 0.53	91.75 ± 1.25	95.64 ± 1.02	71.24 ± 1.13
FedADMM-InSa	97.29 ± 1.10	85.02 ± 0.53	86.64 ± 3.37	94.92 ± 1.26	58.86 ± 1.29
Beyond parameter consensus					
FedProto	98.64 ± 0.27	96.87 ± 0.49	91.09 ± 2.59	95.11 ± 1.41	62.27 ± 1.18
PP-QADMM	90.65 ± 1.12	79.31 ± 2.93	22.29 ± 4.88	18.17 ± 2.55	21.21 ± 2.69
z-SignFedAvg	92.63 ± 0.28	77.37 ± 0.49	26.23 ± 1.97	29.23 ± 1.73	13.00 ± 1.12
pFed1BS	98.04 ± 0.47	95.30 ± 0.89	85.90 ± 3.85	78.75 ± 2.87	49.36 ± 1.59
FICO	98.60 ± 0.27	97.14 ± 0.61	92.50 ± 2.09	96.57 ± 0.79	66.41 ± 1.11

5.1.3 Implementation Details

All methods are implemented in PyTorch (Paszke et al., 2019), and experiments are conducted on an NVIDIA RTX 4090 GPU. All 20 clients participate in every communication round, with a local batch size of 64. For MNIST and FMNIST, we use a DNN and train for 100 communication rounds with 5 local epochs per round. We use VGG8 for SVHN and CIFAR-10 and VGG11 for CIFAR-100. Experiments on these three datasets run for 200 communication rounds with 15 local epochs per round.

For each method, we tune key hyperparameters based on the recommended settings and further refine them on the validation set when necessary, ensuring a fair comparison with competitive configurations. For FICO, a randomized Hadamard sketching operator (Tropp, 2011) is generated locally from a shared random seed and remains fixed throughout each run, so the projection matrix need not be transmitted in each round.

5.1.4 Evaluation Metrics

We use two primary metrics to evaluate the performance of the compared methods:

- **Maximum Accuracy.** For each run, we compute the average client test accuracy at each communication round, weighted by the sizes of the clients' test sets, and take its maximum over all rounds. We report the mean and standard deviation across three independent runs with different random seeds.
- **Communication Cost.** We define the per-round communication cost counts all uplink and downlink payload bits across participating clients, with each broadcast message counted once for each recipient.

5.2 Experimental Results

Tables 1 and 2 summarize the test accuracy results across five datasets under the two Dirichlet non-IID settings. The proposed FICO operates beyond parameter consensus, achieving competitive test accuracy relative to methods based on full or partial parameter consensus and outperforming other methods beyond parameter consensus in several evaluated settings. With $\alpha_{\text{Dir}} = 0.1$, FICO performs particularly well on CIFAR-10, SVHN and CIFAR-100 while remaining competitive on the other datasets. For $\alpha_{\text{Dir}} = 0.5$, FICO achieves higher test accuracy than pFed1BS, which also combines one-bit consensus with bidirectional communication compression for personalized learning. Specifically, FICO improves over pFed1BS by 11.98% and 13.13% on CIFAR-10 and CIFAR-100, respectively. This suggests that transmitting an additional amplitude scalar alongside one-bit signs helps preserve the overall strength of each client's preferences, allowing the server to account for differences in preference strength across clients during aggregation.

Table 2: Maximum test accuracy (%) under the Dirichlet non-IID setting with $\alpha_{\mathrm{Dir}} = 0.5$.

Method	MNIST	FMNIST	CIFAR-10	SVHN	CIFAR-100
Full / Partial parameter consensus					
FedAvg	97.31 ± 0.13	90.54 ± 1.27	79.11 ± 0.74	89.54 ± 0.81	50.47 ± 2.54
FLAME	97.22 ± 0.29	91.29 ± 0.72	77.04 ± 2.06	90.37 ± 0.26	41.66 ± 1.31
FedAPM	97.77 ± 0.14	91.90 ± 0.82	80.78 ± 2.58	90.73 ± 1.41	60.12 ± 3.25
FedADMM-InSa	97.68 ± 0.08	88.87 ± 0.35	75.57 ± 2.82	89.98 ± 1.42	55.23 ± 4.09
Beyond parameter consensus					
FedProto	96.15 ± 0.28	90.46 ± 0.97	76.80 ± 0.23	88.53 ± 0.41	40.64 ± 1.37
PP-QADMM	94.67 ± 0.20	85.76 ± 0.78	44.94 ± 3.79	34.60 ± 2.37	36.75 ± 0.71
z-SignFedAvg	95.10 ± 0.16	83.37 ± 0.34	36.89 ± 3.31	53.18 ± 4.58	17.70 ± 0.98
pFed1BS	94.58 ± 0.09	87.36 ± 1.14	68.09 ± 1.00	51.71 ± 2.91	30.80 ± 0.88
FICO	96.18 ± 0.19	91.28 ± 0.74	80.07 ± 0.40	91.42 ± 0.47	43.93 ± 1.13

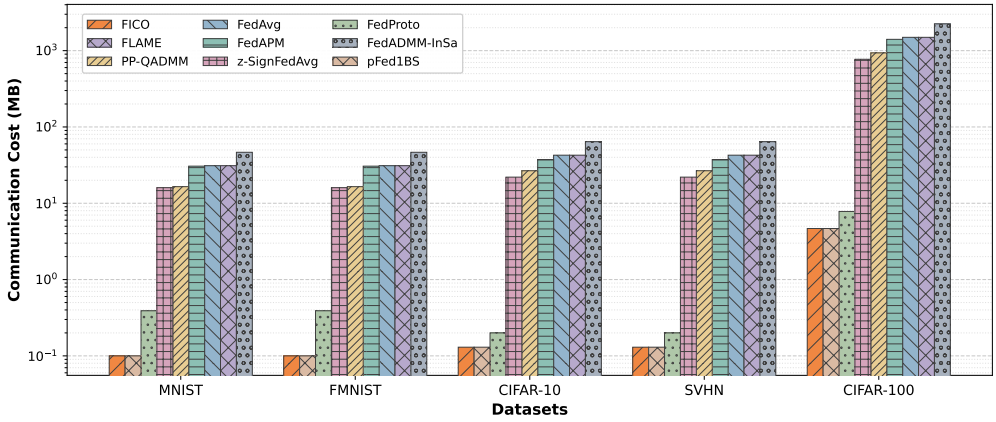

Figure 2: Comparison of communication cost for different algorithms across five datasets.

Fig. 2 compares communication costs of the evaluated methods across the five datasets during a single training round. Compared with FedADMM-InSa, the baseline with the highest per-round communication cost, FICO reduces the per-round communication cost by approximately 99.8% on the five datasets. These findings indicate that FICO achieves competitive personalized performance at a low communication cost.

6 Conclusion

In this paper, we propose FICO, a communication-efficient personalized federated learning framework based on ADMM. The proposed method preserves flexibility for local adaptation through shared quantization-cell constraints and uses approximate local cost differences to guide consensus selection. By avoiding repeated exchanges of full-precision models, FICO reduces communication costs in both directions. Our analysis bounds the average expected squared KKT residual by an $O(1/T)$ term plus a stochastic gradient error term. Under summable gradient-error budgets, we further establish vanishing KKT residuals and stationarity of accumulation points within the final quantization cell. Our experiments demonstrate that FICO achieves competitive test accuracy at a low communication cost.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and language polishing, and to support literature retrieval and the identification of related work. We did not use generative AI tools to generate synthetic datasets or to help develop theoretical models or conceptual frameworks. We reviewed and revised the AI-assisted text and verified the cited references against their original sources. We take full responsibility for the final content of this paper, including its claims, results, and references.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1709–1720, 2017.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Jiacheng Cheng, Xu Zhang, Guanghui Qiu, Yifang Zhang, Yinchuan Li, and Kaiyuan Feng. Personalized federated learning with bidirectional communication compression via one-bit random sketching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20499–20508, 2026.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099, 2021.

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with Moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Anis Elgabli. PP-QADMM: A dual-driven perturbation and quantized ADMM for privacy preserving and communication-efficient federated learning. *IEEE Open Journal of the Communications Society*, 6:4156–4175, 2025.

Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *NeurIPS Workshop on Federated Learning*, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020b.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368, 2021a.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on Non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2021–2031, 2020.

Yongcun Song, Ziqi Wang, and Enrique Zuazua. FedADMM-InSa: An inexact and self-adaptive ADMM for federated learning. *Neural Networks*, 181:106772, 2025.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022.

Zhiwei Tang, Yanmeng Wang, and Tsung-Hui Chang. z-SignFedAvg: A unified stochastic sign-based compression for federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15301–15309, 2024.

Joel A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1–2):115–126, 2011.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Shengkun Zhu, Jinshan Zeng, Sheng Wang, Yuan Sun, Xiaodong Li, Yuan Yao, and Zhiyong Peng. On ADMM in heterogeneous federated learning: Personalization, robustness, and fairness. *arXiv preprint arXiv:2407.16397*, 2024.

Shengkun Zhu, Feiteng Nie, Jinshan Zeng, Sheng Wang, Yuan Sun, Yuan Yao, Shangfeng Chen, Quanqing Xu, and Chuanhui Yang. FedAPM: Federated learning via ADMM with partial model personalization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4192–4202, 2025.

A ALGORITHM

Algorithm 1: FICO: Federated ADMM with One-bit Consensus

Input: communication rounds T, client set $\mathcal{K} = \{1, \dots, K\}$, ρ, β, η, χ, τ, $\epsilon_{\text{tie}} \geq 0$, $\boldsymbol{H}$, $\boldsymbol{v}^0$, step size γ, local step counts $\{N_{k,t}\}$

Client initialization: personalized model $\boldsymbol{w}_k^0$, $\boldsymbol{u}_k^0 = \mathbf{0}$, and $\boldsymbol{s}_k^0 \leftarrow \Pi_{\mathcal{C}^\tau_{\boldsymbol{v}^0}}(\boldsymbol{H}\boldsymbol{w}_k^0)$ for all $k \in \mathcal{K}$;

for $t = 0$ **to** $T - 1$ **do**

  foreach *client* $k \in \mathcal{K}$ ***in parallel*** **do**

    Starting at $\boldsymbol{w}_k^t$, take $N_{k,t}$ mini-batch SGD steps of size γ on the objective in (12) to obtain $\boldsymbol{w}_k^{t+1}$;

    Compute $\boldsymbol{a}_k^{t+1} \leftarrow \frac{\rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{u}_k^t) + \eta \boldsymbol{s}_k^t}{\rho + \eta}$;

    Compute $\boldsymbol{\delta}_k^{t+1} \leftarrow \frac{\rho+\eta}{2}\big([\tau\mathbf{1} - \boldsymbol{a}_k^{t+1}]_+^2 - [\tau\mathbf{1} + \boldsymbol{a}_k^{t+1}]_+^2\big)$;

    $\boldsymbol{q}_k^{t+1} \leftarrow \operatorname{sign}(\boldsymbol{\delta}_k^{t+1})$, $\theta_k^{t+1} \leftarrow \frac{1}{m}\|\boldsymbol{\delta}_k^{t+1}\|_1$;

    Upload $(\boldsymbol{q}_k^{t+1}, \theta_k^{t+1})$ to the server ; `// upload signs and one amplitude`

  Server-side code update:

  $\widehat{\boldsymbol{h}}^{t+1} \leftarrow \sum_{k=1}^K p_k \theta_k^{t+1} \boldsymbol{q}_k^{t+1}$;

  for $j = 1$ **to** m **do**

    $$v_j^{t+1} \leftarrow \begin{cases} -v_j^t, & v_j^t \widehat{h}_j^{t+1} > \chi + \epsilon_{\text{tie}}, \\ v_j^t, & \text{otherwise} \end{cases}$$

  Broadcast $\boldsymbol{v}^{t+1}$; `// shared one-bit code`

  Client-side projection and dual update:

  foreach *client* $k \in \mathcal{K}$ ***in parallel*** **do**

    $\boldsymbol{s}_k^{t+1} \leftarrow \Pi_{\mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}}\left(\boldsymbol{a}_k^{t+1}\right)$

    $\boldsymbol{u}_k^{t+1} \leftarrow \boldsymbol{u}_k^t + \boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}$

B ASSUMPTIONS FOR THE ANALYSIS

The analysis below rests on six assumptions, grouped by role. The first three describe the local objectives, the one-bit consensus cells, and the implemented local SGD updates. The next two control the iterates and mini-batch gradient errors; the last specifies the parameter conditions. Bounded iterates are assumed for the analyzed run rather than inferred from those parameter conditions. For a matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_K]$, we use the client-weighted Frobenius norm $\|\boldsymbol{Z}\|_{\mathrm{F},p}^2 = \sum_k p_k \|\boldsymbol{z}_k\|_2^2$, where $p_k > 0$ and $\sum_k p_k = 1$. The subscript p denotes the client weights; in particular, $\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 = \sum_k p_k \|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2^2$.

Assumption A1 (Local objectives)**.** Each f_k is continuously differentiable and ℓ_f-smooth, meaning that its gradient is Lipschitz with a common constant $\ell_f \geq 0$, and is bounded below by $\underline{f}_k > -\infty$. For some $\mu_f \geq 0$, $f_k + \mu_f \|\cdot\|_2^2/2$ is convex.

Assumption A2 (Communication and cells)**.** All clients participate, $p_k > 0$, and $\sum_k p_k = 1$. The fixed $\boldsymbol{H}$ has full row rank; write $\sigma_{\mathrm{H}} = \lambda_{\min}(\boldsymbol{H}\boldsymbol{H}^\top) > 0$ and $h_{\mathrm{op}} = \|\boldsymbol{H}\|_{\mathrm{op}}$, where $\lambda_{\min}$ denotes the smallest eigenvalue and $\|\cdot\|_{\mathrm{op}}$ the Euclidean induced operator norm. Initialization is deterministic and finite, with $\boldsymbol{s}_k^0 \in \mathcal{C}^\tau_{\boldsymbol{v}^0}$; $\tau, \chi > 0$ and $\epsilon_{\text{tie}} \geq 0$.

Assumption A3 (Finite local updates)**.** Each update in (12) starts at $\boldsymbol{w}_k^t$ and takes $N_{k,t}$ mini-batch SGD steps on its objective, using a fixed step size $\gamma > 0$ and no momentum. The step counts are deterministic, with $1 \leq N_{k,t} \leq N_{\max} < \infty$.

Assumption A4 (Stability)**.** There exist deterministic finite radii $r_{\mathrm{w}}, r_{\mathrm{s}}, r_\lambda \geq 0$ such that, almost surely for all $t \geq 0$,

$$\|\boldsymbol{W}^t\|_{\mathrm{F},p} \leq r_{\mathrm{w}}, \qquad \|\boldsymbol{S}^t\|_{\mathrm{F},p} \leq r_{\mathrm{s}}, \qquad \|\boldsymbol{\Lambda}^t\|_{\mathrm{F},p} \leq r_\lambda. \tag{B.1}$$

Assumption A5 (Gradient errors)**.** Every local step uses a fixed batch size $b \geq 1$. Let $\boldsymbol{\nu}_{k,i}^t = \boldsymbol{g}_{k,i}^t - \nabla f_k(\boldsymbol{x}_{k,i}^t)$, where $\boldsymbol{x}_{k,i}^t$ is the local model before step $i \in \{0, \dots, N_{k,t}-1\}$, starting from $\boldsymbol{x}_{k,0}^t = \boldsymbol{w}_k^t$,

and $\boldsymbol{g}_{k,i}^t$ is its mini-batch task gradient; $\boldsymbol{\nu}_{k,i}^t$ is the resulting gradient error. For deterministic finite $\sigma_k^2 \geq 0$,

$$\mathbb{E}\|\boldsymbol{\nu}_{k,i}^t\|_2^2 \leq \frac{\sigma_k^2}{b}, \qquad \bar{\sigma}^2 = \sum_k p_k \sigma_k^2. \tag{B.2}$$

Assumption A6 (Parameters and local computation)**.** Let $\kappa > 0$ be the relative local-solver accuracy parameter used in Lemma 3. Define $\mu = \beta - \mu_f$ and $\ell_\phi = \ell_f + \rho h_{\mathrm{op}}^2 + \beta$, the strong-convexity and gradient-Lipschitz constants of the local objective ϕ_k^t in (D.3), and set $c_{\mathrm{loc}} = \ell_f + 2\beta + 2\kappa$. The parameters and the bounds in Assumption A4 jointly satisfy

$$\beta > \mu_f, \quad \rho\sigma_{\mathrm{H}} > \max\{4\ell_f, 5c_{\mathrm{loc}}^2/\beta\}, \quad 0 < \gamma \leq 1/\ell_\phi, \quad N_{k,t} \geq N_{\min}, \quad \eta > \eta_{\min}, \tag{B.3}$$

where $N_{\min}$ and $\eta_{\min}$ are given in (D.57) and (D.62).

C Stationarity Theorems and Proofs

We first prove Theorem 1 using the stated descent and residual estimates in Lemma 2, followed by the vanishing-error result in Theorem 2. Appendix D proves all supporting lemmas. The logical dependency order is as follows: Lemma 1 is an algebraic reduction; Lemma 3 follows directly from the local SGD recurrence; Lemma 4 uses that local estimate and stability; these results establish Lemma 2, which in turn proves both theorems. We use the columnwise matrix notation and the client-weighted Frobenius norm defined in Section 2 and Appendix B, respectively. Write $f(\boldsymbol{W}) = \sum_k p_k f_k(\boldsymbol{w}_k)$ for the total objective. We state the optimality identities for each client and then sum them with weights p_k.

C.1 Expected Stationarity with Fixed Mini-Batches

The proof first controls the cumulative changes in the iterates through a descent inequality, and then bounds the KKT residual by those changes. We state the two estimates in Lemma 2 and prove the theorem before deriving the estimates in Appendix D.2.

Let the total squared gradient error in round t be

$$n_t = \sum_k p_k \sum_{i=0}^{N_{k,t}-1} \|\boldsymbol{\nu}_{k,i}^t\|_2^2. \tag{C.1}$$

Restore the dual-dependent constant omitted from the primal-block objective $\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{v}, \boldsymbol{U})$ and define the Lyapunov function ψ^t:

$$\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{v}, \boldsymbol{U}) = \mathcal{L}_\rho(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{v}, \boldsymbol{U}) - \frac{\rho}{2}\|\boldsymbol{U}\|_{\mathrm{F},p}^2, \tag{C.2}$$

$$\psi^t = \widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^t, \boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t) + \frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}}\|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 + 4\rho\|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2, \qquad t \geq 1. \tag{C.3}$$

The two additional terms compensate for the previous-round differences introduced by dual ascent.

We name the model and auxiliary decrease coefficients $a_{\mathrm{w}}, a_{\mathrm{s}}$, and the current- and previous-round error coefficients b_0, b_1:

$$a_{\mathrm{w}} = \frac{\beta}{4} - \frac{5c_{\mathrm{loc}}^2}{4\rho\sigma_{\mathrm{H}}}, \;\; a_{\mathrm{s}} = \frac{\eta - 16\rho}{4}, \;\; b_1 = \frac{20 b_{\mathrm{grad}}}{\rho\sigma_{\mathrm{H}}}, \;\; b_0 = b_{\mathrm{obj}} + b_1 + \frac{2a_{\mathrm{w}} b_{\mathrm{grad}}}{\beta^2 + 2\kappa^2}. \tag{C.4}$$

Here b_{obj} and b_{grad} are the fixed local-solver error coefficients defined in (D.4). The conditions $\rho\sigma_{\mathrm{H}} > 5c_{\mathrm{loc}}^2/\beta$ and $\eta > 16\rho$ imply $a_{\mathrm{w}}, a_{\mathrm{s}} > 0$. For the residual estimate, use the single fixed coefficient

$$c_{\mathrm{res}} = \max\left\{\frac{5(\ell_f + \beta + \kappa)c_{\mathrm{loc}}}{2\rho^2\sigma_{\mathrm{H}}} + 3(\beta+\kappa)^2,\; 4 + 3\rho^2 h_{\mathrm{op}}^2 + \eta^2,\; \left(\frac{20}{\rho^2\sigma_{\mathrm{H}}} + 3\right) b_{\mathrm{grad}}\right\}. \tag{C.5}$$

All these coefficients retain their specified values throughout the proof. Their derivation is given term by term in Appendix D.2.

Lemma 2 (Descent and residual bounds)**.** *Under Assumptions A1–A4 and A6, almost surely, simultaneously for every $t \geq 1$,*

$$\psi^{t+1} \leq \psi^t - a_{\mathrm{w}}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 - a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 - 4\tau^2 a_{\mathrm{s}} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t) + b_0 n_t + b_1 n_{t-1}, \tag{C.6}$$

and

$$r_{t+1}^2 \leq c_{\mathrm{res}}\big(\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 + \|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2 + n_t + n_{t-1}\big). \tag{C.7}$$

Moreover, $\psi^t \geq \psi_{\min} := \sum_k p_k \underline{f}_k - r_\lambda^2/(2\rho)$.

Proof of Theorem 1. **Step 1: Telescope the descent inequality.** The stability assumption and continuity of the objectives make ψ^t and the iterate differences integrable at each fixed round; Assumption A5 makes n_t integrable. For $T \geq 2$, move the decrease terms in (C.6) to the left, sum over $t = 1, \ldots, T-1$, and take expectations:

$$\begin{aligned} &a_{\mathrm{w}} \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + a_{\mathrm{s}} \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + 4\tau^2 a_{\mathrm{s}} \sum_{t=1}^{T-1} \mathbb{E} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t) \\ &\leq \sum_{t=1}^{T-1} \mathbb{E}(\psi^t - \psi^{t+1}) + b_0 \sum_{t=1}^{T-1} \mathbb{E} n_t + b_1 \sum_{t=0}^{T-2} \mathbb{E} n_t \\ &\leq \mathbb{E}\psi^1 - \psi_{\min} + (b_0 + b_1) \sum_{t=0}^{T-1} \mathbb{E} n_t. \end{aligned} \tag{C.8}$$

The energy sum telescopes to $\mathbb{E}\psi^1 - \mathbb{E}\psi^T$. The last line uses $\psi^T \geq \psi_{\min}$ and enlarges both nonnegative noise sums to $t = 0, \ldots, T-1$. In particular, after discarding the nonnegative code-change term,

$$a_{\mathrm{w}} \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + a_{\mathrm{s}} \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 \leq \mathbb{E}\psi^1 - \psi_{\min} + (b_0 + b_1) \sum_{t=0}^{T-1} \mathbb{E} n_t. \tag{C.9}$$

Step 2: Sum the residual bound and retain the boundary terms. For either $\boldsymbol{Z} = \boldsymbol{W}$ or $\boldsymbol{Z} = \boldsymbol{S}$, the exact change of index is

$$\begin{aligned} \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{Z}^t - \boldsymbol{Z}^{t-1}\|_{\mathrm{F},p}^2 &= \mathbb{E}\|\boldsymbol{Z}^1 - \boldsymbol{Z}^0\|_{\mathrm{F},p}^2 + \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_{\mathrm{F},p}^2 - \mathbb{E}\|\boldsymbol{Z}^T - \boldsymbol{Z}^{T-1}\|_{\mathrm{F},p}^2 \\ &\leq \mathbb{E}\|\boldsymbol{Z}^1 - \boldsymbol{Z}^0\|_{\mathrm{F},p}^2 + \sum_{t=1}^{T-1} \mathbb{E}\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_{\mathrm{F},p}^2. \end{aligned} \tag{C.10}$$

Sum (C.7) over $t = 1, \ldots, T-1$ and include the first residual:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E} r_t^2 \leq \mathbb{E} r_1^2 + c_{\mathrm{res}} \sum_{t=1}^{T-1} \mathbb{E}\big(&\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 \\ &+ \|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2\big) + c_{\mathrm{res}} \left(\sum_{t=1}^{T-1} \mathbb{E} n_t + \sum_{t=0}^{T-2} \mathbb{E} n_t \right). \end{aligned} \tag{C.11}$$

Apply (C.10) to the two previous-round sums. Each subsequent squared difference is then counted twice, while the initial terms remain. Enlarging the two error sums gives

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E} r_t^2 \leq \mathbb{E} r_1^2 + c_{\mathrm{res}}\big(&\mathbb{E}\|\boldsymbol{W}^1 - \boldsymbol{W}^0\|_{\mathrm{F},p}^2 + \mathbb{E}\|\boldsymbol{S}^1 - \boldsymbol{S}^0\|_{\mathrm{F},p}^2\big) \\ &+ 2c_{\mathrm{res}} \sum_{t=1}^{T-1} \mathbb{E}\big(\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2\big) + 2c_{\mathrm{res}} \sum_{t=0}^{T-1} \mathbb{E} n_t. \end{aligned} \tag{C.12}$$

Step 3: Use the descent bound and identify the theorem constants. For nonnegative v_1, v_2, positivity of $a_{\mathrm{w}}, a_{\mathrm{s}}$ gives

$$v_1 + v_2 \le \max\{a_{\mathrm{w}}^{-1}, a_{\mathrm{s}}^{-1}\}(a_{\mathrm{w}} v_1 + a_{\mathrm{s}} v_2). \tag{C.13}$$

Apply this inequality to the two subsequent squared-difference sums in (C.12), and then use (C.9). Thus

$$\sum_{t=1}^{T} \mathbb{E} r_t^2 \le c_0 + c_1 \sum_{t=0}^{T-1} \mathbb{E} n_t, \tag{C.14}$$

where the fixed constants are explicitly

$$c_0 = \mathbb{E} r_1^2 + c_{\mathrm{res}}\big(\mathbb{E}\|\boldsymbol{W}^1 - \boldsymbol{W}^0\|_{\mathrm{F},p}^2 + \mathbb{E}\|\boldsymbol{S}^1 - \boldsymbol{S}^0\|_{\mathrm{F},p}^2\big) + 2c_{\mathrm{res}} \max\{a_{\mathrm{w}}^{-1}, a_{\mathrm{s}}^{-1}\}(\mathbb{E}\psi^1 - \psi_{\min}), \tag{C.15}$$

$$c_1 = 2c_{\mathrm{res}} \max\{a_{\mathrm{w}}^{-1}, a_{\mathrm{s}}^{-1}\}(b_0 + b_1) + 2c_{\mathrm{res}}. \tag{C.16}$$

All quantities are finite. To check the first residual explicitly, the third identity in (D.15), applied to the first update, gives $\operatorname{dist}(\boldsymbol{\lambda}_k^1, \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}^1}^\tau}(\boldsymbol{s}_k^1)) \le \eta \|\boldsymbol{s}_k^1 - \boldsymbol{s}_k^0\|_2$. Combining this bound with the residual definition (30), smoothness in Assumption A1, and stability in (B.1), and using $\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 \le 2\|\boldsymbol{x}\|_2^2 + 2\|\boldsymbol{y}\|_2^2$, yields

$$\mathbb{E} r_1^2 \le 4 \sum_k p_k \|\nabla f_k(\mathbf{0})\|_2^2 + (2h_{\mathrm{op}}^2 + 4\ell_f^2) r_{\mathrm{w}}^2 + (2 + 4\eta^2) r_{\mathrm{s}}^2 + 2h_{\mathrm{op}}^2 r_\lambda^2. \tag{C.17}$$

The initial differences and the terms in ψ^1 are bounded by stability and smoothness. Thus c_0, c_1 are finite and independent of T for the fixed run. Since $c_0 \ge \mathbb{E} r_1^2$, (C.14) also holds for $T = 1$.

Step 4: Apply the mini-batch error assumption. By the definition (C.1), the moment bound (B.2), and the step-count bound in Assumption A3,

$$\sum_{t=0}^{T-1} \mathbb{E} n_t = \sum_{t=0}^{T-1} \sum_k p_k \sum_{i=0}^{N_{k,t}-1} \mathbb{E}\|\boldsymbol{\nu}_{k,i}^t\|_2^2 \le \sum_{t=0}^{T-1} \sum_k p_k \frac{N_{k,t}\sigma_k^2}{b} \le \frac{T N_{\max} \bar{\sigma}^2}{b}. \tag{C.18}$$

Substitute this into (C.14) and divide by T. The minimum is no greater than the average, proving (31). □

C.2 Stationarity under Vanishing Gradient Errors

The proof of Theorem 1 uses the fixed-batch condition only in Step 4. Thus (C.14) also applies to other error schedules. For deterministic nonnegative bounds $\mathbb{E}\|\boldsymbol{\nu}_{k,i}^t\|_2^2 \le \omega_{k,t,i}$, define

$$\omega_t = \sum_k p_k \sum_{i=0}^{N_{k,t}-1} \omega_{k,t,i}, \qquad \mathbb{E} n_t \le \omega_t < \infty. \tag{C.19}$$

Theorem 2 (Vanishing and summable errors)**.** *Suppose Assumptions A1–A4 and A6 hold. Replace Assumption A5 by deterministic nonnegative bounds $\mathbb{E}\|\boldsymbol{\nu}_{k,i}^t\|_2^2 \le \omega_{k,t,i}$ with finite ω_t in* (C.19)*. Then*

$$\min_{1 \le t \le T} \mathbb{E} r_t^2 \le \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} r_t^2 \le \frac{c_0}{T} + \frac{c_1}{T} \sum_{t=0}^{T-1} \omega_t, \qquad T \ge 1, \tag{C.20}$$

for finite c_0, c_1 independent of T. If $T^{-1}\sum_{t<T} \omega_t \to 0$, the average expected squared residual tends to zero. If $\sum_{t=0}^{\infty} \omega_t < \infty$, then almost surely $r_t \to 0$, the code changes finitely many times, and $f(\boldsymbol{W}^t)$ converges. Every accumulation point satisfies

$$\boldsymbol{H}\boldsymbol{w}_k^\star = \boldsymbol{s}_k^\star \in \mathcal{C}_{\boldsymbol{v}^\star}^\tau, \quad \nabla f_k(\boldsymbol{w}_k^\star) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^\star = \mathbf{0}, \quad \boldsymbol{\lambda}_k^\star \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}^\star}^\tau}(\boldsymbol{s}_k^\star), \quad k = 1, \dots, K, \tag{C.21}$$

where $\boldsymbol{v}^\star$ is the final code.

Proof. Replacing $\mathbb{E} n_t$ by its upper bound ω_t in (C.14), and dividing by T, gives (C.20) with the same explicit c_0, c_1 in (C.15)–(C.16). If $T^{-1}\sum_{t<T} \omega_t \to 0$, both terms on the right tend to zero.

Suppose now that $\sum_t \omega_t < \infty$. Since all n_t are nonnegative, Tonelli's theorem gives

$$\mathbb{E}\left[\sum_{t=0}^{\infty} n_t\right] = \sum_{t=0}^{\infty} \mathbb{E} n_t \le \sum_{t=0}^{\infty} \omega_t < \infty. \tag{C.22}$$

Thus $\sum_t n_t < \infty$ almost surely. Fix an outcome on this event where the stability bounds also hold.

Summability of the changes. Add the exact remaining noise tail from (C.6) to the Lyapunov function:

$$\widehat{\psi}^t = \psi^t + b_0 \sum_{\ell=t}^{\infty} n_\ell + b_1 \sum_{\ell=t}^{\infty} n_{\ell-1}, \qquad t \geq 1. \tag{C.23}$$

Both series are finite. Passing from t to $t+1$ removes exactly the two error terms on the right of (C.6). Consequently,

$$\widehat{\psi}^{t+1} \leq \widehat{\psi}^t - a_{\mathrm{w}} \|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 - a_{\mathrm{s}} \|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 - 4\tau^2 a_{\mathrm{s}} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t). \tag{C.24}$$

By Lemma 2, $\psi^t \geq \psi_{\min}$; the added tails are nonnegative. Thus (C.24) makes $\widehat{\psi}^t$ nonincreasing and bounded below, so it converges. Summing the preceding inequality and taking the limit gives

$$\begin{aligned} a_{\mathrm{w}} \sum_{t=1}^{\infty} \|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + a_{\mathrm{s}} \sum_{t=1}^{\infty} \|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 \\ + 4\tau^2 a_{\mathrm{s}} \sum_{t=1}^{\infty} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t) \leq \widehat{\psi}^1 - \psi_{\min} < \infty. \end{aligned} \tag{C.25}$$

All three coefficients are strictly positive. Each of the three series is therefore finite. In particular, $\boldsymbol{W}^{t+1} - \boldsymbol{W}^t \to \boldsymbol{0}$ and $\boldsymbol{S}^{t+1} - \boldsymbol{S}^t \to \boldsymbol{0}$. Each nonzero Hamming distance is at least one, so only finitely many code changes occur. Denote the final code by $\boldsymbol{v}^\star$. The tails in (C.23) tend to zero, so ψ^t converges to the same limit as $\widehat{\psi}^t$.

Residual and objective convergence. The previous-round squared differences are summable by the same index shift as (C.10), now applied pathwise to (C.25); the initial difference is finite by stability. The noise terms are summable by (C.22). Thus summing (D.21) shows $\sum_{t\geq 1} \|\boldsymbol{\Lambda}^{t+1} - \boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2 < \infty$. Every term on the right of the explicit bound (C.7) is summable, which gives $\sum_{t\geq 1} r_{t+1}^2 < \infty$ and hence $r_t \to 0$. Expand (C.2) in the energy definition (C.3); feasibility of $\boldsymbol{S}^t$ removes the indicators and gives

$$\begin{aligned} \psi^t - f(\boldsymbol{W}^t) = \sum_k p_k \langle \boldsymbol{\lambda}_k^t, \boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t \rangle + \frac{\rho}{2} \sum_k p_k \|\boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t\|_2^2 \\ + \frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}} \|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 + 4\rho \|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2. \end{aligned} \tag{C.26}$$

The quadratic feasibility term tends to zero by (30); the two increment terms tend to zero by (C.25). For the first term, weighted Cauchy–Schwarz and (B.1) give $|\sum_k p_k \langle \boldsymbol{\lambda}_k^t, \boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t \rangle| \leq r_\lambda \left(\sum_k p_k \|\boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t\|_2^2\right)^{1/2} \to 0$. Thus $\psi^t - f(\boldsymbol{W}^t) \to 0$, and convergence of ψ^t implies convergence of $f(\boldsymbol{W}^t)$.

Accumulation points. Assumption A4, $p_{\min} > 0$, and finite dimensionality ensure boundedness and hence existence of accumulation points. For any such point, choose a convergent subsequence after the final code change:

$$(\boldsymbol{W}^{t_j}, \boldsymbol{S}^{t_j}, \boldsymbol{\Lambda}^{t_j}) \to (\boldsymbol{W}^\star, \boldsymbol{S}^\star, \boldsymbol{\Lambda}^\star). \tag{C.27}$$

Since $p_k > 0$, the residual definition gives, for every k,

$$\|\boldsymbol{H}\boldsymbol{w}_k^{t_j} - \boldsymbol{s}_k^{t_j}\|_2 \leq \frac{r_{t_j}}{\sqrt{p_k}} \to 0, \qquad \|\nabla f_k(\boldsymbol{w}_k^{t_j}) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^{t_j}\|_2 \leq \frac{r_{t_j}}{\sqrt{p_k}} \to 0. \tag{C.28}$$

Continuity of the linear map $\boldsymbol{H}$ and of ∇f_k under Assumption A1 permits passage to the limit, proving primal feasibility and model stationarity. Closedness of $\mathcal{C}_{\boldsymbol{v}^\star}^\tau$ gives $\boldsymbol{s}_k^\star \in \mathcal{C}_{\boldsymbol{v}^\star}^\tau$. Finally, the third identity in (D.15), eventual constancy of the code, and vanishing auxiliary increments from (C.25) imply

$$\boldsymbol{\lambda}_k^{t_j} - \eta(\boldsymbol{s}_k^{t_j} - \boldsymbol{s}_k^{t_j-1}) \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}^\star}^\tau}(\boldsymbol{s}_k^{t_j}), \qquad \boldsymbol{\lambda}_k^{t_j} - \eta(\boldsymbol{s}_k^{t_j} - \boldsymbol{s}_k^{t_j-1}) \to \boldsymbol{\lambda}_k^\star. \tag{C.29}$$

For any $\boldsymbol{z} \in \mathcal{C}_{\boldsymbol{v}^\star}^\tau$, the associated normal-cone inequality passes to the limit and gives $\langle \boldsymbol{\lambda}_k^\star, \boldsymbol{z} - \boldsymbol{s}_k^\star \rangle \leq 0$. This is precisely $\boldsymbol{\lambda}_k^\star \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{v}^\star}^\tau}(\boldsymbol{s}_k^\star)$ and proves (C.21). □

A fixed batch size alone does not guarantee a vanishing or summable error budget. The result establishes stationary accumulation points; it does not establish convergence of the entire continuous-state sequence.

A sufficient sampling condition. Let $\boldsymbol{\zeta}_\ell \in \mathbb{R}^d$ be the gradient error from the ℓ-th sample in the current batch, and let $\mathscr{F}_{k,t,i}$ be the σ-algebra generated by the history before this gradient call. Suppose $\boldsymbol{\nu}_{k,i}^t = b^{-1}\sum_{\ell=1}^b \boldsymbol{\zeta}_\ell$ and, conditional on $\mathscr{F}_{k,t,i}$, these errors are independent, have mean zero, and have second moments at most σ_k^2. Then

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\nu}_{k,i}^t\|_2^2 \mid \mathscr{F}_{k,t,i}] &= \frac{1}{b^2}\sum_{\ell=1}^b \mathbb{E}[\|\boldsymbol{\zeta}_\ell\|_2^2 \mid \mathscr{F}_{k,t,i}] + \frac{2}{b^2}\sum_{\ell<j}\mathbb{E}[\langle \boldsymbol{\zeta}_\ell, \boldsymbol{\zeta}_j\rangle \mid \mathscr{F}_{k,t,i}] \\ &= \frac{1}{b^2}\sum_{\ell=1}^b \mathbb{E}[\|\boldsymbol{\zeta}_\ell\|_2^2 \mid \mathscr{F}_{k,t,i}] \le \frac{\sigma_k^2}{b}.\end{aligned} \tag{C.30}$$

The cross terms vanish by conditional independence and zero conditional means. Taking expectations proves Assumption A5. The convergence argument itself only uses the stated mean-square bound.

D Supporting Lemmas and Proofs

We now prove the lemmas used in Appendix C.

D.1 Equivalence of the Binary Code Update

For a fixed consensus code $\boldsymbol{v}$, the optimal auxiliary variable is obtained from (14). Substituting this optimal solution into (13) eliminates the variables $\{\boldsymbol{s}_k\}_{k=1}^K$ and leaves an optimization problem with respect to $\boldsymbol{v}$ only.

For each coordinate j, $v_j = +1$ and $v_j = -1$ correspond to two different feasible half-lines $[\tau, \infty)$ and $(-\infty, -\tau]$ for $s_{k,j}$, by (8). The minimum squared distances to these two regions are $[\tau - a_{k,j}^{t+1}]_+^2$ and $[\tau + a_{k,j}^{t+1}]_+^2$, respectively. Therefore, the difference between the two candidate costs is exactly $\delta_{k,j}^{t+1}$ defined in (18).

For a binary variable $v_j \in \{-1, +1\}$, the cost associated with this coordinate equals half the sum of the two candidate costs plus $v_j/2$ times their difference. Thus it can be decomposed into a term depending on v_j and a constant term independent of v_j:

$$\frac{1}{2}\delta_{k,j}^{t+1} v_j + \text{const.} \tag{D.1}$$

Summing over all coordinates and clients yields

$$\frac{1}{2}\left\langle \sum_{k=1}^K p_k \boldsymbol{\delta}_k^{t+1}, \boldsymbol{v}\right\rangle = \frac{1}{2}\langle \boldsymbol{h}^{t+1}, \boldsymbol{v}\rangle. \tag{D.2}$$

Combining this term with the code variation penalty gives (17), which completes the proof.

D.2 Descent and KKT Residual Bounds

Using $\boldsymbol{\lambda}_k^t = \rho \boldsymbol{u}_k^t$, or equivalently $\boldsymbol{u}_k^t = \boldsymbol{\lambda}_k^t/\rho$, we write the local objective and its terminal gradient as

$$\begin{aligned}\phi_k^t(\boldsymbol{x}) &= f_k(\boldsymbol{x}) + \frac{\rho}{2}\left\|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{s}_k^t + \frac{\boldsymbol{\lambda}_k^t}{\rho}\right\|_2^2 + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{w}_k^t\|_2^2 \\ &= f_k(\boldsymbol{x}) + \langle \boldsymbol{\lambda}_k^t, \boldsymbol{H}\boldsymbol{x} - \boldsymbol{s}_k^t\rangle + \frac{\rho}{2}\|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{s}_k^t\|_2^2 + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{w}_k^t\|_2^2 + \frac{\|\boldsymbol{\lambda}_k^t\|_2^2}{2\rho}, \\ \boldsymbol{e}_k^{t+1} &= \nabla\phi_k^t(\boldsymbol{w}_k^{t+1}).\end{aligned} \tag{D.3}$$

The last term in the expanded objective is independent of $\boldsymbol{x}$ and therefore contributes neither to the terminal gradient nor to differences of objective values within round t. We use two fixed local-solver coefficients, one for the objective error and one for the squared gradient error:

$$b_{\text{obj}} = \frac{\gamma}{2}, \qquad b_{\text{grad}} = \frac{\gamma(\ell_\phi + \kappa)^2}{\mu}. \tag{D.4}$$

Appendix D.3 derives both coefficients directly from the local SGD update. These coefficients keep the same values at every occurrence.

For the local errors below, let $\boldsymbol{E}^{t+1} = [\boldsymbol{e}_1^{t+1}, \dots, \boldsymbol{e}_K^{t+1}] \in \mathbb{R}^{d\times K}$ be the local residual matrix and let $\boldsymbol{\varepsilon}^{t+1} = (\varepsilon_1^{t+1}, \dots, \varepsilon_K^{t+1})^\top$ collect the scalar error bounds defined in (D.5), with $\|\boldsymbol{\varepsilon}^{t+1}\|_p^2 = \sum_k p_k(\varepsilon_k^{t+1})^2$.

The following estimate is proved in Appendix D.3 directly from the local SGD recurrence, without using Lemma 2.

Lemma 3 (Local residual and descent)**.** *Under Assumptions A1–A3 and the local conditions in Assumption A6, define*

$$\varepsilon_k^{t+1} = \left(b_{\mathrm{grad}} \sum_{i=0}^{N_{k,t}-1} \|\boldsymbol{\nu}_{k,i}^t\|_2^2 \right)^{1/2}. \tag{D.5}$$

For every round $t \geq 0$, pathwise,

$$\begin{aligned} \|\boldsymbol{e}_k^{t+1}\|_2 \leq \kappa\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + \varepsilon_k^{t+1}, \quad \|\boldsymbol{\varepsilon}^{t+1}\|_p^2 = b_{\mathrm{grad}} n_t, \\ \sum_k p_k \phi_k^t(\boldsymbol{w}_k^{t+1}) \leq \sum_k p_k \phi_k^t(\boldsymbol{w}_k^t) + b_{\mathrm{obj}} n_t. \end{aligned} \tag{D.6}$$

The coefficients are those in (D.4)*.*

Proof of Lemma 2. **Step 1: Derive the optimality identities.** Work on the probability-one event in Assumption A4. Differentiating the local objective in (D.3) gives

$$\boldsymbol{e}_k^{t+1} = \nabla f_k(\boldsymbol{w}_k^{t+1}) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^t + \rho \boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^t) + \beta(\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t). \tag{D.7}$$

Multiplying the scaled dual update (26) by ρ gives $\boldsymbol{\lambda}_k^{t+1} = \boldsymbol{\lambda}_k^t + \rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1})$. Hence $\boldsymbol{\lambda}_k^t + \rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^t) = \boldsymbol{\lambda}_k^{t+1} + \rho(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t)$. The update (25) projects onto the nonempty closed convex cell (8). Its first-order optimality condition, obtained from (16), is

$$\langle \boldsymbol{a}_k^{t+1} - \boldsymbol{s}_k^{t+1}, \boldsymbol{z} - \boldsymbol{s}_k^{t+1} \rangle \leq 0, \qquad \boldsymbol{z} \in \mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}. \tag{D.8}$$

Since $\rho + \eta > 0$, multiplying the variational inequality by this scalar preserves its direction:

$$\big\langle (\rho+\eta)(\boldsymbol{a}_k^{t+1} - \boldsymbol{s}_k^{t+1}), \boldsymbol{z} - \boldsymbol{s}_k^{t+1} \big\rangle \leq 0, \qquad \boldsymbol{z} \in \mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}. \tag{D.9}$$

First, substitute $\rho\boldsymbol{u}_k^t = \boldsymbol{\lambda}_k^t$ into the projection-center definition (15):

$$\boldsymbol{a}_k^{t+1} = \frac{\rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{u}_k^t) + \eta \boldsymbol{s}_k^t}{\rho+\eta} = \frac{\rho\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{\lambda}_k^t + \eta\boldsymbol{s}_k^t}{\rho+\eta}. \tag{D.10}$$

Hence the vector in the first argument of the inner product is

$$\begin{aligned} (\rho+\eta)(\boldsymbol{a}_k^{t+1} - \boldsymbol{s}_k^{t+1}) &= (\rho+\eta)\boldsymbol{a}_k^{t+1} - (\rho+\eta)\boldsymbol{s}_k^{t+1} \\ &= \rho\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \boldsymbol{\lambda}_k^t + \eta\boldsymbol{s}_k^t - \rho\boldsymbol{s}_k^{t+1} - \eta\boldsymbol{s}_k^{t+1} \\ &= \boldsymbol{\lambda}_k^t + \rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}) - \eta(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t). \end{aligned} \tag{D.11}$$

Next, the scaled dual update and the same relation between $\boldsymbol{u}$ and $\boldsymbol{\lambda}$ give

$$\boldsymbol{\lambda}_k^{t+1} = \rho\boldsymbol{u}_k^{t+1} = \rho\big(\boldsymbol{u}_k^t + \boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}\big) = \boldsymbol{\lambda}_k^t + \rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}). \tag{D.12}$$

Substituting this identity into the preceding expansion yields

$$(\rho+\eta)(\boldsymbol{a}_k^{t+1} - \boldsymbol{s}_k^{t+1}) = \boldsymbol{\lambda}_k^{t+1} - \eta(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t). \tag{D.13}$$

Therefore the multiplied variational inequality becomes

$$\big\langle \boldsymbol{\lambda}_k^{t+1} - \eta(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t), \boldsymbol{z} - \boldsymbol{s}_k^{t+1} \big\rangle \leq 0, \qquad \boldsymbol{z} \in \mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}, \tag{D.14}$$

which is exactly the normal-cone inclusion in the last line below. Together these relations imply, for every client $k = 1, \dots, K$,

$$\begin{aligned} \boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1} &= (\boldsymbol{\lambda}_k^{t+1} - \boldsymbol{\lambda}_k^t)/\rho, \\ \nabla f_k(\boldsymbol{w}_k^{t+1}) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^{t+1} &= \boldsymbol{e}_k^{t+1} - \beta(\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t) - \rho\boldsymbol{H}^\top(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t), \\ \boldsymbol{\lambda}_k^{t+1} - \eta(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t) &\in \mathcal{N}_{\mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}}(\boldsymbol{s}_k^{t+1}). \end{aligned} \tag{D.15}$$

We use $\mathcal{N}_{\mathcal{C}}(\boldsymbol{s}) = \{\boldsymbol{y} : \langle \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{s}\rangle \le 0 \text{ for all } \boldsymbol{z} \in \mathcal{C}\}$ when $\boldsymbol{s} \in \mathcal{C}$, and $\mathcal{N}_{\mathcal{C}}(\boldsymbol{s}) = \varnothing$ otherwise.

Step 2: Bound the dual increment with explicit coefficients. Set $\boldsymbol{P} = (\boldsymbol{H}\boldsymbol{H}^\top)^{-1}\boldsymbol{H} \in \mathbb{R}^{m\times d}$. With $\boldsymbol{I}_m$ denoting the $m \times m$ identity matrix, Assumption A2 gives $\boldsymbol{P}\boldsymbol{H}^\top = \boldsymbol{I}_m$ and $\boldsymbol{P}\boldsymbol{P}^\top = (\boldsymbol{H}\boldsymbol{H}^\top)^{-1}$, hence $\|\boldsymbol{P}\|_{\mathrm{op}}^2 = 1/\sigma_{\mathrm{H}}$. Subtract the model identity in (D.15) at rounds t and $t-1$ and apply $\boldsymbol{P}$:

$$\begin{aligned}\boldsymbol{\lambda}_k^{t+1} - \boldsymbol{\lambda}_k^t = \boldsymbol{P}\Big[&- (\nabla f_k(\boldsymbol{w}_k^{t+1}) - \nabla f_k(\boldsymbol{w}_k^t)) - \beta(\boldsymbol{w}_k^{t+1} - 2\boldsymbol{w}_k^t + \boldsymbol{w}_k^{t-1}) + \boldsymbol{e}_k^{t+1} - \boldsymbol{e}_k^t\Big] \\ &- \rho(\boldsymbol{s}_k^{t+1} - 2\boldsymbol{s}_k^t + \boldsymbol{s}_k^{t-1}), \qquad t \ge 1.\end{aligned} \tag{D.16}$$

Assumption A1 and the triangle inequality bound the norm of the bracketed expression in (D.16) by

$$(\ell_f + \beta)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + \beta\|\boldsymbol{w}_k^t - \boldsymbol{w}_k^{t-1}\|_2 + \|\boldsymbol{e}_k^{t+1}\|_2 + \|\boldsymbol{e}_k^t\|_2. \tag{D.17}$$

Apply the first inequality of (D.6) to both $\boldsymbol{e}_k^{t+1}$ and $\boldsymbol{e}_k^t$, using $t \ge 1$. Using $(\upsilon_1 + \upsilon_2)^2 \le 2\upsilon_1^2 + 2\upsilon_2^2$ to separate the $\boldsymbol{P}$ term from the auxiliary differences gives

$$\begin{aligned}\frac{1}{\rho}\|\boldsymbol{\lambda}_k^{t+1} - \boldsymbol{\lambda}_k^t\|_2^2 \le \frac{2}{\rho\sigma_{\mathrm{H}}}\Big[(\ell_f + \beta + \kappa)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + (\beta+\kappa)\|\boldsymbol{w}_k^t - \boldsymbol{w}_k^{t-1}\|_2 + \varepsilon_k^{t+1} + \varepsilon_k^t\Big]^2 \\ + 2\rho\big(\|\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t\|_2 + \|\boldsymbol{s}_k^t - \boldsymbol{s}_k^{t-1}\|_2\big)^2.\end{aligned} \tag{D.18}$$

Next use $(\upsilon_1 + \upsilon_2)^2 \le (5/4)\upsilon_1^2 + 5\upsilon_2^2$ to separate the model differences from the two errors. Applying $(\upsilon_1 + \upsilon_2)^2 \le 2\upsilon_1^2 + 2\upsilon_2^2$ to the error pair and to the auxiliary pair gives

$$\begin{aligned}\frac{1}{\rho}\|\boldsymbol{\lambda}_k^{t+1} - \boldsymbol{\lambda}_k^t\|_2^2 \le \frac{5}{2\rho\sigma_{\mathrm{H}}}\Big[(\ell_f + \beta + \kappa)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + (\beta+\kappa)\|\boldsymbol{w}_k^t - \boldsymbol{w}_k^{t-1}\|_2\Big]^2 \\ + \frac{20}{\rho\sigma_{\mathrm{H}}}\big((\varepsilon_k^{t+1})^2 + (\varepsilon_k^t)^2\big) + 4\rho\big(\|\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t\|_2^2 + \|\boldsymbol{s}_k^t - \boldsymbol{s}_k^{t-1}\|_2^2\big).\end{aligned} \tag{D.19}$$

The weighted Cauchy–Schwarz inequality and $c_{\mathrm{loc}} = \ell_f + 2\beta + 2\kappa$ imply

$$\begin{aligned}&\Big[(\ell_f + \beta + \kappa)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + (\beta+\kappa)\|\boldsymbol{w}_k^t - \boldsymbol{w}_k^{t-1}\|_2\Big]^2 \\ &\quad\le c_{\mathrm{loc}}\Big[(\ell_f + \beta + \kappa)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2^2 + (\beta+\kappa)\|\boldsymbol{w}_k^t - \boldsymbol{w}_k^{t-1}\|_2^2\Big].\end{aligned} \tag{D.20}$$

Insert (D.20) into (D.19), multiply by p_k, and sum over k. The weighted Frobenius norm definition and the error identities in (D.6), at the current and previous rounds, give

$$\begin{aligned}\frac{1}{\rho}\|\boldsymbol{\Lambda}^{t+1} - \boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2 \le \frac{5(\ell_f + \beta + \kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}}\|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 \\ + 4\rho\big(\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + \|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2\big) + \frac{20 b_{\mathrm{grad}}}{\rho\sigma_{\mathrm{H}}}(n_t + n_{t-1}).\end{aligned} \tag{D.21}$$

Step 3: Compare the successive primal and dual updates. For the model block, keep $\boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t$ fixed. The equality below follows by subtracting the definitions (11), (C.2), and (D.3): the cell indicators and dual correction cancel, while the local proximal term is zero at $\boldsymbol{w}_k^t$ and must be subtracted at $\boldsymbol{w}_k^{t+1}$. The last inequality is exactly the objective bound in Lemma 3, (D.6), with b_{obj} and n_t defined in (D.4) and (C.1). Thus

$$\begin{aligned}&\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1}, \boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t) - \widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^t, \boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t) \\ &= \sum_k p_k\big(\phi_k^t(\boldsymbol{w}_k^{t+1}) - \phi_k^t(\boldsymbol{w}_k^t)\big) - \frac{\beta}{2}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 \\ &\le -\frac{\beta}{2}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + b_{\mathrm{obj}} n_t.\end{aligned} \tag{D.22}$$

For the auxiliary block, we first state the estimate that is needed for the one-round descent argument:

$$\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1}, \boldsymbol{S}^{t+1}, \boldsymbol{v}^{t+1}, \boldsymbol{U}^t) - \widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1}, \boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t) \le -\frac{\eta}{2}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 - \sum_{j : v_j^{t+1} \ne v_j^t} v_j^t h_j^{t+1}. \tag{D.23}$$

We prove this estimate by comparing the proximalized auxiliary objective in the new and old code cells. First, completing the square in the auxiliary variable client by client and summing with weights p_k gives the exact identity

$$\begin{aligned}&\frac{\rho}{2}\sum_k p_k\|\boldsymbol{H}\boldsymbol{w}_k^{t+1}-\boldsymbol{s}_k+\boldsymbol{u}_k^t\|_2^2+\frac{\eta}{2}\|\boldsymbol{S}-\boldsymbol{S}^t\|_{\mathrm{F},p}^2\\&=\frac{\rho+\eta}{2}\sum_k p_k\|\boldsymbol{s}_k-\boldsymbol{a}_k^{t+1}\|_2^2+\frac{\rho\eta}{2(\rho+\eta)}\sum_k p_k\|\boldsymbol{H}\boldsymbol{w}_k^{t+1}+\boldsymbol{u}_k^t-\boldsymbol{s}_k^t\|_2^2.\end{aligned}\tag{D.24}$$

By the half-line projection calculation in Lemma 1, for a fixed $\boldsymbol{v}$ the minimum of the first term of (D.24) over $\boldsymbol{s}_k\in\mathcal{C}_{\boldsymbol{v}}^\tau$ is $\frac{\rho+\eta}{2}\sum_k p_k\sum_j[\tau-v_ja_{k,j}^{t+1}]_+^2$. The second term is independent of $\boldsymbol{v}$. As in Lemma 1, subtracting the old-cell minimum from the new-cell minimum gives

$$\frac{1}{2}\langle\boldsymbol{h}^{t+1},\boldsymbol{v}^{t+1}-\boldsymbol{v}^t\rangle=-\sum_{j:v_j^{t+1}\neq v_j^t}v_j^th_j^{t+1}.\tag{D.25}$$

In (D.25), unchanged coordinates contribute zero; when a sign flips, $v_j^{t+1}-v_j^t=-2v_j^t$. We now compare the two auxiliary objective values. The new point $\boldsymbol{S}^{t+1}$ is the projection onto the new cell, whereas the old point $\boldsymbol{S}^t$ is feasible in the old cell. Thus the value at the new point equals the new-cell minimum, and the value at the old point is at least the old-cell minimum. Applying the completed-square identity (D.24) to both values cancels its code-independent term and gives

$$\begin{aligned}&\frac{\rho}{2}\sum_k p_k\|\boldsymbol{H}\boldsymbol{w}_k^{t+1}-\boldsymbol{s}_k^{t+1}+\boldsymbol{u}_k^t\|_2^2+\frac{\eta}{2}\|\boldsymbol{S}^{t+1}-\boldsymbol{S}^t\|_{\mathrm{F},p}^2-\frac{\rho}{2}\sum_k p_k\|\boldsymbol{H}\boldsymbol{w}_k^{t+1}-\boldsymbol{s}_k^t+\boldsymbol{u}_k^t\|_2^2\\&\le\frac{\rho+\eta}{2}\sum_k p_k\Big[\operatorname{dist}^2(\boldsymbol{a}_k^{t+1},\mathcal{C}_{\boldsymbol{v}^{t+1}}^\tau)-\operatorname{dist}^2(\boldsymbol{a}_k^{t+1},\mathcal{C}_{\boldsymbol{v}^t}^\tau)\Big]\\&=\frac{1}{2}\langle\boldsymbol{h}^{t+1},\boldsymbol{v}^{t+1}-\boldsymbol{v}^t\rangle=-\sum_{j:v_j^{t+1}\neq v_j^t}v_j^th_j^{t+1}.\end{aligned}\tag{D.26}$$

The equality with $\boldsymbol{h}^{t+1}$ follows from the fixed-code minimum in Lemma 1; no exact minimization by the compressed code is used. At the two feasible endpoints, the cell indicators vanish. With $\boldsymbol{W}^{t+1}$ and $\boldsymbol{U}^t$ fixed, the model objective and dual correction also cancel. The left side of (D.26) is therefore the corrected augmented-Lagrangian difference in (D.23) plus $\eta\|\boldsymbol{S}^{t+1}-\boldsymbol{S}^t\|_{\mathrm{F},p}^2/2$. Subtracting this proximal term proves (D.23).

Finally, expanding the corrected augmented Lagrangian in (C.2) leaves only the term $\sum_k p_k\langle\boldsymbol{\lambda}_k,\boldsymbol{H}\boldsymbol{w}_k-\boldsymbol{s}_k\rangle$ dependent on the dual variable. This gives the first equality below; the second uses the dual recurrence (D.12) and the weighted Frobenius norm definition:

$$\begin{aligned}&\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1},\boldsymbol{S}^{t+1},\boldsymbol{v}^{t+1},\boldsymbol{U}^{t+1})-\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1},\boldsymbol{S}^{t+1},\boldsymbol{v}^{t+1},\boldsymbol{U}^t)\\&=\sum_k p_k\langle\boldsymbol{\lambda}_k^{t+1}-\boldsymbol{\lambda}_k^t,\boldsymbol{H}\boldsymbol{w}_k^{t+1}-\boldsymbol{s}_k^{t+1}\rangle=\frac{1}{\rho}\|\boldsymbol{\Lambda}^{t+1}-\boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2.\end{aligned}\tag{D.27}$$

Adding (D.22), (D.23), and (D.27) gives

$$\begin{aligned}&\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1},\boldsymbol{S}^{t+1},\boldsymbol{v}^{t+1},\boldsymbol{U}^{t+1})-\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^t,\boldsymbol{S}^t,\boldsymbol{v}^t,\boldsymbol{U}^t)\\&\quad\le-\frac{\beta}{2}\|\boldsymbol{W}^{t+1}-\boldsymbol{W}^t\|_{\mathrm{F},p}^2-\frac{\eta}{2}\|\boldsymbol{S}^{t+1}-\boldsymbol{S}^t\|_{\mathrm{F},p}^2+\frac{1}{\rho}\|\boldsymbol{\Lambda}^{t+1}-\boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2\\&\qquad-\sum_{j:v_j^{t+1}\neq v_j^t}v_j^th_j^{t+1}+b_{\mathrm{obj}}n_t.\end{aligned}\tag{D.28}$$

Step 4: Account for every term of the Lyapunov decrease. Subtracting the two instances of the Lyapunov definition (C.3) gives

$$\begin{aligned}\psi^{t+1}-\psi^t=\;&\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^{t+1},\boldsymbol{S}^{t+1},\boldsymbol{v}^{t+1},\boldsymbol{U}^{t+1})-\widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^t,\boldsymbol{S}^t,\boldsymbol{v}^t,\boldsymbol{U}^t)\\&+\frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}}\big(\|\boldsymbol{W}^{t+1}-\boldsymbol{W}^t\|_{\mathrm{F},p}^2-\|\boldsymbol{W}^t-\boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2\big)\\&+4\rho\big(\|\boldsymbol{S}^{t+1}-\boldsymbol{S}^t\|_{\mathrm{F},p}^2-\|\boldsymbol{S}^t-\boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2\big).\end{aligned}\tag{D.29}$$

Insert (D.28) and then (D.21). The coefficient of the previous model difference is $5(\beta + \kappa)c_{\mathrm{loc}}/(2\rho\sigma_{\mathrm{H}}) - 5(\beta + \kappa)c_{\mathrm{loc}}/(2\rho\sigma_{\mathrm{H}}) = 0$; the coefficient of the previous auxiliary difference is $4\rho - 4\rho = 0$. For the current differences, use $c_{\mathrm{loc}} = \ell_f + 2\beta + 2\kappa$ and the coefficient definitions (C.4) to obtain

$$\begin{aligned} -\frac{\beta}{2} + \frac{5(\ell_f + \beta + \kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}} + \frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho\sigma_{\mathrm{H}}} &= -\frac{\beta}{2} + \frac{5c_{\mathrm{loc}}^2}{2\rho\sigma_{\mathrm{H}}} = -2a_{\mathrm{w}}, \\ -\frac{\eta}{2} + 4\rho + 4\rho &= -\frac{\eta - 16\rho}{2} = -2a_{\mathrm{s}}. \end{aligned} \tag{D.30}$$

Consequently,

$$\begin{aligned} \psi^{t+1} - \psi^t \le & -2a_{\mathrm{w}}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 - 2a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 \\ & - \sum_{j : v_j^{t+1} \neq v_j^t} v_j^t h_j^{t+1} + b_{\mathrm{obj}} n_t + b_1(n_t + n_{t-1}). \end{aligned} \tag{D.31}$$

Lemma 4 in Appendix D.4 gives the following bound directly from the transmitted signs and Lemma 3:

$$\left[- \sum_{j : v_j^{t+1} \neq v_j^t} v_j^t h_j^{t+1} \right]_+ \le a_{\mathrm{w}}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \frac{2a_{\mathrm{w}} b_{\mathrm{grad}}}{\beta^2 + 2\kappa^2} n_t. \tag{D.32}$$

Since any scalar is at most its positive part, (D.32) also bounds the signed switching term in (D.31). Combining the bounds, the model coefficient becomes $-2a_{\mathrm{w}} + a_{\mathrm{w}} = -a_{\mathrm{w}}$. The current error coefficient is

$$b_{\mathrm{obj}} + b_1 + \frac{2a_{\mathrm{w}} b_{\mathrm{grad}}}{\beta^2 + 2\kappa^2} = b_0, \tag{D.33}$$

while the previous error coefficient remains b_1. For every changed code coordinate, the opposite margins in the cell definition (8) and feasibility of the old and new auxiliary iterates imply $|s_{k,j}^{t+1} - s_{k,j}^t| \ge 2\tau$ for every client k. Since $\sum_k p_k = 1$,

$$\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 \ge \sum_k p_k \sum_{j : v_j^{t+1} \neq v_j^t} 4\tau^2 = 4\tau^2 d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t). \tag{D.34}$$

Therefore

$$\begin{aligned} -2a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 = & -a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 - a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 \\ \le & -a_{\mathrm{s}}\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 - 4\tau^2 a_{\mathrm{s}} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t). \end{aligned} \tag{D.35}$$

These substitutions prove every term of (C.6). For the lower bound, use (C.2) and cell feasibility to obtain the equality below. Drop the nonnegative square and apply the objective lower bounds in Assumption A1 and the dual bound in (B.1) to obtain the inequality:

$$\begin{aligned} \widetilde{\mathcal{L}}_\rho(\boldsymbol{W}^t, \boldsymbol{S}^t, \boldsymbol{v}^t, \boldsymbol{U}^t) &= f(\boldsymbol{W}^t) + \frac{\rho}{2}\sum_k p_k \|\boldsymbol{H}\boldsymbol{w}_k^t - \boldsymbol{s}_k^t + \boldsymbol{\lambda}_k^t/\rho\|_2^2 - \frac{1}{2\rho}\|\boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2 \\ &\ge \sum_k p_k \underline{f}_k - \frac{r_\lambda^2}{2\rho} = \psi_{\min}. \end{aligned} \tag{D.36}$$

The remaining terms in ψ^t are nonnegative.

Step 5: Bound the three components of the KKT residual. The first identity in (D.15) gives the feasibility bound below. For the normal-cone distance, the third identity supplies the candidate $\boldsymbol{\lambda}_k^{t+1} - \eta(\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t)$ in the cone; the infimum defining distance is at most the distance to this candidate. Squaring and summing with weights p_k gives

$$\sum_k p_k \|\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^{t+1}\|_2^2 = \rho^{-2}\|\boldsymbol{\Lambda}^{t+1} - \boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2, \tag{D.37}$$

$$\sum_k p_k \operatorname{dist}^2\big(\boldsymbol{\lambda}_k^{t+1}, \mathcal{N}_{\mathcal{C}^\tau_{\boldsymbol{v}^{t+1}}}(\boldsymbol{s}_k^{t+1})\big) \le \eta^2\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2. \tag{D.38}$$

For model stationarity, use the second identity in (D.15), $\|\boldsymbol{H}^\top\|_{\mathrm{op}} = h_{\mathrm{op}}$, the triangle inequality, and then (D.6):

$$\begin{aligned}\|\nabla f_k(\boldsymbol{w}_k^{t+1}) + \boldsymbol{H}^\top \boldsymbol{\lambda}_k^{t+1}\|_2 &\le \|\boldsymbol{e}_k^{t+1}\|_2 + \beta\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + \rho h_{\mathrm{op}}\|\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t\|_2 \\ &\le (\beta+\kappa)\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + \rho h_{\mathrm{op}}\|\boldsymbol{s}_k^{t+1} - \boldsymbol{s}_k^t\|_2 + \varepsilon_k^{t+1}.\end{aligned} \tag{D.39}$$

Square using $(\upsilon_1+\upsilon_2+\upsilon_3)^2 \le 3(\upsilon_1^2+\upsilon_2^2+\upsilon_3^2)$, multiply by p_k, and sum. Adding (D.37) and (D.38), and using (30) and the error identity in (D.6), yields

$$\begin{aligned}r_{t+1}^2 \le \rho^{-2}\|\boldsymbol{\Lambda}^{t+1} - \boldsymbol{\Lambda}^t\|_{\mathrm{F},p}^2 + 3(\beta+\kappa)^2\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 \\ + (3\rho^2 h_{\mathrm{op}}^2 + \eta^2)\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + 3b_{\mathrm{grad}} n_t.\end{aligned} \tag{D.40}$$

Divide (D.21) by ρ and substitute it for the first term. Collecting each coefficient separately gives

$$\begin{aligned}r_{t+1}^2 \le &\left[\frac{5(\ell_f+\beta+\kappa)c_{\mathrm{loc}}}{2\rho^2\sigma_{\mathrm{H}}} + 3(\beta+\kappa)^2\right]\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 \\ &+ \frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho^2\sigma_{\mathrm{H}}}\|\boldsymbol{W}^t - \boldsymbol{W}^{t-1}\|_{\mathrm{F},p}^2 \\ &+ (4 + 3\rho^2 h_{\mathrm{op}}^2 + \eta^2)\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^t\|_{\mathrm{F},p}^2 + 4\|\boldsymbol{S}^t - \boldsymbol{S}^{t-1}\|_{\mathrm{F},p}^2 \\ &+ \left(\frac{20}{\rho^2\sigma_{\mathrm{H}}} + 3\right) b_{\mathrm{grad}} n_t + \frac{20 b_{\mathrm{grad}}}{\rho^2\sigma_{\mathrm{H}}} n_{t-1}.\end{aligned} \tag{D.41}$$

The three current-round coefficients are precisely the entries in the maximum defining c_{res} in (C.5). Each previous-round coefficient is no larger than its current-round counterpart:

$$\begin{aligned}\frac{5(\beta+\kappa)c_{\mathrm{loc}}}{2\rho^2\sigma_{\mathrm{H}}} \le \frac{5(\ell_f+\beta+\kappa)c_{\mathrm{loc}}}{2\rho^2\sigma_{\mathrm{H}}} + 3(\beta+\kappa)^2 \le c_{\mathrm{res}}, \\ 4 \le 4 + 3\rho^2 h_{\mathrm{op}}^2 + \eta^2 \le c_{\mathrm{res}}, \\ \frac{20 b_{\mathrm{grad}}}{\rho^2\sigma_{\mathrm{H}}} \le \left(\frac{20}{\rho^2\sigma_{\mathrm{H}}} + 3\right) b_{\mathrm{grad}} \le c_{\mathrm{res}}.\end{aligned} \tag{D.42}$$

Replacing each of the six nonnegative coefficients in (D.41) by this fixed upper bound proves (C.7). □

D.3 Local SGD Error Bounds

Fix k, t, write $\phi = \phi_k^t$ and $N = N_{k,t}$, and suppress these fixed indices by writing $\boldsymbol{x}_i = \boldsymbol{x}_{k,i}^t$ and $\boldsymbol{\nu}_i = \boldsymbol{\nu}_{k,i}^t$. By Assumptions A1 and A6, $f_k + \mu_f\|\cdot\|_2^2/2$ is convex and $\beta > \mu_f$; hence ϕ is $\mu = \beta - \mu_f$ strongly convex. Its gradient is $\ell_\phi = \ell_f + \rho h_{\mathrm{op}}^2 + \beta$ Lipschitz. Let $\boldsymbol{x}_\star$ be its unique minimizer, so $\nabla\phi(\boldsymbol{x}_\star) = \boldsymbol{0}$. The local SGD updates start at $\boldsymbol{x}_0 = \boldsymbol{w}_k^t$, end at $\boldsymbol{x}_N = \boldsymbol{w}_k^{t+1}$, and satisfy

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma\big(\nabla\phi(\boldsymbol{x}_i) + \boldsymbol{\nu}_i\big), \qquad i = 0, \dots, N-1. \tag{D.43}$$

Here only the task gradient is sampled; the gradients of the quadratic terms in ϕ are computed exactly. Thus $\boldsymbol{\nu}_i$ is precisely the gradient error defined in Assumption A5. Only this definition is used here: the following local bounds hold for each realization and do not use its moment bound or assume unbiased gradients.

Step 1: Bound the local objective change. The smoothness inequality for ϕ, the recurrence (D.43), and the step-size condition $\gamma\ell_\phi \le 1$ in (B.3) give, respectively, the first three lines below:

$$\begin{aligned}\phi(\boldsymbol{x}_{i+1}) - \phi(\boldsymbol{x}_i) &\le \langle\nabla\phi(\boldsymbol{x}_i), \boldsymbol{x}_{i+1} - \boldsymbol{x}_i\rangle + \frac{\ell_\phi}{2}\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 \\ &= -\gamma\langle\nabla\phi(\boldsymbol{x}_i), \nabla\phi(\boldsymbol{x}_i) + \boldsymbol{\nu}_i\rangle + \frac{\ell_\phi\gamma^2}{2}\|\nabla\phi(\boldsymbol{x}_i) + \boldsymbol{\nu}_i\|_2^2 \\ &\le -\gamma\langle\nabla\phi(\boldsymbol{x}_i), \nabla\phi(\boldsymbol{x}_i) + \boldsymbol{\nu}_i\rangle + \frac{\gamma}{2}\|\nabla\phi(\boldsymbol{x}_i) + \boldsymbol{\nu}_i\|_2^2 \\ &= -\frac{\gamma}{2}\|\nabla\phi(\boldsymbol{x}_i)\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\nu}_i\|_2^2.\end{aligned} \tag{D.44}$$

In the last equality, the cross terms $-\gamma\langle\nabla\phi(\boldsymbol{x}_i),\boldsymbol{\nu}_i\rangle$ and $\gamma\langle\nabla\phi(\boldsymbol{x}_i),\boldsymbol{\nu}_i\rangle$ cancel. Therefore

$$\phi(\boldsymbol{x}_{i+1}) \le \phi(\boldsymbol{x}_i) - \frac{\gamma}{2}\|\nabla\phi(\boldsymbol{x}_i)\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.45}$$

Sum over $i = 0, \dots, N-1$; the objective differences telescope:

$$\begin{aligned}\phi(\boldsymbol{x}_N) - \phi(\boldsymbol{x}_0) &= \sum_{i=0}^{N-1}\big(\phi(\boldsymbol{x}_{i+1}) - \phi(\boldsymbol{x}_i)\big) \\ &\le -\frac{\gamma}{2}\sum_{i=0}^{N-1}\|\nabla\phi(\boldsymbol{x}_i)\|_2^2 + \frac{\gamma}{2}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2 \le \frac{\gamma}{2}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2.\end{aligned} \tag{D.46}$$

Hence, with $b_{\mathrm{obj}} = \gamma/2$,

$$\phi(\boldsymbol{x}_N) \le \phi(\boldsymbol{x}_0) + b_{\mathrm{obj}}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.47}$$

Step 2: Establish contraction of the objective gap. Strong convexity and completion of the square give

$$\begin{aligned}\phi(\boldsymbol{x}_\star) &\ge \phi(\boldsymbol{x}_i) + \langle\nabla\phi(\boldsymbol{x}_i), \boldsymbol{x}_\star - \boldsymbol{x}_i\rangle + \frac{\mu}{2}\|\boldsymbol{x}_\star - \boldsymbol{x}_i\|_2^2 \\ &= \phi(\boldsymbol{x}_i) - \frac{1}{2\mu}\|\nabla\phi(\boldsymbol{x}_i)\|_2^2 + \frac{\mu}{2}\|\boldsymbol{x}_\star - \boldsymbol{x}_i + \nabla\phi(\boldsymbol{x}_i)/\mu\|_2^2.\end{aligned} \tag{D.48}$$

Dropping the last nonnegative term and rearranging yields $\|\nabla\phi(\boldsymbol{x}_i)\|_2^2 \ge 2\mu\big(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_\star)\big)$. Substituting this lower bound into the negative gradient term in (D.45) gives

$$\phi(\boldsymbol{x}_{i+1}) - \phi(\boldsymbol{x}_\star) \le (1-\gamma\mu)\big(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_\star)\big) + \frac{\gamma}{2}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.49}$$

Define the contraction factor

$$\vartheta = 1 - \gamma\mu. \tag{D.50}$$

Assumption A2 gives $h_{\mathrm{op}} > 0$ and $\rho > 0$. Since $\ell_\phi - \mu = \ell_f + \rho h_{\mathrm{op}}^2 + \mu_f > 0$ and $0 < \gamma \le 1/\ell_\phi$, we have $0 < \gamma\mu \le \mu/\ell_\phi < 1$, hence $0 < \vartheta < 1$. Thus

$$\phi(\boldsymbol{x}_{i+1}) - \phi(\boldsymbol{x}_\star) \le \vartheta\big(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_\star)\big) + \frac{\gamma}{2}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.51}$$

Applying this recurrence for $i = 0, \dots, N-1$ gives

$$\phi(\boldsymbol{x}_N) - \phi(\boldsymbol{x}_\star) \le \vartheta^N\big(\phi(\boldsymbol{x}_0) - \phi(\boldsymbol{x}_\star)\big) + \frac{\gamma}{2}\sum_{i=0}^{N-1}\vartheta^{N-1-i}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.52}$$

Each error at step i is multiplied by ϑ in the subsequent $N-1-i$ steps, which accounts for its weight in the sum.

Step 3: Convert contraction to the relative local residual. Smoothness and strong convexity at $\boldsymbol{x}_\star$, where $\nabla\phi(\boldsymbol{x}_\star) = \mathbf{0}$, imply

$$\phi(\boldsymbol{x}_0) - \phi(\boldsymbol{x}_\star) \le \frac{\ell_\phi}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2, \qquad \phi(\boldsymbol{x}_N) - \phi(\boldsymbol{x}_\star) \ge \frac{\mu}{2}\|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2^2. \tag{D.53}$$

Insert both bounds into (D.52) and multiply by $2/\mu$:

$$\|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2^2 \le \frac{\ell_\phi}{\mu}\vartheta^N\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{\gamma}{\mu}\sum_{i=0}^{N-1}\vartheta^{N-1-i}\|\boldsymbol{\nu}_i\|_2^2. \tag{D.54}$$

Taking square roots using $\sqrt{v_1 + v_2} \le \sqrt{v_1} + \sqrt{v_2}$ and $\vartheta^{N-1-i} \le 1$ yields

$$\|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2 \le \sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 + \left(\frac{\gamma}{\mu}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2\right)^{1/2}. \tag{D.55}$$

To bound the terminal gradient in terms of the actual local displacement, choose N so that $\sqrt{\ell_\phi/\mu}\,\vartheta^{N/2} \le \kappa/(\ell_\phi+\kappa)$. Taking logarithms shows that this condition is equivalent to

$$N(-\log\vartheta) \ge \log(\ell_\phi/\mu) + 2\log((\ell_\phi+\kappa)/\kappa). \tag{D.56}$$

It is therefore ensured by

$$N_{\min} = \max\left\{1, \left\lceil \frac{\log(\ell_\phi/\mu) + 2\log((\ell_\phi+\kappa)/\kappa)}{-\log\vartheta} \right\rceil\right\}. \tag{D.57}$$

For $N \ge N_{\min}$, use the triangle inequality $\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 \le \|\boldsymbol{x}_N - \boldsymbol{x}_0\|_2 + \|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2$ in (D.55) and move the last distance to the left:

$$\left(1 - \sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}\right)\|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2 \le \sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}\|\boldsymbol{x}_N - \boldsymbol{x}_0\|_2 + \left(\frac{\gamma}{\mu}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2\right)^{1/2}. \tag{D.58}$$

The step-count condition (D.57) makes the coefficient on the left of (D.58) at least $\ell_\phi/(\ell_\phi+\kappa) > 0$. Since $\boldsymbol{x}_\star$ minimizes the differentiable unconstrained objective ϕ, we have $\nabla\phi(\boldsymbol{x}_\star) = \boldsymbol{0}$. The ℓ_ϕ-Lipschitz continuity of $\nabla\phi$ therefore gives

$$\|\nabla\phi(\boldsymbol{x}_N)\|_2 = \|\nabla\phi(\boldsymbol{x}_N) - \nabla\phi(\boldsymbol{x}_\star)\|_2 \le \ell_\phi\|\boldsymbol{x}_N - \boldsymbol{x}_\star\|_2. \tag{D.59}$$

Divide (D.58) by its positive left-hand coefficient and apply (D.59). The next inequality uses the threshold in (D.57); the last equality uses b_{grad} from (D.4):

$$\begin{aligned}
\|\nabla\phi(\boldsymbol{x}_N)\|_2 &\le \frac{\ell_\phi\sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}}{1-\sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}}\|\boldsymbol{x}_N - \boldsymbol{x}_0\|_2 + \frac{\ell_\phi}{1-\sqrt{\ell_\phi/\mu}\,\vartheta^{N/2}}\left(\frac{\gamma}{\mu}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2\right)^{1/2} \\
&\le \kappa\|\boldsymbol{x}_N - \boldsymbol{x}_0\|_2 + (\ell_\phi+\kappa)\left(\frac{\gamma}{\mu}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2\right)^{1/2} \\
&= \kappa\|\boldsymbol{x}_N - \boldsymbol{x}_0\|_2 + \left(b_{\mathrm{grad}}\sum_{i=0}^{N-1}\|\boldsymbol{\nu}_i\|_2^2\right)^{1/2},
\end{aligned} \tag{D.60}$$

where $b_{\mathrm{grad}} = \gamma(\ell_\phi+\kappa)^2/\mu$. Restoring the indices k, t gives the first claim of Lemma 3 with (D.5). Finally, multiplying the squared error definition by p_k and summing gives $\|\boldsymbol{\varepsilon}^{t+1}\|_p^2 = b_{\mathrm{grad}} n_t$ exactly. Multiplying (D.47) by p_k and summing proves the objective bound with coefficient b_{obj}.

D.4 Code-Switching Error Control

The following lemma supplies the switching estimate used in Step 4 of the proof of Lemma 2.

Lemma 4 (Control of the code-switching error)**.** *Under Assumptions A1–A4 and A6, the iterates of FICO satisfy, almost surely for every $t \ge 0$,*

$$\left[-\sum_{j:v_j^{t+1}\neq v_j^t} v_j^t h_j^{t+1}\right]_+ \le a_{\mathrm{w}}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + \frac{2a_{\mathrm{w}}b_{\mathrm{grad}}}{\beta^2+2\kappa^2}n_t,$$

where $a_{\mathrm{w}} = \beta/4 - 5c_{\mathrm{loc}}^2/(4\rho\sigma_{\mathrm{H}}) > 0$, and n_t and b_{grad} are defined in (C.1) *and* (D.4)*, respectively.*

Proof. Work on the probability-one event on which the stability bounds hold for all rounds. The argument uses the transmitted signs and the local-solver estimate in Lemma 3, without invoking the descent or residual conclusions of Lemma 2.

Step 1: Derive uniform bounds for the gradients and projection centers. Define

$$\begin{aligned}
g_{\mathrm{bd}} &= \max_k \|\nabla f_k(\boldsymbol{0})\|_2 + \frac{\ell_f r_{\mathrm{w}}}{\sqrt{p_{\min}}}, \\
a_{\mathrm{bd}} &= \frac{\max\{r_{\mathrm{s}}, h_{\mathrm{op}}r_{\mathrm{w}} + r_\lambda/\rho\}}{\sqrt{p_{\min}}}, \qquad p_{\min} = \min_k p_k.
\end{aligned} \tag{D.61}$$

For any round ℓ, the weighted Frobenius norm definition and Assumption A4 give $p_k\|\boldsymbol{w}_k^\ell\|_2^2 \le \sum_i p_i\|\boldsymbol{w}_i^\ell\|_2^2 = \|\boldsymbol{W}^\ell\|_{\mathrm{F},p}^2 \le r_{\mathrm{w}}^2$. Applying the same argument to $\boldsymbol{S}^\ell$ and $\boldsymbol{\Lambda}^\ell$, and using $p_k \ge p_{\min} > 0$, yields the individual bounds

$$\|\boldsymbol{w}_k^\ell\|_2 \le \frac{r_{\mathrm{w}}}{\sqrt{p_{\min}}}, \qquad \|\boldsymbol{s}_k^\ell\|_2 \le \frac{r_{\mathrm{s}}}{\sqrt{p_{\min}}}, \qquad \|\boldsymbol{\lambda}_k^\ell\|_2 \le \frac{r_\lambda}{\sqrt{p_{\min}}}.$$

The ℓ_f-Lipschitz continuity in Assumption A1 then implies

$$\begin{aligned}\|\nabla f_k(\boldsymbol{w}_k^{t+1})\|_2 &\le \|\nabla f_k(\boldsymbol{0})\|_2 + \|\nabla f_k(\boldsymbol{w}_k^{t+1}) - \nabla f_k(\boldsymbol{0})\|_2 \\ &\le \|\nabla f_k(\boldsymbol{0})\|_2 + \ell_f\|\boldsymbol{w}_k^{t+1}\|_2 \\ &\le \max_\ell \|\nabla f_\ell(\boldsymbol{0})\|_2 + \frac{\ell_f r_{\mathrm{w}}}{\sqrt{p_{\min}}} = g_{\mathrm{bd}}.\end{aligned}$$

For the projection center, use $\boldsymbol{\lambda}_k^t = \rho\boldsymbol{u}_k^t$ to write

$$\boldsymbol{a}_k^{t+1} = \frac{\rho}{\rho+\eta}\left(\boldsymbol{H}\boldsymbol{w}_k^{t+1} + \frac{\boldsymbol{\lambda}_k^t}{\rho}\right) + \frac{\eta}{\rho+\eta}\boldsymbol{s}_k^t.$$

The two coefficients are nonnegative and sum to one. Since $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2$ and $\|\boldsymbol{H}\boldsymbol{w}\|_2 \le h_{\mathrm{op}}\|\boldsymbol{w}\|_2$, the preceding individual bounds give

$$\begin{aligned}\|\boldsymbol{a}_k^{t+1}\|_\infty &\le \frac{\rho}{\rho+\eta}\left(h_{\mathrm{op}}\|\boldsymbol{w}_k^{t+1}\|_2 + \frac{\|\boldsymbol{\lambda}_k^t\|_2}{\rho}\right) + \frac{\eta}{\rho+\eta}\|\boldsymbol{s}_k^t\|_2 \\ &\le \frac{\rho(h_{\mathrm{op}}r_{\mathrm{w}} + r_\lambda/\rho) + \eta r_{\mathrm{s}}}{(\rho+\eta)\sqrt{p_{\min}}} \\ &\le \frac{\max\{r_{\mathrm{s}}, h_{\mathrm{op}}r_{\mathrm{w}} + r_\lambda/\rho\}}{\sqrt{p_{\min}}} = a_{\mathrm{bd}}.\end{aligned}$$

The last inequality bounds a convex combination of two nonnegative scalars by their maximum. The threshold used in Assumption A6 is

$$\eta_{\min} = \max\left\{16\rho,\ \frac{2g_{\mathrm{bd}}}{\tau\sqrt{\sigma_{\mathrm{H}}}},\ \frac{16(\beta^2+2\kappa^2)(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}\big(\beta - 5c_{\mathrm{loc}}^2/(\rho\sigma_{\mathrm{H}})\big)\tau^2}\right\}. \tag{D.62}$$

The bounds $g_{\mathrm{bd}}, a_{\mathrm{bd}}$ may depend on η, so this threshold and Assumption A4 are joint conditions on the same run. The term 16ρ ensures $a_{\mathrm{s}} > 0$ in the overall descent estimate; the other two terms control the task-gradient contribution in Step 2 and the switching-error coefficient in Step 3 below.

Step 2: Bound the number of switches by displacement and local error. If coordinate j changes, the server rule (24), the message definitions (20)–(21), and $\chi > 0$ from Assumption A2 give

$$\sum_k p_k\theta_k^{t+1}v_j^t q_{k,j}^{t+1} = v_j^t\widehat{h}_j^{t+1} > \chi + \epsilon_{\mathrm{tie}} > 0. \tag{D.63}$$

Since $p_k > 0$, $\theta_k^{t+1} \ge 0$, and $v_j^t q_{k,j}^{t+1} \in \{-1,+1\}$, a positive sum requires at least one strictly positive summand. Choose one such client $k(j)$; it satisfies $\theta_{k(j)}^{t+1} > 0$ and $v_j^t q_{k(j),j}^{t+1} = 1$. To interpret this sign condition, consider (18) as a function of its scalar input υ. For $\upsilon > 0$, $[\tau-\upsilon]_+ < [\tau+\upsilon]_+$, so the cost difference is negative; for $\upsilon < 0$, the inequality reverses and the cost difference is positive. At $\upsilon = 0$ the difference is zero. Thus $q_{k(j),j}^{t+1}$ has the opposite sign to a nonzero $a_{k(j),j}^{t+1}$, and $v_j^t q_{k(j),j}^{t+1} = 1$ implies $v_j^t a_{k(j),j}^{t+1} \le 0$; the latter also holds at zero under either fixed binary sign convention. Meanwhile, membership of the old auxiliary iterate in $\mathcal{C}_{\boldsymbol{v}^t}^\tau$ gives $v_j^t s_{k(j),j}^t \ge \tau$. Subtracting yields $v_j^t(s_{k(j),j}^t - a_{k(j),j}^{t+1}) \ge \tau$. Because $|v_j^t| = 1$, we obtain $|a_{k(j),j}^{t+1} - s_{k(j),j}^t| \ge \tau$. Hence every flipped coordinate has at least one client whose projection center is separated from its old auxiliary value by the margin τ.

To relate this displacement to the local error, recall $\boldsymbol{e}_k^{t+1} = \nabla\phi_k^t(\boldsymbol{w}_k^{t+1})$ and $\boldsymbol{P} = (\boldsymbol{H}\boldsymbol{H}^\top)^{-1}\boldsymbol{H}$. Expand (D.3) and apply $\boldsymbol{P}\boldsymbol{H}^\top = \boldsymbol{I}_m$ for the first equality below; the second follows from the center definition (15):

$$\begin{aligned}\boldsymbol{P}\big(\boldsymbol{e}_k^{t+1} - \nabla f_k(\boldsymbol{w}_k^{t+1}) - \beta(\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t)\big) &= \boldsymbol{\lambda}_k^t + \rho(\boldsymbol{H}\boldsymbol{w}_k^{t+1} - \boldsymbol{s}_k^t) \\ &= (\rho+\eta)(\boldsymbol{a}_k^{t+1} - \boldsymbol{s}_k^t).\end{aligned} \tag{D.64}$$

Move the task-gradient term in (D.64) to the right-hand side and apply the reverse triangle inequality to coordinate j. Since $\|\boldsymbol{P}\|_{\mathrm{op}} = 1/\sqrt{\sigma_{\mathrm{H}}}$ and $\|\nabla f_k(\boldsymbol{w}_k^{t+1})\|_2 \le g_{\mathrm{bd}}$, every coordinate of $\boldsymbol{P}\nabla f_k(\boldsymbol{w}_k^{t+1})$ has absolute value at most $g_{\mathrm{bd}}/\sqrt{\sigma_{\mathrm{H}}}$. Together with the displacement bound above, this gives, for each selected pair $(k(j), j)$,

$$\begin{aligned}\Big|[\boldsymbol{P}(\boldsymbol{e}_{k(j)}^{t+1} - \beta\big(\boldsymbol{w}_{k(j)}^{t+1} - \boldsymbol{w}_{k(j)}^{t}\big))]_j\Big| &\ge (\rho+\eta)|a_{k(j),j}^{t+1} - s_{k(j),j}^{t}| - \Big|[\boldsymbol{P}\nabla f_{k(j)}(\boldsymbol{w}_{k(j)}^{t+1})]_j\Big| \\ &\ge (\rho+\eta)\tau - \frac{g_{\mathrm{bd}}}{\sqrt{\sigma_{\mathrm{H}}}} \\ &> \frac{(\rho+\eta)\tau}{2}.\end{aligned} \tag{D.65}$$

The last inequality follows from $\eta > 2g_{\mathrm{bd}}/(\tau\sqrt{\sigma_{\mathrm{H}}})$, which implies $g_{\mathrm{bd}}/\sqrt{\sigma_{\mathrm{H}}} < \eta\tau/2 < (\rho+\eta)\tau/2$. Thus the bounded task-gradient term cannot cancel more than half of the displacement contribution; a flip requires a sufficiently large transformed local residual or model step. The selected pairs have distinct coordinates even if a client is selected more than once. The sum over these pairs is therefore at most the sum over all clients and coordinates. Using $\sum_k \|\boldsymbol{z}_k\|_2^2 \le p_{\min}^{-1}\sum_k p_k\|\boldsymbol{z}_k\|_2^2$ and $\|\boldsymbol{P}\|_{\mathrm{op}}^2 = \sigma_{\mathrm{H}}^{-1}$ therefore yields

$$\begin{aligned}\frac{(\rho+\eta)^2\tau^2}{4} d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t) &\le \sum_k \|\boldsymbol{P}(\boldsymbol{e}_k^{t+1} - \beta\big(\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\big))\|_2^2 \\ &\le \frac{2}{p_{\min}\sigma_{\mathrm{H}}}\big(\|\boldsymbol{E}^{t+1}\|_{\mathrm{F},p}^2 + \beta^2\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2\big).\end{aligned} \tag{D.66}$$

Step 3: Bound and absorb the switching error. Both $[\tau - a_{k,j}^{t+1}]_+^2$ and $[\tau + a_{k,j}^{t+1}]_+^2$ belong to $[0, (\tau + a_{\mathrm{bd}})^2]$. Their absolute difference is at most $(\tau + a_{\mathrm{bd}})^2$, so $|\delta_{k,j}^{t+1}| \le (\rho+\eta)(\tau+a_{\mathrm{bd}})^2/2$. Using $\boldsymbol{h}^{t+1} = \sum_k p_k\boldsymbol{\delta}_k^{t+1}$, $p_k > 0$, and $\sum_k p_k = 1$, the triangle inequality gives

$$\left[-\sum_{j:v_j^{t+1}\neq v_j^t} v_j^t h_j^{t+1}\right]_+ \le \sum_{j:v_j^{t+1}\neq v_j^t} |h_j^{t+1}| \le \frac{\rho+\eta}{2}(\tau+a_{\mathrm{bd}})^2 d_{\mathrm{H}}(\boldsymbol{v}^{t+1}, \boldsymbol{v}^t). \tag{D.67}$$

Apply (D.6) client by client, use $(\upsilon_1+\upsilon_2)^2 \le 2\upsilon_1^2 + 2\upsilon_2^2$, and substitute the error identity from the same lemma:

$$\begin{aligned}\|\boldsymbol{E}^{t+1}\|_{\mathrm{F},p}^2 &\le \sum_k p_k\big(\kappa\|\boldsymbol{w}_k^{t+1} - \boldsymbol{w}_k^t\|_2 + \varepsilon_k^{t+1}\big)^2 \\ &\le 2\kappa^2\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + 2\|\boldsymbol{\varepsilon}^{t+1}\|_p^2 = 2\kappa^2\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 + 2b_{\mathrm{grad}}n_t.\end{aligned} \tag{D.68}$$

Next solve (D.66) for the Hamming distance and insert this bound into (D.67). The coefficients before and after the local-error substitution are

$$\begin{aligned}\left[-\sum_{j:v_j^{t+1}\neq v_j^t} v_j^t h_j^{t+1}\right]_+ &\le \frac{4(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}(\rho+\eta)\tau^2}\big(\|\boldsymbol{E}^{t+1}\|_{\mathrm{F},p}^2 + \beta^2\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2\big) \\ &\le \frac{4(\beta^2+2\kappa^2)(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}(\rho+\eta)\tau^2}\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|_{\mathrm{F},p}^2 \\ &\quad + \frac{8(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}(\rho+\eta)\tau^2} b_{\mathrm{grad}}n_t.\end{aligned} \tag{D.69}$$

The requirement $\eta > \eta_{\min}$ in (B.3), the last entry in (D.62), and $\rho + \eta > \eta$ give

$$\frac{4(\beta^2+2\kappa^2)(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}(\rho+\eta)\tau^2} < \frac{\beta - 5c_{\mathrm{loc}}^2/(\rho\sigma_{\mathrm{H}})}{4} = \frac{\beta}{4} - \frac{5c_{\mathrm{loc}}^2}{4\rho\sigma_{\mathrm{H}}} = a_{\mathrm{w}}. \tag{D.70}$$

Multiplying this inequality by $2/(\beta^2+2\kappa^2)$ also yields

$$\frac{8(\tau+a_{\mathrm{bd}})^2}{p_{\min}\sigma_{\mathrm{H}}(\rho+\eta)\tau^2} < \frac{\beta/2 - 5c_{\mathrm{loc}}^2/(2\rho\sigma_{\mathrm{H}})}{\beta^2+2\kappa^2} = \frac{2a_{\mathrm{w}}}{\beta^2+2\kappa^2}. \tag{D.71}$$

Substituting (D.70) and (D.71) into (D.69) proves the stated estimate, which is (D.32). □