FEDLORE: COMMUNICATION AND MEMORY EFFICIENT FEDERATED LEARNING VIA SHARED GRADIENT LOW-RANK PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated training of foundation models is constrained by client memory and communication costs. LoRA-based methods reduce these costs through low-rank adapters, but their fixed rank budget can limit adaptation. Gradient low-rank optimization offers greater flexibility, yet independently chosen client subspaces create a problem we term *subspace fragmentation*: local projections interact with data heterogeneity to bias aggregated directions, while aggregation can increase update rank and communication cost. Thus, accurate local gradient compression need not preserve global descent. We propose `FedLore`, which shares a low-rank optimization basis within each round and refreshes it across rounds. The shared basis enables exact aggregation in low-rank coordinates and eliminates the identified projection bias. Subspace refresh allows the accumulated model update to exceed the per-round rank budget. We characterize the aggregation bias and establish an $O(T^{-1/2})$ stationarity bound for the projected-SGD variant under a global-gradient coverage condition and standard smoothness and variance assumptions, with bounded gradient heterogeneity. Experiments on vision and language tasks, including federated pre-training, show that `FedLore` outperforms the evaluated low-rank adapter baselines and matches or exceeds full-parameter training, while reducing communication and optimizer-state memory.

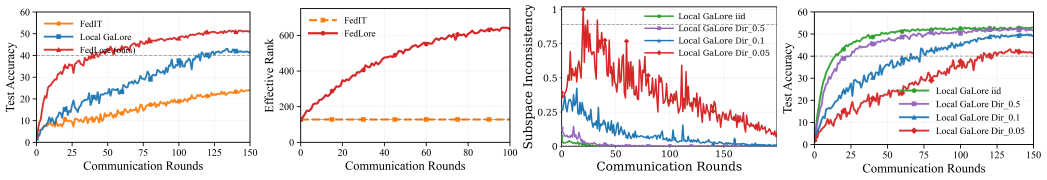
INTRODUCTION

Foundation models have become central to modern vision and language systems, yet much of the data needed for further training resides on private devices or across institutions with restricted data access Bommasani et al. (2021). Federated learning (FL) enables collaborative training without centralizing these data McMahan et al. (2016); Kairouz & McMahan (2021). Scaling this paradigm to foundation models, however, requires addressing two coupled resource bottlenecks: maintaining optimizer states on memory-constrained clients and repeatedly exchanging high-dimensional model updates. Full-parameter federated optimization (`FedFull`) incurs substantial costs on both fronts, particularly with adaptive optimizers such as Adam/AdamW Kingma & Ba (2014).

A common approach combines FL with parameter-efficient fine-tuning, particularly LoRA (Hu et al., 2021). Methods such as FedIT Zhang et al. (2023a) reduce optimizer-state memory and communication by training and exchanging low-rank adapters (Wu et al., 2024; Wang et al., 2024; Yi et al., 2023). For a standard rank-r adapter, however, the model correction remains rank-constrained, even though its row and column spaces can evolve during training. This constraint can limit adaptation when heterogeneous data or training from scratch requires updates spanning many directions. The key question is therefore: *Can federated training retain the resource efficiency of low-rank updates without imposing a fixed rank budget on the accumulated model change?*

Gradient low-rank optimization offers a promising starting point. GaLore (Zhao et al., 2024) maintains optimizer states in projected gradient coordinates while updating model weights in the original parameter space. This separates the dimension of the optimization state from the rank of the accumulated model change: successive low-rank increments can span an increasingly rich set of directions as the projection changes. In FL, however, local compression and global aggregation must be designed together. A subspace that captures a client's dominant gradient directions need not preserve the directions required for progress on the global objective.

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.



(a) Test accuracy (%) (b) Effective Rank (c) Space fragmentation (d) Test accuracy (%)

Figure 1: Comparison between FedLore and Local GaLore and FedIT on ViT-Tiny over CIFAR-100 with non-iid data. (a) Test accuracy: FedLore vs FedIT vs Local GaLore. (b) Effective Rank: Dynamic expansion of optimization space of FedLore. (c) Subspace fragmentation of Local GaLore. We measure subspace fragmentation by the difference between projection matrices: $\Delta_{\mathcal{D}} = \frac{1}{Nr}\sum_{i\in\mathcal{N}} \|\Pi_i - \bar{\Pi}\|_F^2, \Pi_i = P_i P_i^\top, \bar{\Pi} = \frac{1}{N}\sum_{i\in\mathcal{N}} \Pi_i$. Client-specific projections introduce severe subspace fragmentation under heterogeneous data. (d) Local GaLore degrades under heterogeneity due to misaligned client-wise optimization subspaces.

We study this mismatch through *subspace fragmentation*: heterogeneous clients select different gradient subspaces, and these client-dependent projections interact with gradient heterogeneity during aggregation (Figure 1). The central issue is not that updates expressed in different coordinates cannot be averaged, they can be reconstructed and aggregated in the original parameter space. Rather, independently chosen projections alter which components of each client gradient survive aggregation, potentially suppressing useful global directions. This interaction admits an exact decomposition. At a common model W, let $G_i = \nabla f_i(W)$, $G = N^{-1}\sum_i G_i$, $\Pi_i = P_i P_i^\top$, and $\bar{\Pi} = N^{-1}\sum_i \Pi_i$. Then $\frac{1}{N}\sum_{i=1}^N \Pi_i G_i = \bar{\Pi} G + b$, $b := \frac{1}{N}\sum_{i=1}^N (\Pi_i - \bar{\Pi})(G_i - G)$. Here, b captures the coupling between projection disagreement and gradient heterogeneity. The second term captures the interaction between projection disagreement and data heterogeneity. It vanishes when all clients share a projector, but need not vanish when each local projector captures most of its own gradient energy. Thus, *accurate local gradient compression does not by itself ensure an informative global update*. Our analysis makes this distinction explicit through a sharp bias bound and a quadratic example in which exact local SVD projections stall away from stationarity.

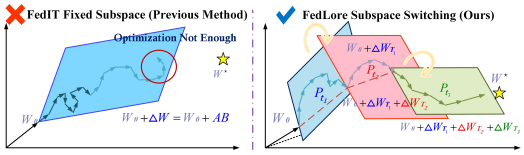


Figure 2: Fixed-rank LoRA vs ours low-rank updates. FedLore shares a low-rank update basis within each round and refreshes it across rounds, allowing the accumulated model change to exceed the per-round rank.

Subspace fragmentation also affects communication. With a basis fixed within each round and zero weight decay, each client's accumulated update remains rank at most r and admits an exact low-rank factorization. Nevertheless, aggregating updates from N distinct client subspaces can produce a matrix of rank up to $\min\{m, n, Nr\}$. Low-rank client updates therefore do not automatically yield a rank-r global update. This exposes two connected challenges:

Challenge 1: *Client-specific subspaces increase the cost of exact communication.*

Challenge 2: *Federated subspace fragmentation under heterogeneous clients.*

We propose FedLore, a federated optimization framework that shares optimization coordinates within each round and refreshes them across rounds. Clients optimize in a common orthonormal basis, upload their accumulated low-rank coordinates, and provide candidate subspaces for the next round. The server constructs the next shared basis from the leading eigenspace of the averaged candidate projectors, making subspace aggregation invariant to the signs and rotations of the candidate bases.

This design separates two roles of low-rank structure. *Within a round*, sharing the basis makes coordinate aggregation exactly equivalent to averaging the local endpoint models, without additional rank truncation. Under full participation with synchronized model caches, the aggregated coordinates and the next basis suffice for low-rank model synchronization. *Across rounds*, refreshing the basis allows the accumulated model change to extend beyond a single rank-r correction (Figures 2). Sharing removes the projection–heterogeneity coupling term, whereas subspace selection determines how

much of the global gradient is retained. We distinguish these properties explicitly: our convergence guarantee requires global-gradient coverage, which agreement alone does not ensure.

Contributions. Our main contributions are as follows:
• We identify *subspace fragmentation*: client-specific projections couple with data heterogeneity, allowing locally accurate compression to suppress global descent and cost of exact communication.
• We propose `FedLore`, which shares a low-rank basis within each round and refreshes it across rounds. This enables exact coordinate aggregation, compact optimizer states, and accumulated updates beyond rank r. With synchronized caches and full participation, communication remains low-rank.
• We show that shared projections eliminate projection–heterogeneity coupling in projected SGD and establish an $O(T^{-1/2})$ stationarity bound under standard smoothness, variance, and heterogeneity assumptions with global-gradient coverage. Across vision and language tasks, including federated pre-training, FedLore combines competitive model quality with reduced communication and optimizer-state memory, while allowing accumulated updates beyond the per-round rank budget.

Related Work

Parameter-Efficient Fine-Tuning for Large Models. As pretrained models scale, parameter-efficient fine-tuning (PEFT) enables adaptation under limited computational and memory budgets (Che et al., 2023). LoRA parameterizes updates to frozen pretrained weights using trainable low-rank factors, substantially reducing gradient and optimizer-state memory (Hu et al., 2021). Subsequent methods extend this approach through adaptive rank allocation and alternative parameterizations (Zhang et al., 2023b; Liu et al., 2024). However, although standard LoRA's update subspaces evolve during optimization, its accumulated weight updates remain rank-constrained, potentially limiting expressivity under substantial distribution shift or when training from scratch. Low-rank factorization can also introduce optimization challenges (Cosson et al., 2023), motivating memory-efficient methods with more flexible weight updates (Zhao et al., 2024).

Extending LoRA to Federated Learning. LoRA is widely used for parameter-efficient federated fine-tuning because it reduces communication and optimizer-state memory costs. FedIT (Zhang et al., 2023a) aggregates client LoRA factors, while subsequent methods refine aggregation, initialization, or factor sharing. FLoRA (Wang et al., 2024) and LoRA-FAIR (Bian et al., 2025) improve aggregation, whereas FedSA-LoRA (Guo et al., 2025) and FFA-LoRA (Sun et al., 2024) modify how the A and B factors are shared or optimized. FR-LoRA (Yan et al., 2025) combines factor reinitialization with residual low-rank updates, allowing the accumulated model change to exceed the per-round rank budget. In contrast, FedLore optimizes projected gradients in a shared, evolving basis and enables exact aggregation of accumulated updates in low-rank coordinates.

Why Local Gradient Projection Fails in Federated Learning

FL Problem Setup. FL aims to optimize model parameters, i.e., minimizing the following problem:

$$f_i(\boldsymbol{W}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\big[F_i(\boldsymbol{W}; \xi_i)\big], f(\boldsymbol{W}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{W}).$$

Here, f_i denotes client i's loss, with expectation over $\xi_i \sim \mathcal{D}_i$; N is the client count, and $\boldsymbol{W}$ is the global model.

FL with Gradient Low-Rank Projection in Local

As a motivating example, we describe low-rank optimizers using a single linear layer W. In round t, client i receives the global weights W^t and computes the gradient of its local objective f_i: $G_i^t = -\nabla_W f_i(W_i^{t,0}; \xi_i^{t,k}) \in \mathbb{R}^{m\times n}$. To construct a local low-rank subspace tailored to its data distribution, the client performs a truncated singular value decomposition (SVD) of G_i^t: $G_i^t \approx U_i^t S_i^t (V_i^t)^\top$, where $U_i^t \in \mathbb{R}^{m\times r}$, $S_i^t \in \mathbb{R}^{r\times r}$, and $V_i^t \in \mathbb{R}^{n\times r}$ contain the top-r singular components. We define the left singular vectors as the projection basis for local optimization: $P_i^t = U_i^t$.

Given the subspace basis P_i^t, the client projects gradients onto this subspace. The projected gradient at step k is $R_i^{t,k} = -(P_i^t)^\top \nabla_W f_i(W_i^{t,k}; \xi_i^{t,k}) \in \mathbb{R}^{r\times n}$. To retain adaptive optimization, we maintain

the optimizer states in the projected coordinates:

$$\begin{aligned} M_i^{t,k+1} &= \beta_1 M_i^{t,k} + (1-\beta_1) R_i^{t,k}, & V_i^{t,k+1} &= \beta_2 V_i^{t,k} + (1-\beta_2)(R_i^{t,k} \odot R_i^{t,k}), \\ \widehat{M}_i^{t,k+1} &= M_i^{t,k+1}/(1-\beta_1^{k+1}), & \widehat{V}_i^{t,k+1} &= V_i^{t,k+1}/(1-\beta_2^{k+1}), \\ N_i^{t,k+1} &= \widehat{M}_i^{t,k+1}/(\sqrt{\widehat{V}_i^{t,k+1}} + \varepsilon), & W_i^{t,k+1} &= W_i^{t,k} + \eta\alpha\, P_i^t N_i^{t,k+1}. \end{aligned} \tag{1}$$

After K local steps, the client uploads full-parameter update $\Delta W_i^t = W_i^{t,K} - W^t$ to the server.

Challenges of Gradient Low-Rank Projection in FL

Despite the widespread use of GaLore in centralized deep learning, its adaptation to federated settings remains largely unexplored. In this subsection, we analyze two fundamental challenges in FL.

Challenge 1: *Client-specific subspaces increase the cost of exact communication.*

Low-rank local updates do not necessarily admit a rank-r representation after aggregation. Consider zero weight decay and a client-specific basis $P_i^t \in \mathbb{R}^{m\times r}$ fixed within round t. After K local steps, client i produces $\Delta W_i^t = \eta\alpha P_i^t \Delta Z_i^t, \Delta Z_i^t = \sum_{k=0}^{K-1} N_i^{t,k+1}$. Each client can therefore upload its update exactly using the factors $(P_i^t, \Delta Z_i^t)$, requiring $r(m+n)$ scalars. However, with N clients, the aggregated update is

$$\Delta W^t = \frac{\eta\alpha}{N} \sum_{i\in\mathcal{S}_t} P_i^t \Delta Z_i^t, \quad \operatorname{rank}(\Delta W^t) \le \min\{m, n, Nr\}.$$

Client-specific subspaces can produce an aggregate of rank exceeding r, precluding an exact rank-r representation. Direct transmission requires either mn scalars for the dense aggregate or $Nr(m+n)$ for concatenated client factors; rank-r truncation introduces error when the aggregate rank exceeds r. With a shared basis, $\Delta W^t = \eta\alpha P^t\big(N^{-1}\sum_{i\in\mathcal{N}} \Delta Z_i^t\big)$ has rank at most r. Clients caching (W^t, P^t) reconstruct the next model from rn aggregated coordinates. Including the next basis costs $r(m+n)$ scalars per client, independently of N.

Challenge 2: *Federated subspace fragmentation under heterogeneous clients.*

In Local GaLore, each client constructs an orthonormal basis P_i from its local gradient via truncated SVD. Heterogeneous data can produce different projection subspaces, a phenomenon we term *subspace fragmentation*. Although local updates remain exactly aggregable in the ambient space, projection disagreement can couple with gradient heterogeneity to impair global descent. At a common model W, let $G_i = \nabla f_i(W)$, $G = N^{-1}\sum_i G_i$, $\Pi_i = P_i P_i^\top$, and $\bar{\Pi} = N^{-1}\sum_i \Pi_i$. Then

$$\frac{1}{N}\sum_{i=1}^N \Pi_i G_i = \bar{\Pi} G + b, \quad b := \frac{1}{N}\sum_{i=1}^N (\Pi_i - \bar{\Pi})(G_i - G).$$

The coupling term b can weaken or cancel global-gradient alignment despite accurate local compression; disagreement alone does not imply harmful interference. Figure 1(c,d) illustrates subspace disagreement and performance degradation under stronger heterogeneity. Shared projections eliminate b exactly, while global descent additionally requires sufficient gradient coverage.

Federated Training via Shared Gradient Low-Rank Projection

We propose `FedLore`, a federated optimization framework using a shared, evolving gradient subspace. It maintains full model weights while storing local optimizer states and communicating updates in low-rank coordinates. The shared basis enables exact aggregation, while refreshing it allows accumulated updates to exceed rank r. Algorithm 1 presents the procedure for a weight matrix $W \in \mathbb{R}^{m\times n}$ under full participation and a constant learning rate.

Shared-subspace construction. Let $P^t \in \mathbb{R}^{m\times r}$ denote the shared orthonormal basis used throughout round t, with $(P^t)^\top P^t = I_r$. The corresponding orthogonal projector is $\Pi^t = P^t(P^t)^\top$. At the end of the round, client i extracts a candidate basis P_i^t from the leading r left singular vectors

Algorithm 1: `FedLore` Training Procedure

Require: Global weights W^0; learning rate η; scale α; rank r; Adam coefficients β_1, β_2; $\epsilon > 0$; communication rounds T; local steps K; client set $\mathcal{N}$ with $|\mathcal{N}| = N$; weight decay $\lambda = 0$.

Ensure: Trained global model W^T.

Initialize a shared orthonormal basis $P^0 \in \mathbb{R}^{m\times r}$
Server broadcasts (W^0, P^0) to all clients
Each client initializes global-model cache: $\widetilde{W}_i \leftarrow W^0$
for $t = 0$ to $T-1$ **do**
  for each client $i \in \mathcal{N}$ **in parallel do**
    `//Low-Rank Model Synchronization`
    if $t > 0$ **then**
      Receive (P^t, Δ^{t-1}) from the server
      $\widetilde{W}_i \leftarrow \widetilde{W}_i + \eta\alpha P^{t-1}\Delta^{t-1}$
    end if
    $W_i^{t,0} \leftarrow \widetilde{W}_i$, $M_i^{t,0} \leftarrow 0$, $V_i^{t,0} \leftarrow 0$, $\Delta Z_i^t \leftarrow 0$
    `//Local Projected Adam Updates`
    for $k = 0$ to $K-1$ **do**
      $G_i^{t,k} \leftarrow -\nabla_W f_i(W_i^{t,k}; \xi_i^{t,k}) \in \mathbb{R}^{m\times n}$; $R_i^{t,k} \leftarrow (P^t)^\top G_i^{t,k} \in \mathbb{R}^{r\times n}$
      $M_i^{t,k+1} \leftarrow \beta_1 M_i^{t,k} + (1-\beta_1)R_i^{t,k}$; $V_i^{t,k+1} \leftarrow \beta_2 V_i^{t,k} + (1-\beta_2)(R_i^{t,k} \odot R_i^{t,k})$
      $\widehat{M}_i^{t,k+1} \leftarrow M_i^{t,k+1}/(1-\beta_1^{k+1})$; $\widehat{V}_i^{t,k+1} \leftarrow V_i^{t,k+1}/(1-\beta_2^{k+1})$
      $N_i^{t,k+1} \leftarrow \widehat{M}_i^{t,k+1}/(\sqrt{\widehat{V}_i^{t,k+1}} + \epsilon)$; $W_i^{t,k+1} \leftarrow W_i^{t,k} + \eta\alpha P^t N_i^{t,k+1}$
      $\Delta Z_i^t \leftarrow \Delta Z_i^t + N_i^{t,k+1}$
    end for
    `//Local Candidate Subspace`
    Reuse the last local gradient $G_i^{t,K-1}$; $P_i^t \leftarrow \mathrm{TopLeftSVD}_r(G_i^{t,K-1}) \in \mathbb{R}^{m\times r}$
    Upload $(\Delta Z_i^t, P_i^t)$ to the server; Cache the shared basis P^t for reconstructing the aggregated update.
  end for
  `//Server-Side Update Aggregation`
  $\Delta^t \leftarrow \frac{1}{N}\sum_{i\in\mathcal{N}} \Delta Z_i^t$; $W^{t+1} \leftarrow W^t + \eta\alpha P^t \Delta^t$
  `//Shared-Subspace Construction`
  $B^t \leftarrow \frac{1}{\sqrt{N}}[P_1^t, \ldots, P_N^t] \in \mathbb{R}^{m\times Nr}$; $P^{t+1} \leftarrow \mathrm{TopLeftSVD}_r(B^t) \in \mathbb{R}^{m\times r}$
end for

of its last local gradient $G_i^{t,K-1}$. Directly averaging these bases is sensitive to arbitrary signs and within-subspace rotations. We instead aggregate their orthogonal projectors:

$$A_{\mathrm{proj}}^t = \frac{1}{N}\sum_{i\in\mathcal{N}} P_i^t (P_i^t)^\top, P^{t+1} \in \operatorname*{arg\,max}_{P^\top P = I_r} \mathrm{tr}\big(P^\top A_{\mathrm{proj}}^t P\big).$$

Thus, P^{t+1} spans a leading rank-r eigenspace of A_{proj}^t and maximizes the average overlap with the client candidate subspaces. The aggregated operator is invariant to replacing any P_i^t with $P_i^t Q_i$ for an orthogonal matrix Q_i. To avoid explicitly forming the dense projector average, we use $B^t = \frac{1}{\sqrt{N}}[P_1^t, \ldots, P_N^t]$, $A_{\mathrm{proj}}^t = B^t(B^t)^\top$, and obtain P^{t+1} from the leading left singular vectors of B^t. Randomized truncated SVD can approximate these subspaces without forming A_{proj}^t. The server broadcasts the resulting basis for round $t+1$.

Local optimization in shared coordinates. Each client starts from the synchronized global model, $W_i^{t,0} = W^t$, and initializes $M_i^{t,0} = V_i^{t,0} = 0$. For local step k, it computes a stochastic negative gradient and its low-rank coordinates:

$$G_i^{t,k} = -\nabla_W f_i(W_i^{t,k}; \xi_i^{t,k}), R_i^{t,k} = (P^t)^\top G_i^{t,k} \in \mathbb{R}^{r\times n}.$$

The first- and second-moment states are maintained entirely in this coordinate system:

$$\begin{aligned} M_i^{t,k+1} &= \beta_1 M_i^{t,k} + (1-\beta_1)R_i^{t,k}, & V_i^{t,k+1} &= \beta_2 V_i^{t,k} + (1-\beta_2)\big(R_i^{t,k} \odot R_i^{t,k}\big), \\ \widehat{M}_i^{t,k+1} &= M_i^{t,k+1}/(1-\beta_1^{k+1}), & \widehat{V}_i^{t,k+1} &= V_i^{t,k+1}/(1-\beta_2^{k+1}). \end{aligned} \tag{2}$$

The normalized direction and weight update are $N_i^{t,k+1} = \frac{\widehat{M}_i^{t,k+1}}{\sqrt{\widehat{V}_i^{t,k+1}}+\epsilon}$, $W_i^{t,k+1} = W_i^{t,k} + \eta\alpha P^t N_i^{t,k+1}$. All operations in the normalization are elementwise. This is AdamW with weight

decay set to zero. Resetting the moments each round avoids transferring optimizer states between changing coordinate systems. Each weight increment has rank at most r.

Exact low-rank aggregation and synchronization. With P^t fixed within each round, local updates factorize exactly as $\Delta Z_i^t = \sum_{k=0}^{K-1} N_i^{t,k+1}$ and $W_i^{t,K} - W^t = \eta\alpha P^t \Delta Z_i^t$. Clients upload $(\Delta Z_i^t, P_i^t)$ for aggregation and basis construction. The server computes $\Delta^t = N^{-1}\sum_{i\in\mathcal{N}} \Delta Z_i^t$ and $W^{t+1} = W^t + \eta\alpha P^t \Delta^t$, exactly matching endpoint averaging without additional compression. Under full participation, clients cache (W^t, P^t) and reconstruct W^{t+1} upon receiving (Δ^t, P^{t+1}). Excluding initialization, communication costs $r(m+n)$ scalars per matrix, per client, per round in each direction. The two optimizer moments occupy $2rn$ scalars, excluding weights, gradients, bases, activations.

THEORETICAL ANALYSIS

Let $f(W) := \frac{1}{N}\sum_{i=1}^N f_i(W)$, $f^\star := \inf_W f(W)$. For each client i and round t, let $P_i^t \in \mathbb{R}^{m\times r}$ have orthonormal columns, and define $\Pi_i^t := P_i^t (P_i^t)^\top$, $A_t := \frac{1}{N}\sum_{i=1}^N \Pi_i^t$. In the shared-projection setting, $\Pi_i^t = \Pi^t$ for all i, so $A_t = \Pi^t$.

Assumption 1 (Smoothness and lower boundedness)**.** *Each local objective $f_i : \mathbb{R}^{m\times n} \to \mathbb{R}$ is differentiable and L-smooth: for all X, Y, $\|\nabla f_i(X) - \nabla f_i(Y)\|_F \le L\|X - Y\|_F$. The global objective satisfies $f^\star > -\infty$.*

Assumption 2 (Stochastic gradient oracle)**.** *Define $\varepsilon_i^{t,k} := g_i^{t,k} - \nabla f_i(W_i^{t,k})$. For every client i, round t, and step k, $\mathbb{E}[\varepsilon_i^{t,k} \mid \mathcal{H}_{t,k}] = 0, \mathbb{E}[\|\varepsilon_i^{t,k}\|_F^2 \mid \mathcal{H}_{t,k}] \le \sigma^2$. For distinct clients $i \ne j$, projected noises at same local step are conditionally uncorrelated: $\mathbb{E}\Big[\Big\langle \Pi_i^t \varepsilon_i^{t,k}, \Pi_j^t \varepsilon_j^{t,k}\Big\rangle_F \,\Big|\, \mathcal{H}_{t,k}\Big] = 0$.*

Assumption 3 (Bounded heterogeneity)**.** *There exists $\zeta \ge 0$ such that, for all W, $\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(W) - \nabla f(W)\|_F^2 \le \zeta^2$.*

Assumption 4 (Aggregate global-gradient coverage)**.** *There exists $\mu_P \in (0, 1]$ such that, almost surely, for every round t, $\langle \nabla f(W^t), A_t \nabla f(W^t)\rangle_F = \frac{1}{N}\sum_{i=1}^N \|\Pi_i^t \nabla f(W^t)\|_F^2 \ge \mu_P \|\nabla f(W^t)\|_F^2$.*

Assumption 4 requires coverage only along the current global gradient. For shared projections, it becomes $\|\Pi^t \nabla f(W^t)\|_F^2 \ge \mu_P \|\nabla f(W^t)\|_F^2$; for client-specific projections, coverage is through A_t. Both methods are compared using a common valid μ_P. Orthogonality directly implies $\|\Pi_i^t X\|_F \le \|X\|_F$ and $0 \preceq A_t \preceq I$, without additional assumptions.

We analyze the projected-SGD variants of `Local GaLore` and `FedLore` under the sampling and conditioning convention stated above. Let S denote the number of participating clients per round, and define $\Delta := f(W^0) - f^\star, \omega_S := \frac{N-S}{S(N-1)}$. Full participation corresponds to $S = N$ and $\omega_S = 0$.

For each round t, write $G_t := \nabla f(W^t), d_i^t := \nabla f_i(W^t) - G_t, A_t := \frac{1}{N}\sum_{i=1}^N \Pi_i^t$, where $\Pi_i^t = P_i^t (P_i^t)^\top$. Define the projection-induced aggregation bias $b_t := \frac{1}{N}\sum_{i=1}^N (\Pi_i^t - A_t) d_i^t$. Since $\sum_i d_i^t = 0$, the round-start projected direction admits the exact decomposition $\frac{1}{N}\sum_{i=1}^N \Pi_i^t \nabla f_i(W^t) = A_t G_t + b_t$. For a horizon T, choose $\eta = \frac{c}{\alpha L K\sqrt{T}}, 0 < c \le \frac{\mu_P}{16}$, and define the common remainder $\mathcal{R}_T := \frac{4L\Delta}{\mu_P c\sqrt{T}} + \frac{40c^2\zeta^2}{3\mu_P^2 T} + \frac{12c^2\sigma^2}{\mu_P^2 KT} + \frac{4c}{\mu_P\sqrt{T}}\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right)$.

Theorem 1 (Nonconvex convergence of `Local GaLore-SGD`)**.** *Suppose Assumptions 1–4. With the stepsize specified above, `Local GaLore-SGD` satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W^t)\|_F^2 \le \mathcal{R}_T + \frac{8}{\mu_P^2 T}\sum_{t=0}^{T-1} \mathbb{E}\|b_t\|_F^2. \tag{3}$$

Moreover, let $\bar{v}_T := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|A_t - A_t^2\|_2$. Then $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W^t)\|_F^2 \le \mathcal{R}_T + \frac{8\zeta^2}{\mu_P^2}\bar{v}_T$.

Theorem 2 (Nonconvex convergence of `FedLore-SGD`)**.** *Suppose Assumptions 1– 4 and the sampling and conditioning convention hold. If all clients share the same round-wise projector, $\Pi_i^t = \Pi^t$, then $b_t = 0$ for every round. With the same stepsize, `FedLore-SGD` satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W^t)\|_F^2 \le \mathcal{R}_T. \tag{4}$$

For fixed problem parameters, K, S, and c, its averaged squared gradient norm is $O(T^{-1/2})$.

Table 1: Test accuracy (%) under Dir-0.1 over $T = 100$. Experiments are conducted with 50 clients, batch size 16, local steps $K = 50$, and LoRA rank $r = 8$, projection rank $r = 8$. Best results are shown in bold and second-best results are underlined.

Method	Venue	CIFAR-100		Tiny-ImageNet		Food-101		Avg.
		Swin-Base	ViT-Base	Swin-Base	ViT-Base	Swin-Base	ViT-Base	
FedIT	ICASSP'24	90.17 ± 0.28	89.57 ± 0.28	86.50 ± 0.22	84.88 ± 0.12	91.55 ± 0.18	91.02 ± 0.20	88.95
FFA-LoRA	ICLR'24	90.07 ± 0.18	86.26 ± 0.14	85.34 ± 0.18	77.89 ± 0.11	91.25 ± 0.12	90.68 ± 0.22	86.92
FlexLoRA	ICLR'25	90.91 ± 0.19	89.59 ± 0.19	85.94 ± 0.19	85.55 ± 0.19	91.82 ± 0.16	91.34 ± 0.10	89.19
LoRA-FAIR	ICCV'25	90.26 ± 0.19	89.86 ± 0.19	86.59 ± 0.15	85.11 ± 0.19	91.41 ± 0.12	91.23 ± 0.14	89.08
RoLoRA	NeurIPS'25	90.55 ± 0.26	88.55 ± 0.26	86.03 ± 0.28	82.97 ± 0.28	90.54 ± 0.15	90.32 ± 0.08	88.16
FLoRA	NeurIPS'24	86.32 ± 0.19	85.48 ± 0.19	87.20 ± 0.12	83.69 ± 0.15	87.12 ± 0.21	86.56 ± 0.11	86.06
FRLoRA	ICLR'25	87.17 ± 0.16	66.80 ± 0.19	76.45 ± 0.22	42.04 ± 0.08	88.12 ± 0.26	87.96 ± 0.17	74.76
Local GaLore	-	90.22 ± 0.12	89.62 ± 0.21	86.47 ± 0.16	84.96 ± 0.12	91.61 ± 0.11	91.27 ± 0.15	89.03
FedFull	AISTATS'17	$\underline{90.81 \pm 0.12}$	$\underline{90.26 \pm 0.10}$	$\underline{87.50 \pm 0.08}$	$\underline{87.02 \pm 0.20}$	$\underline{92.25 \pm 0.12}$	$\underline{92.12 \pm 0.10}$	$\underline{89.99}$
FedLore	Ours	$\mathbf{90.85 \pm 0.08}$	$\mathbf{90.55 \pm 0.08}$	$\mathbf{87.52 \pm 0.10}$	$\mathbf{87.11 \pm 0.10}$	$\mathbf{92.32 \pm 0.16}$	$\mathbf{92.21 \pm 0.17}$	**90.09**

Table 2: Test accuracy (%) of RoBERTa-Base with LoRA on six text classification tasks under FL. Experiments are conducted for $T = 100$ rounds with 20 clients, Dir-0.5 data partition, batch size 16, local steps $K = 50$, and LoRA rank $r = 8$.

Method	Venue	SNLI	AG News	QQP	DBPedia 14	QNLI	MNLI	Avg.
FedIT	ICASSP'24	86.50 ± 0.13	92.68 ± 0.09	85.06 ± 0.14	98.94 ± 0.07	89.47 ± 0.11	81.77 ± 0.15	89.07
FFA-LoRA	ICLR'24	80.70 ± 0.17	90.14 ± 0.12	81.48 ± 0.16	98.22 ± 0.10	85.61 ± 0.14	72.44 ± 0.18	84.77
FlexLoRA	ICLR'25	86.64 ± 0.15	93.30 ± 0.11	63.18 ± 0.19	98.06 ± 0.09	49.46 ± 0.20	35.44 ± 0.18	71.01
LoRA-FAIR	ICCV'25	86.76 ± 0.10	92.76 ± 0.08	85.33 ± 0.12	98.88 ± 0.07	89.47 ± 0.09	82.04 ± 0.11	89.21
RoLoRA	NeurIPS'25	85.95 ± 0.12	92.06 ± 0.10	84.51 ± 0.13	98.78 ± 0.08	88.30 ± 0.12	80.58 ± 0.14	88.36
FLoRA	NeurIPS'24	69.36 ± 0.18	89.77 ± 0.13	79.78 ± 0.15	98.22 ± 0.11	81.93 ± 0.16	78.23 ± 0.17	82.88
FRLoRA	ICLR'25	86.67 ± 0.11	93.00 ± 0.09	85.04 ± 0.12	98.90 ± 0.08	89.16 ± 0.10	81.81 ± 0.13	89.10
Local GaLore	-	86.65 ± 0.11	92.71 ± 0.12	85.21 ± 0.15	98.96 ± 0.08	89.52 ± 0.12	82.07 ± 0.16	89.19
FedFull	AISTATS'17	$\underline{88.02 \pm 0.09}$	$\underline{93.72 \pm 0.08}$	$\underline{86.31 \pm 0.14}$	$\underline{99.02 \pm 0.17}$	$\underline{90.16 \pm 0.10}$	$\underline{83.85 \pm 0.16}$	$\underline{90.18}$
FedLore	Ours	$\mathbf{88.19 \pm 0.08}$	$\mathbf{93.97 \pm 0.07}$	$\mathbf{86.79 \pm 0.14}$	$\mathbf{99.10 \pm 0.17}$	$\mathbf{90.48 \pm 0.10}$	$\mathbf{84.11 \pm 0.16}$	**90.44**

Theoretical interpretation. Theorems 1 and 2 isolate the effect of projection disagreement under a common global-gradient coverage condition. Client-specific projections introduce the coupling term b_t between projection disagreement and gradient heterogeneity. Its magnitude satisfies $\|b_t\|_F^2 \leq \|A_t - A_t^2\|_2 \frac{1}{N} \sum_{i=1}^N \|d_i^t\|_F^2$. With a shared projector, the zero-mean identity $\sum_i d_i^t = 0$ is preserved after projection, so this bias vanishes exactly. Using common valid constants for both methods, the shared-projection bound therefore removes the additional bias term in equation 3. The remaining terms account for initial suboptimality, local drift, client sampling, and stochastic-gradient noise, and decay under the stated horizon-dependent stepsize. The local-projection bound also vanishes if its average bias energy tends to zero; a nonvanishing term in an upper bound alone does not establish an unavoidable error floor. The strict quadratic separation in Theorem 6 shows that the obstruction can nevertheless occur: exact local SVD projections may stall at a nonstationary point, whereas a globally informed shared projection converges to the optimum.

A sharper guarantee for shared projections. The shared-projection structure yields a stronger result than simply setting $b_t = 0$ in the generic bound. Let $h_t = K\gamma_t$, $\Lambda_T = \sum_{t<T} h_t$, and assume $Lh_t \leq 1/8$. As shown in Theorem 5, without a global-gradient coverage assumption 4, $\frac{1}{\Lambda_T} \sum_{t<T} h_t \mathbb{E}\big[\|\Pi^t \nabla f(W^t)\|_F^2\big] \leq B_T$. Consequently, $\frac{1}{\Lambda_T} \sum_{t<T} h_t \mathbb{E}\big[\|\nabla f(W^t)\|_F^2\big] \leq B_T + \frac{1}{\Lambda_T} \sum_{t<T} h_t \mathbb{E}\big[\|(I - \Pi^t)\nabla f(W^t)\|_F^2\big]$. Thus, shared projections control projected stationarity without coverage, while full-gradient stationarity additionally requires control of the uncovered component. If $\mathbb{E}\|\Pi^t \nabla f(W^t)\|_F^2 \geq \mu \mathbb{E}\|\nabla f(W^t)\|_F^2$ for every t, the full-gradient bound becomes B_T/μ. The stepsize condition $Lh_t \leq 1/8$ does not depend on μ.

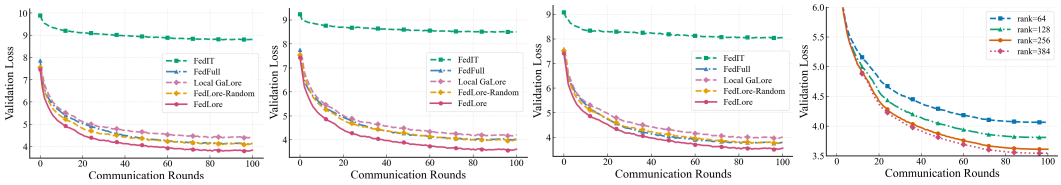
EXPERIMENTS

EXPERIMENTS SETTINGS

Datasets. We evaluate FedLore on both vision and language tasks. (*i*) For image classification, we use CIFAR-100 Krizhevsky et al. (2009), Tiny-ImageNet Le & Yang (2015), Food-101 Bossard

Table 3: Comparison with low-rank federated algorithms for pre-training LLaMA models on the C4 dataset. We report validation loss (Val Loss), validation perplexity (Val PPL), communication cost (Comm.), and local training time. All methods are trained for 100 communication rounds (20 clients, batch size 16, $K = 50$). The projection rank of `FedLore` is identical to LoRA rank.

Method	**60M**		**130M**		**350M**		**1B**		**Comm.** $\downarrow$	**Time/Round** $\downarrow$ on LLaMA 350M (s)
	Loss	PPL	Loss	PPL	Loss	PPL	Loss	PPL		
FedIT	8.811	6707.62	8.544	5135.85	8.072	3203.50	7.561	1921.77	$O(rd)$	118.6
Local GaLore	4.326	75.64	4.215	67.69	3.962	52.56	3.781	43.86	$O(d^2)$	136.6
FedFull	4.134	62.43	4.056	57.74	3.798	44.61	3.561	35.20	$O(d^2)$	132.1
`FedLore-Random`	4.112	61.07	4.036	56.60	3.781	43.86	3.523	33.89	$O(rd)$	119.8
`FedLore`	**3.833**	**46.20**	**3.611**	**37.00**	**3.554**	**34.95**	**3.345**	**28.36**	$O(rd)$	122.4
r/d_{model}	128 / 512		256 / 768		256 / 1024		512 / 2048			



(a) LLaMA 60M, C4 (b) LLaMA 130M, C4 (c) LLaMA 350M, C4 (d) Effect of rank r

Figure 3: (a,b,c): Validation loss convergence curves on C4 with LLaMA models of different scales. (d): Effect of the projection rank r on validation loss for LLaMA 350M.

et al. (2014). (*ii*) For NLP tasks pretraining, we adopt C4 Raffel et al. (2020) dataset. For language understanding, we evaluate SNLI, AG News, QQP, DBPedia-14, QNLI, and MNLI. To simulate data heterogeneity across clients, we follow the Dirichlet partitioning scheme Hsu et al. (2019).

Model Architectures. We explore a variety of model types: (*i*) ViT-Base Dosovitskiy et al. (2020), Swin-Base Liu et al. (2021) and ViT-Tiny Dosovitskiy et al. (2020), and (*ii*) LLaMA Touvron et al. (2023) for large-scale language model and RoBERTa-Base Liu et al. (2019).

Baselines. We compare our method against sota FL algorithms: FedIT Zhang et al. (2023a), FFA-LoRA Sun et al. (2024), FlexLoRA Bai et al. (2024), RoLoRA Chen et al. (2024), FLoRA Wang et al. (2024), FRLoRA Yan et al. (2025), LoRA-FAIR Bian et al. (2025), FedFull, and Local Galore.

Hyperparameter Settings. The lr is selected from $\{10^{-4},\ 3\times10^{-4},\ 5\times10^{-4},\ 8\times10^{-4},\ 10^{-3},\ 3\times10^{-3},\ 5\times10^{-3}\}$. Their weight decay is selected from $\{0,\ 10^{-3},\ 3\times10^{-3},\ 1\times10^{-2},\ 3\times10^{-2}\}$. For `FedLore`, $\alpha = 1$. We apply cosine learning rate decay. Additional hyperparameter configurations are detailed in the **Appendix**. We release all code to ensure full reproducibility upon acceptance. All results are averaged over 5 runs with seeds 42, 43, 44, 45, 46 on NVIDIA RTX 4090 GPU.

Results on Fine-tuning.

Vision Benchmark Results. As shown in Table 1, `FedLore` consistently matches or exceeds the performance of existing federated fine-tuning approaches under the challenging Dir-0.1 heterogeneous setting. It achieves the highest average accuracy of 90.09%, compared with 89.99% for the full-parameter baseline `FedFull`. Across all six dataset combinations, `FedLore` either achieves the best result or remains statistically competitive with the strongest baseline. The advantages are particularly noticeable on ViT-Base, where `FedLore` reaches 90.55% on CIFAR-100 and 87.11% on Tiny-ImageNet. Overall, these results indicate that dynamic low-rank subspace optimization provides a more flexible and effective alternative to LoRA training in heterogeneous federated environments.

Language Benchmark Results. As shown in Table 2, `FedLore` delivers the strongest overall performance on federated language understanding. It ranks first on all six tasks and achieves the best average accuracy of 90.44%, surpassing `FedFull` by 0.26 points and the best LoRA-based baseline by more than 1 point. The improvements are particularly pronounced on QQP, QNLI, and MNLI, suggesting that `FedLore` is more effective than LoRA training in handling heterogeneous NLP tasks. These results confirm that `FedLore` is not limited to vision models, but also provides a strong and communication-efficient alternative to FL for large language transformers.

Table 4: Communication and optimizer memory on LLaMA 350M (left), and component ablations on C4 pre-training (right). All costs are in GB.

Method	Comm.	Opt. Memory	Ablation Setting	Val Loss / PPL	Comm. ↓
FedIT	0.36	0.72	Local GaLore	3.962 / 52.56	0.68
FedFull	0.68	1.37	FedLore	**3.554 / 34.95**	0.36
Local GaLore	0.68	0.54	w/o Shared Projector	3.890 / 48.91	0.68
FedLore	0.36	0.54	w/o Dynamic Subspace	5.952 / 384.7	0.18

RESULTS ON LLAMA PRE-TRAINING.

As shown in Table 3 and Figure 3, FedLore consistently outperforms existing federated baselines across different LLaMA scales. Compared with FedIT, FedLore significantly reduces validation perplexity, from 6707.62 to 46.20 on LLaMA-60M and from 1921.77 to 28.36 on LLaMA-1B, demonstrating that fixed low-rank adapters severely limit the optimization capacity for federated pre-training. Meanwhile, FedLore achieves better performance than full-parameter training while retaining low-rank communication and memory efficiency, e.g., obtaining a PPL of 28.36 on LLaMA-1B compared with 35.20 of FedFull. The inferior performance of FedLore−Random, which uses a shared random projection, further verifies that the gains come from adaptive and data-driven subspace evolution rather than shared projection alone.

ABLATION STUDY

Effect of Subspace Rank. Figure 3(d) studies the effect of the projection rank r on LLaMA-350M pre-training. Larger ranks accelerate convergence and reduce the final validation loss, since the projected subspace captures more informative update directions. The improvement is significant from $r = 64$ to $r = 256$, while increasing r to 384 brings only marginal gains. This indicates that FedLore achieves a favorable accuracy and communication trade-off with a moderate rank.

Ablation study. Table 4 examines the roles of shared projections and dynamic subspace updates in LLaMA-350M pre-training on C4. Removing the shared projector increases validation loss from 3.554 to 3.890 and communication from 0.36 to 0.68 GB under the evaluated protocol, supporting the benefits of shared optimization coordinates for both optimization and communication efficiency. Fixing the shared subspace reduces communication to 0.18 GB but increases validation loss to 5.952, indicating that subspace refresh is important for maintaining optimization capacity throughout training. Together, these results support the complementary roles of subspace sharing and refresh in balancing model quality and communication cost.

Communication and Memory Efficiency Analysis. Table 4 summarizes the communication and memory efficiency on LLaMA 350M. FedFull incurs the largest cost, requiring 0.68 GB communication cost and 1.37 GB optimizer memory. Local GaLore reduces optimizer memory to 0.54 GB, but still communicates full updates. FedIT lowers communication to 0.36 GB, while requiring 0.72 GB optimizer memory. By contrast, FedLore achieves the best trade-off, with 0.36 GB total communication and 0.54 GB optimizer memory. This represents approximately 47% lower communication than FedFull/Local GaLore.

CONCLUSION

We introduced FedLore, a communication- and memory-efficient framework for FL foundation-model training. A shared, evolving basis enables exact low-rank aggregation without imposing a fixed rank constraint on accumulated updates. Our analysis characterizes projection–heterogeneity coupling and establishes projected-SGD convergence under global-gradient coverage. Experiments across vision, language, and federated pre-training support the approach. Future work will develop coverage-aware subspace selection and efficient synchronization for intermittent participation.

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

T. Che, J. Liu, Y. Zhou, et al. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7871–7888, 2023.

Shuangyi Chen, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish J Khisti. Robust federated finetuning of foundation models via alternating minimization of lora. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*, 2024.

Romain Cosson, Ali Jadbabaie, Anuran Makur, et al. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, volume 2025, pp. 99003–99027, 2025.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

H. B. McMahan, Eider Moore, Daniel Ramage, S.C.D. Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016. URL https://api.semanticscholar.org/CorpusID:14955348.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*, volume abs/2403.12313, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024.

Xinghao Wu, Xuefeng Liu, Jianwei Niu, Haolin Wang, Shaojie Tang, and Guogang Zhu. Fedlora: When personalized federated learning meets low-rank adaptation. 2024.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pfedlora: Model-heterogeneous personalized federated learning with lora tuning. 2023. URL https://api.semanticscholar.org/CorpusID:267627316.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919, 2023a. URL https://api.semanticscholar.org/CorpusID:258564501.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. In *Forty-first International Conference on Machine Learning*, 2024.

ETHICS STATEMENT

This work studies communication- and memory-efficient federated learning using existing vision and language benchmarks: CIFAR-100, Tiny-ImageNet, Food-101, C4, SNLI, AG News, QQP, DBPedia-14, QNLI, and MNLI. Our experiments simulate federated clients using benchmark partitions rather than collecting data from real users. Although FedLore keeps training data local within the federated protocol, communicated updates and shared subspaces may still expose information about local data. Neither federated training nor low-rank projection alone provides a formal privacy guarantee; FedLore does not establish differential privacy or resistance to inference attacks. Models trained with this approach may also inherit biases or harmful content from the underlying datasets and pretrained models. Deployment with sensitive data therefore requires additional privacy safeguards and application-specific evaluation.

REPRODUCIBILITY STATEMENT

The paper provides the FedLore algorithm and its theoretical assumptions, with detailed derivations and proofs in the appendix. The experimental settings in the main text and the appendices on federated pre-training and fine-tuning describe the datasets, model architectures, client partitions, participation settings, communication rounds, local optimization steps, projection ranks, optimizer configurations, and hyperparameter selection. The paper also reports the hardware and the five random seeds used for repeated experiments. We will publicly release the implementation and experiment scripts upon acceptance.

AI USE STATEMENT

We used large language models to assist with language editing, including sentence rephrasing, grammar correction, and readability improvements. We did not use these tools for research ideation, methodology or experiment design, theoretical derivations, method implementation, data generation or processing, or interpretation of results. The authors reviewed and revised all AI-assisted text and take full responsibility for the final manuscript, including its claims, analyses, and conclusions.

APPENDIX CONTENTS

Details of Pre-Training Experiments . . . 14
Architecture and Hyperparameters . . . 14
Federated Pre-Training Setup . . . 14
Optimization Details . . . 14
Communication and Memory Analysis . . . 15
Hardware and Reproducibility . . . 15
Learning Rate Selection . . . 15
Details of Fine-Tuning Experiments . . . 16
Vision Benchmark . . . 16
Language Benchmark . . . 16
Hyperparameter Selection . . . 16
Learning Rate Selection . . . 16
Communication and Memory Overhead . . . 18
Communication Cost . . . 18
Memory Estimates . . . 18
Shared Projections in Federated Optimization . . . 19
Setting and assumptions . . . 19
Projection disagreement and aggregation bias . . . 20
Convergence under partial participation . . . 22
A strict separation with exact local SVD projections . . . 29
Coverage residual of the implemented projector average . . . 30
Coverage of an approximate shared subspace . . . 32
Exact synchronization and optimizer scope . . . 34
Baseline Algorithm . . . 37

DETAILS OF PRE-TRAINING EXPERIMENTS

ARCHITECTURE AND HYPERPARAMETERS

We provide detailed configurations of the LLaMA models and hyperparameters used in federated pre-training experiments. We adopt the LLaMA architecture and evaluate four model scales, including LLaMA-60M, LLaMA-130M, LLaMA-350M, and LLaMA-1B. All models are pretrained on the C4 dataset under the federated setting. For all experiments, we use a maximum sequence length of 256 and a batch size of 16 for each client. The training process contains 100 communication rounds, and each round performs 50 local optimization steps.

Table 5: LLaMA architecture configurations used in federated pre-training. Model names follow the GaLore implementation.

Model	Hidden	Intermediate	Heads	Layers
LLaMA-60M	512	1376	8	8
LLaMA-130M	768	2048	12	12
LLaMA-350M	1024	2736	16	24
LLaMA-1B	2048	5461	32	24

For optimization, we adopt AdamW in the shared low-rank subspace. The learning rate is selected from $\{10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 8\times10^{-4}, 10^{-3}, 3\times10^{-3}, 5\times10^{-3}\}$ according to the validation perplexity. For FedLore, we set the projection rank according to the model scale and use $\alpha = 1$. The learning rate follows cosine decay during training.

FEDERATED PRE-TRAINING SETUP

We follow the standard federated pre-training protocol. The C4 dataset is partitioned among clients using a Dirichlet distribution to simulate heterogeneous data distributions. The detailed federated training configurations are summarized below: Different from LoRA-based federated fine-tuning methods that restrict model updates to fixed adapter spaces, FedLore performs optimization in dynamically evolving shared low-rank subspaces while reconstructing updates in the original parameter space. This enables full-parameter adaptation with low communication and memory overhead.

Table 6: Federated pre-training configurations.

Configuration	Value
Dataset	C4
Number of clients	20
Client participation	100%
Communication rounds	100
Local steps per round	50
Batch size	16
Sequence length	256
Optimizer	AdamW
Weight decay	0
LR scheduler	Cosine decay

OPTIMIZATION DETAILS

For all LLaMA pre-training experiments, we use AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The weight decay is set to zero since weight decay introduces a full-rank component into accumulated local updates, which breaks the low-rank communication property. Specifically, after K local optimization steps, the local update can be written as

$$\Delta W_i^t = \sum_{k=0}^{K-1} \eta \alpha P^t \Delta Z_i^{t,k}, \tag{5}$$

where $\Delta Z_i^{t,k}$ is the low-rank update in the shared projection space. Therefore, only the low-rank representation is communicated between clients and the server.

Communication and Memory Analysis

For a weight matrix $W \in \mathbb{R}^{m \times n}$, FedLore communicates the projected update $\Delta Z \in \mathbb{R}^{r \times n}$, while the shared projection matrix is maintained by the server. The communication complexity is therefore

$$O(r(m+n)), \tag{6}$$

which is comparable to LoRA-based methods. The optimizer states are also maintained in the projected space with memory complexity

$$O(rn). \tag{7}$$

The comparison with existing methods is shown in Table 7.

Table 7: Communication and optimizer memory comparison. $r \ll \min(m, n)$.

Method	Communication Cost	Optimizer Memory
FedIT	$O(2r(m+n))$	$O(r(m+n))$
FedFull	$O(2mn)$	$O(mn)$
Local GaLore	$O(2mn)$	$O(rn)$
FedLore	$O(2r(m+n))$	$O(rn)$

Hardware and Reproducibility

All experiments are conducted on NVIDIA RTX 4090 GPUs. For each experiment, we run five independent trials with random seeds $\{42, 43, 44, 45, 46\}$ and report the average performance. The implementation will be released to facilitate reproducibility.

Learning Rate Selection

To ensure a fair comparison, we perform learning rate tuning for all methods under each model scale. The learning rate is selected from $\{10^{-4},\ 3 \times 10^{-4},\ 5 \times 10^{-4},\ 8 \times 10^{-4},\ 10^{-3},\ 3 \times 10^{-3},\ 5 \times 10^{-3}\}$ according to the validation perplexity on the C4 validation set. Table 8 reports the optimal learning rates used by different methods for different LLaMA model sizes. We observe that full-parameter methods generally require smaller learning rates due to the large optimization space, while low-rank optimization methods can adopt larger learning rates owing to the reduced optimization dimension. FedLore remains stable across different model scales and consistently achieves competitive performance with a relatively large learning rate.

Table 8: Optimal learning rates for different methods and LLaMA model scales. The learning rate is selected according to validation perplexity on C4.

Method	LLaMA-60M	LLaMA-130M	LLaMA-350M	LLaMA-1B
FedIT	1×10^{-3}	8×10^{-4}	5×10^{-4}	5×10^{-4}
Local GaLore	3×10^{-3}	3×10^{-3}	1×10^{-3}	1×10^{-3}
FedFull	1×10^{-3}	8×10^{-4}	5×10^{-4}	5×10^{-4}
FedLore-Random	3×10^{-3}	3×10^{-3}	3×10^{-3}	1×10^{-3}
FedLore	3×10^{-3}	3×10^{-3}	3×10^{-3}	1×10^{-3}

DETAILS OF FINE-TUNING EXPERIMENTS

We provide detailed experimental configurations for the vision and language federated fine-tuning experiments. All experiments are conducted under heterogeneous federated learning settings, where client data are partitioned using the Dirichlet distribution.

VISION BENCHMARK

For vision federated fine-tuning, we evaluate FedLore on three benchmark datasets, including CIFAR-100, Tiny-ImageNet, and Food-101. We adopt two representative vision transformer architectures, i.e., Swin-Base and ViT-Base, as backbone models. Following previous federated fine-tuning works, we simulate data heterogeneity using a Dirichlet distribution with concentration parameter $\alpha = 0.1$. The federated training process contains 100 communication rounds with 50 clients. All clients participate in each communication round. The detailed experimental configurations are summarized in Table 9.

Table 9: Experimental configurations for vision federated fine-tuning.

Configuration	Value
Datasets	CIFAR-100, Tiny-ImageNet, Food-101
Backbone	Swin-Base, ViT-Base
Number of clients	50
Data partition	Dirichlet ($\alpha = 0.1$)
Communication rounds	100
Local steps K	50
Batch size	16
LoRA rank	8
Projection rank	8
Optimizer	AdamW
Learning rate	Grid search
LR scheduler	Cosine decay
Weight decay	0

LANGUAGE BENCHMARK

For language federated fine-tuning, we evaluate FedLore with RoBERTa-Base on six text classification benchmarks: SNLI, AG News, QQP, DBPedia-14, QNLI, and MNLI. The data among clients are partitioned using a Dirichlet distribution with $\alpha = 0.5$. We use 20 clients and train the model for 100 communication rounds. Each client performs 50 local optimization steps in each round. The detailed configurations are summarized in Table 10.

HYPERPARAMETER SELECTION

For all federated fine-tuning experiments, the learning rate is selected from

$$\{10^{-4},\ 3\times10^{-4},\ 5\times10^{-4},\ 8\times10^{-4},\ 10^{-3},\ 3\times10^{-3},\ 5\times10^{-3}\}$$

according to the validation performance. For all methods, we use the same search space to ensure a fair comparison. The LoRA rank and projection rank are both set to 8 following previous parameter-efficient federated fine-tuning approaches. FedLore adopts AdamW optimizer with zero weight decay. The learning rate follows cosine decay during training. All reported results are averaged over five independent runs with random seeds $\{42, 43, 44, 45, 46\}$.

LEARNING RATE SELECTION

To ensure a fair comparison, we perform learning rate tuning for all methods. The learning rate is selected from $\{10^{-4},\ 3\times10^{-4},\ 5\times10^{-4},\ 8\times10^{-4},\ 10^{-3},\ 3\times10^{-3},\ 5\times10^{-3}\}$ according to the validation performance. Tables 11 reports the selected learning rates for vision and language federated fine-tuning experiments, respectively. Different methods use their individually optimized learning rates.

Table 10: Experimental configurations for language federated fine-tuning.

Configuration	Value
Datasets	SNLI, AG News, QQP, DBPedia-14, QNLI, MNLI
Backbone	RoBERTa-Base
Number of clients	20
Data partition	Dirichlet ($\alpha = 0.5$)
Communication rounds	100
Local steps K	50
Batch size	16
LoRA rank	8
Projection rank	8
Optimizer	AdamW
Learning rate	Grid search
LR scheduler	Cosine decay
Weight decay	0

Table 11: Selected learning rates for vision federated fine-tuning experiments.

Method	CIFAR-100		Tiny-ImageNet		Food-101	
	Swin-Base	ViT-Base	Swin-Base	ViT-Base	Swin-Base	ViT-Base
FedIT	$1e^{-3}$	$8e^{-4}$	$1e^{-3}$	$8e^{-4}$	$1e^{-3}$	$8e^{-4}$
FFA-LoRA	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$
FlexLoRA	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$
LoRA-FAIR	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$
RoLoRA	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$
FLoRA	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$
FRLoRA	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$
Local GaLore	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$
FedFull	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$
FedLore	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$

Table 12: Selected learning rates for RoBERTa-Base federated fine-tuning on different NLP datasets.

Method	SNLI	AG News	QQP	DBPedia-14	QNLI	MNLI
FedIT	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$5e^{-4}$
FFA-LoRA	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$1e^{-3}$
FlexLoRA	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$5e^{-4}$
LoRA-FAIR	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$1e^{-3}$
RoLoRA	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$1e^{-3}$
FLoRA	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$3e^{-3}$	$1e^{-3}$	$1e^{-3}$
FRLoRA	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$5e^{-4}$
Local GaLore	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$5e^{-3}$	$3e^{-3}$	$3e^{-3}$
FedFull	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$1e^{-3}$	$5e^{-4}$	$5e^{-4}$
FedLore	$3e^{-3}$	$3e^{-3}$	$3e^{-3}$	$5e^{-3}$	$3e^{-3}$	$3e^{-3}$

COMMUNICATION AND MEMORY OVERHEAD

COMMUNICATION COST

We analyze the communication and memory overhead of FedLore compared with existing federated optimization methods. Consider a weight matrix $W \in \mathbb{R}^{m\times n}$ and a low-rank projection dimension $r \ll \min(m, n)$. For full-parameter federated training (FedFull), each client needs to upload and download the complete model update, resulting in a communication cost of $\mathcal{O}(2mn)$. For LoRA-based federated fine-tuning methods, only the low-rank adapter parameters are communicated. The communication cost is $\mathcal{O}(2r(m + n))$.

The communication comparison is summarized in Table 13.

Table 13: Communication and optimizer memory complexity comparison.

Method	Communication Cost	Optimizer Memory
FedIT	$\mathcal{O}(2r(m+n))$	$\mathcal{O}(r(m+n))$
FedFull	$\mathcal{O}(2mn)$	$\mathcal{O}(mn)$
Local GaLore	$\mathcal{O}(2mn)$	$\mathcal{O}(rn)$
FedLore	$\mathcal{O}(2r(m+n))$	$\mathcal{O}(rn)$

MEMORY ESTIMATES

Since the GPU memory consumption of individual components is difficult to measure precisely, we estimate the memory usage of model weights and optimizer states based on the number of trainable parameters. All memory estimations are calculated using BF16 precision, where each parameter requires 2 bytes. For AdamW optimization, two additional first- and second-order moment states are maintained for trainable parameters. For FedLore, the model weights are stored in the original full-parameter space, while optimizer states are maintained only in the shared low-rank projection space. Therefore, FedLore preserves the expressive capability of full-parameter training while requiring significantly less optimizer memory. Table 14 report the estimated memory usage for different LLaMA model sizes.

Table 14: Memory estimation of optimizer states under BF16 precision.

Method	60M	130M	350M	1B
Full-Rank	0.23G	0.51G	1.37G	5.20G
GaLore	0.13G	0.28G	0.54G	1.78G
Low-Rank	0.17G	0.37G	0.72G	2.38G
LoRA	0.17G	0.37G	0.72G	2.38G
FedLore	0.13G	0.28G	0.54G	1.78G

Overall, FedLore achieves the best trade-off between memory efficiency and optimization capability. Unlike LoRA-based methods, FedLore updates the full model parameters through dynamically evolving low-rank gradient subspaces. Meanwhile, compared with full-parameter training, FedLore significantly reduces optimizer memory by maintaining adaptive states only in the projected space.

SHARED PROJECTIONS IN FEDERATED OPTIMIZATION

We analyze projected SGD to isolate the effect of shared optimization subspaces. We first characterize the exact coupling between client-specific projectors and gradient heterogeneity. A unified convergence bound separates this coupling from stochastic noise, client sampling, and local drift. For shared projectors, a sharper analysis controls projected stationarity without a coverage assumption and exposes the uncovered gradient as an explicit residual. We then bound that residual for FedLore's candidate-projector average and identify sufficient conditions for full-gradient convergence. Exact coordinate aggregation is also established for adaptive updates; the stationarity results themselves do not analyze coordinatewise Adam.

SETTING AND ASSUMPTIONS

Let

$$f(W) = \frac{1}{N}\sum_{i=1}^{N} f_i(W), \qquad W \in \mathbb{R}^{m\times n}, \qquad N \geq 2.$$

We use the Frobenius inner product $\langle X, Y\rangle_F = \operatorname{tr}(X^\top Y)$, the Frobenius norm $\|X\|_F$, and the operator norm $\|X\|_2$. For each round t, let $\mathcal{F}_t$ contain the complete history and the round-start model W^t and projectors

$$\Pi_i^t = P_i^t (P_i^t)^\top, \qquad (P_i^t)^\top P_i^t = I_r, \qquad P_i^t \in \mathbb{R}^{m\times r}.$$

The rank satisfies $1 \leq r \leq m$. Here P_i^t is a basis and Π_i^t is an orthogonal projection operator. Projectors may depend on past iterates, past data, or a separate construction batch already included in $\mathcal{F}_t$. They remain fixed during the round. The shared case is $\Pi_i^t = \Pi^t$ for all i.

Sampling and the stochastic oracle. After $\mathcal{F}_t$ is fixed, a subset $\mathcal{S}_t$ of size S is sampled uniformly without replacement. Conditional on $\mathcal{F}_t$, sampling is independent of the fresh within-round oracle streams. Each virtual local trajectory is a function only of $\mathcal{F}_t$ and its client oracle stream, not of $\mathcal{S}_t$; only selected trajectories are executed. This conditional independence justifies averaging virtual trajectories over the uniformly sampled subset. Basis construction uses earlier draws or a separate construction batch; the local optimization draws below are fresh. We use positive stochastic gradients throughout. Unless a result explicitly allows round-dependent stepsizes, let $\gamma = \eta\alpha > 0$ be constant. There is no weight decay in the stationarity analysis. The updates are

$$W_i^{t,0} = W^t,$$
$$W_i^{t,k+1} = W_i^{t,k} - \gamma \Pi_i^t g_i^{t,k}, \quad k = 0, \ldots, K-1, \tag{8}$$
$$W^{t+1} = W^t - \frac{\gamma}{S}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1} \Pi_i^t g_i^{t,k}. \tag{9}$$

Let $\mathcal{H}_{t,k}$ include $\mathcal{F}_t$, the sampled subset, and all virtual oracle draws at steps strictly less than k. In particular, $W_i^{t,k}$ is $\mathcal{H}_{t,k}$-measurable. The current noises $\varepsilon_i^{t,k} = g_i^{t,k} - \nabla f_i(W_i^{t,k})$ satisfy

$$\mathbb{E}[\varepsilon_i^{t,k} \mid \mathcal{H}_{t,k}] = 0, \qquad \mathbb{E}[\left\|\varepsilon_i^{t,k}\right\|_F^2 \mid \mathcal{H}_{t,k}] \leq \sigma^2. \tag{10}$$

For distinct clients $i \neq j$, assume

$$\mathbb{E}[\left\langle \Pi_i^t \varepsilon_i^{t,k}, \Pi_j^t \varepsilon_j^{t,k} \right\rangle_F \mid \mathcal{H}_{t,k}] = 0.$$

Independent client oracle draws conditional on $\mathcal{H}_{t,k}$ suffice. Across steps, the martingale-difference property in equation 10 supplies the required orthogonality. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. All objectives are real-valued. The finite second moments needed below follow by induction from deterministic W^0, smoothness, bounded oracle second moments, finite deterministic stepsizes, and contraction of the projectors. Smoothness then also ensures integrability of the objectives.

Assumption 5 (Smoothness and a finite lower bound)**.** *Each f_i is differentiable and L-smooth, with $L > 0$:*

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_F \leq L \|X - Y\|_F .$$

The global objective satisfies $f^\star = \inf_W f(W) > -\infty$. Let W^0 be fixed and $\Delta_0 = f(W^0) - f^\star$.

Assumption 6 (Bounded gradient heterogeneity)**.** *For all W,*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(W)-\nabla f(W)\|_F^2\le\zeta^2.$$

It suffices that this bound hold almost surely at the round-start iterates.

At a round-start point define

$$G_t=\nabla f(W^t),\qquad d_i^t=\nabla f_i(W^t)-G_t,\qquad A_t=\frac{1}{N}\sum_{i=1}^{N}\Pi_i^t,$$

$$b_t=\frac{1}{N}\sum_{i=1}^{N}(\Pi_i^t-A_t)d_i^t,\qquad q_t=\frac{1}{N}\sum_{i=1}^{N}\Pi_i^t\nabla f_i(W^t),\qquad v_t=\left\|A_t-A_t^2\right\|_2. \tag{11}$$

The scalar v_t measures projection dispersion. It is independent of the choice of orthonormal bases for the subspaces. The Frobenius dispersion used in the motivating experiment satisfies

$$\frac{1}{N}\sum_i\left\|\Pi_i^t-A_t\right\|_F^2=\operatorname{tr}(A_t-A_t^2).$$

This trace and $v_t=\|A_t-A_t^2\|_2$ are different quantities; $v_t\le\operatorname{tr}(A_t-A_t^2)\le mv_t$.

Assumption 7 (Aggregate global-gradient coverage)**.** *There exists $\mu\in(0,1]$ such that almost surely, for every round,*

$$\langle G_t,A_tG_t\rangle_F\ge\mu\|G_t\|_F^2. \tag{12}$$

Condition equation 12 concerns only the current global gradient. It takes the respective forms

$$\left\|\Pi^tG_t\right\|_F^2\ge\mu\|G_t\|_F^2,\qquad\frac{1}{N}\sum_i\left\|\Pi_i^tG_t\right\|_F^2\ge\mu\|G_t\|_F^2$$

for shared and client-specific projections. Neither statement follows from coverage of each client's own gradient. The same form of coverage is used throughout, and comparisons below use a common valid value of μ.

The order of conditioning matters. A projector constructed from current oracle noise need not preserve its zero mean, and a projector constructed from the current participating subset may be correlated with that subset. The results below apply when projectors are chosen before sampling and the fresh oracle draws, as specified above. Reusing earlier rounds is permitted. Approximate coverage is treated in Proposition 7.

Projection disagreement and aggregation bias

Theorem 3 (Bias identity, sharp worst-case bound, and cancellation)**.** *Fix any orthogonal projectors $\Pi_1,\dots,\Pi_N$ and set $A=N^{-1}\sum_i\Pi_i$. For matrices d_i with $\sum_i d_i=0$, define $b=N^{-1}\sum_i(\Pi_i-A)d_i$. Then:*

1. *For any G and $G_i=G+d_i$,*

$$\frac{1}{N}\sum_{i=1}^{N}\Pi_iG_i=AG+b. \tag{13}$$

2. *The following operator identity and bounds hold:*

$$\frac{1}{N}\sum_{i=1}^{N}(\Pi_i-A)^2=A-A^2,\qquad 0\preceq A-A^2\preceq\tfrac{1}{4}I. \tag{14}$$

Moreover, for every $z\ge0$,

$$\sup_{\substack{\sum_i d_i=0\\ N^{-1}\sum_i\|d_i\|_F^2\le z^2}}\left\|\frac{1}{N}\sum_{i=1}^{N}(\Pi_i-A)d_i\right\|_F^2=z^2\left\|A-A^2\right\|_2. \tag{15}$$

3. *Bias vanishes for every zero-mean collection (d_i) if and only if all projectors coincide. Equivalently,*

$$[b = 0 \textit{ for every } (d_i)] \iff A - A^2 = 0 \iff \Pi_1 = \cdots = \Pi_N. \tag{16}$$

Consequently, under Assumption 6, $\|b_t\|_F^2 \le v_t \zeta^2$, and shared projections give $b_t = 0$ exactly.

Proof. Since $\frac{1}{N}\sum_{i=1}^N d_i = 0$,

$$\frac{1}{N}\sum_{i=1}^N \Pi_i(G + d_i) = AG + \frac{1}{N}\sum_{i=1}^N \Pi_i d_i = AG + \frac{1}{N}\sum_{i=1}^N (\Pi_i - A)d_i.$$

All matrices $\Pi_i - A$ are symmetric. Expanding their squares gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N (\Pi_i - A)^2 &= \frac{1}{N}\sum_{i=1}^N \Pi_i^2 - \left(\frac{1}{N}\sum_{i=1}^N \Pi_i\right) A - A\left(\frac{1}{N}\sum_{i=1}^N \Pi_i\right) + A^2 \\ &= A - A^2.\end{aligned}$$

An equivalent pairwise form is

$$A - A^2 = \frac{1}{2N^2}\sum_{i,j=1}^N (\Pi_i - \Pi_j)^2. \tag{17}$$

Indeed, the sum of the two squared terms is $2N^2 A$, while the sum of the two ordered cross terms is $2N^2 A^2$. Taking traces gives

$$\operatorname{tr}(A - A^2) = \frac{1}{2N^2}\sum_{i,j=1}^N \|\Pi_i - \Pi_j\|_F^2 .$$

Thus the dispersion has an intrinsic subspace interpretation, independent of basis coordinates. Since $0 \preceq A \preceq I$, its eigenvalues belong to $[0, 1]$. The eigenvalues of $A - A^2$ are $a(1-a) \in [0, 1/4]$. For any test matrix Z with $\|Z\|_F = 1$, symmetry and Cauchy–Schwarz yield

$$\begin{aligned}|\langle Z, b\rangle_F| &= \left|\frac{1}{N}\sum_{i=1}^N \langle (\Pi_i - A)Z, d_i\rangle_F\right| \\ &\le \left(\frac{1}{N}\sum_{i=1}^N \|(\Pi_i - A)Z\|_F^2\right)^{1/2} \left(\frac{1}{N}\sum_{i=1}^N \|d_i\|_F^2\right)^{1/2} \\ &= \left\langle Z, (A - A^2)Z\right\rangle_F^{1/2} \left(\frac{1}{N}\sum_{i=1}^N \|d_i\|_F^2\right)^{1/2} \\ &\le \sqrt{\|A - A^2\|_2}\, z.\end{aligned}$$

Taking the supremum over Z proves the upper bound in equation 15. Let $v = \left\|A - A^2\right\|_2 > 0$, let $u \in \mathbb{R}^m$ be a unit eigenvector with eigenvalue v, and let $a \in \mathbb{R}^n$ be a unit vector. Set $Z = ua^\top$ and

$$d_i = \frac{z}{\sqrt{v}}(\Pi_i - A)Z.$$

Then $\frac{1}{N}\sum_{i=1}^N d_i = 0$ and

$$\frac{1}{N}\sum_{i=1}^N \|d_i\|_F^2 = \frac{z^2}{v}\left\langle Z, (A - A^2)Z\right\rangle_F = z^2.$$

Furthermore,

$$b = \frac{z}{\sqrt{v}}(A - A^2)Z = z\sqrt{v}\, Z,$$

which attains $\|b\|_F^2 = z^2 v$. The cases $v = 0$ or $z = 0$ are immediate. If $A - A^2 = 0$, then for every vector u,

$$0 = u^\top (A - A^2)u = \frac{1}{N} \sum_{i=1}^{N} \|(\Pi_i - A)u\|_2^2.$$

Every summand is nonnegative, so each $\Pi_i - A$ is zero. Conversely, identical projectors give $A^2 = A$ and $b = 0$. If some projectors differ, then $v > 0$, and the maximizing construction above gives a nonzero bias for every $z > 0$. □

Remark 1 (Worst-case interpretation)**.** *The supremum in equation 15 fixes the projectors and varies the zero-mean gradient perturbations. Every admissible perturbation is realizable by smooth quadratic objectives, for example $f_i(W) = \frac{1}{2}\|W\|_F^2 + \langle G + d_i, W\rangle_F$ at $W = 0$. The identity does not assert attainability under a prescribed rule that recomputes the projectors as the gradients vary. An example with exact local SVD selection is given in Theorem 6.*

The norm bound does not retain the orientation of the bias. The next result characterizes its effect on the descent direction without this relaxation.

Proposition 1 (Sharp directional alignment)**.** *Fix the projectors and $G \in \mathbb{R}^{m\times n}$. Let $V = A - A^2$ and $s_G = \sqrt{\langle G, VG\rangle_F}$. For every $z \geq 0$,*

$$\inf_{\substack{\sum_i d_i = 0 \\ N^{-1}\sum_i \|d_i\|_F^2 \leq z^2}} \left\langle G, \frac{1}{N}\sum_i \Pi_i (G + d_i) \right\rangle_F = \langle G, AG\rangle_F - z s_G. \tag{18}$$

Consequently, for $G \neq 0$, the negative aggregate is a strict descent direction for every admissible perturbation if and only if

$$\langle G, AG\rangle_F > z\sqrt{\langle G, (A - A^2)G\rangle_F}. \tag{19}$$

For a shared projector the right-hand side is zero.

Proof. The bias identity and Cauchy–Schwarz imply

$$\begin{aligned}
\langle G, AG + b\rangle_F &= \langle G, AG\rangle_F + \frac{1}{N}\sum_i \langle (\Pi_i - A)G, d_i\rangle_F \\
&\geq \langle G, AG\rangle_F - \left(\frac{1}{N}\sum_i \|(\Pi_i - A)G\|_F^2\right)^{1/2} \left(\frac{1}{N}\sum_i \|d_i\|_F^2\right)^{1/2} \\
&\geq \langle G, AG\rangle_F - z s_G.
\end{aligned}$$

If $s_G > 0$, equality holds for $d_i = -z(\Pi_i - A)G/s_G$: their mean is zero and their mean squared norm is z^2. If $s_G = 0$, every $(\Pi_i - A)G$ is zero, so the directional bias vanishes for all admissible (d_i). This proves the equality. The directional derivative of f along $-q$ is $-\langle G, q\rangle_F$. Since the infimum is attained, uniform strict descent is equivalent to equation 19. □

Convergence under partial participation

Write

$$h = \gamma K, \qquad x = Lh, \qquad \omega_S = \frac{N - S}{S(N - 1)}, \tag{20}$$

and define

$$a_K = \frac{5}{6}L^2\gamma^2(K - 1)(2K - 1), \qquad e_K = \frac{3}{2}L^2\gamma^2(K - 1). \tag{21}$$

Here $a_K \leq 5x^2/3$, $e_K \leq 3x^2/(2K)$, and both vanish when $K = 1$. The two coefficients distinguish drift from deterministic gradients and from accumulated martingale noise.

Theorem 4 (Convergence with projection bias)**.** *Suppose the model in Section and Assumptions 5–7 hold, and*

$$0 < x = L\gamma K \leq \frac{\mu}{16}. \tag{22}$$

For every $T \geq 1$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|G_t\right\|_F^2 \leq \frac{4\Delta_0}{\mu h T} + \frac{8}{\mu^2 T}\sum_{t=0}^{T-1}\mathbb{E}\left\|b_t\right\|_F^2 + \frac{8}{\mu^2}(a_K\zeta^2 + e_K\sigma^2) + \frac{4x}{\mu}\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right). \tag{23}$$

The projection-bias term is at most

$$\frac{8\zeta^2}{\mu^2}\bar{v}_T, \qquad \bar{v}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|A_t - A_t^2\right\|_2, \tag{24}$$

and is zero for shared projections.

The term $\omega_S\zeta^2$ accounts for client sampling and vanishes under full participation. The variance $\sigma^2/(SK)$ is the accumulated oracle noise in the server direction. Its indirect contribution through local drift is controlled by $e_K\sigma^2 = O(x^2\sigma^2/K)$, whereas deterministic heterogeneity contributes $a_K\zeta^2 = O(x^2\zeta^2)$. No independence between the local drift and the accumulated noise is used.

Auxiliary estimates

Lemma 1 (Finite-population sampling)**.** *For any $\mathcal{F}_t$-measurable matrices $Q_1, \ldots, Q_N$, let $q = N^{-1}\sum_i Q_i$ and $q_{\mathcal{S}} = S^{-1}\sum_{i\in\mathcal{S}_t} Q_i$. Then*

$$\mathbb{E}_t q_{\mathcal{S}} = q, \qquad \mathbb{E}_t\left\|q_{\mathcal{S}} - q\right\|_F^2 = \omega_S\frac{1}{N}\sum_{i=1}^{N}\left\|Q_i - q\right\|_F^2. \tag{25}$$

For $Q_i = \Pi_i^t\nabla f_i(W^t)$, this implies

$$\mathbb{E}_t\left\|q_{\mathcal{S}}\right\|_F^2 \leq 3\left\|G_t\right\|_F^2 + 2\left\|b_t\right\|_F^2 + \omega_S\zeta^2. \tag{26}$$

Proof. Let $I_i = \mathbf{1}\{i \in \mathcal{S}_t\}$ and $X_i = Q_i - q$. Uniform sampling gives $\mathbb{E}_t I_i = S/N$ and, for $i \neq j$, $\mathbb{E}_t I_i I_j = S(S-1)/(N(N-1))$. Since $\sum_i X_i = 0$, $\sum_{i\neq j}\langle X_i, X_j\rangle_F = -\sum_i\left\|X_i\right\|_F^2$. Expanding the square of $S^{-1}\sum_i I_i X_i$ gives

$$\frac{1}{S^2}\left(\frac{S}{N} - \frac{S(S-1)}{N(N-1)}\right)\sum_i\left\|X_i\right\|_F^2 = \omega_S\frac{1}{N}\sum_{i=1}^{N}\left\|X_i\right\|_F^2.$$

For the stated application, $q = A_tG_t + b_t$, and hence $\left\|q\right\|_F^2 \leq 2\left\|G_t\right\|_F^2 + 2\left\|b_t\right\|_F^2$ because $\left\|A_t\right\|_2 \leq 1$. Also,

$$\frac{1}{N}\sum_{i=1}^{N}\left\|Q_i - q\right\|_F^2 \leq \frac{1}{N}\sum_{i=1}^{N}\left\|Q_i\right\|_F^2 \leq \frac{1}{N}\sum_{i=1}^{N}\left\|G_t + d_i^t\right\|_F^2 = \left\|G_t\right\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|d_i^t\right\|_F^2 \leq \left\|G_t\right\|_F^2 + \zeta^2.$$

Combining these bounds and $\omega_S \leq 1$ gives equation 26. □

Define the average direction, accumulated noise, and drift error by

$$U_t = \frac{1}{SK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\Pi_i^t g_i^{t,k},$$

$$B_t = \frac{1}{SK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\Pi_i^t \varepsilon_i^{t,k},$$

$$C_t = \frac{1}{SK}\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\Pi_i^t[\nabla f_i(W_i^{t,k}) - \nabla f_i(W^t)]. \tag{27}$$

Then the exact identities are

$$U_t = q_{\mathcal{S}_t} + B_t + C_t, \qquad W^{t+1} = W^t - hU_t. \tag{28}$$

Neither C_t nor $B_t + C_t$ is assumed to be independent of the local trajectory. In particular, C_t need not have zero mean.

Lemma 2 (Accumulated oracle noise)**.** *Under equation 10 and the sampling convention,*

$$\mathbb{E}[B_t \mid \mathcal{F}_t, \mathcal{S}_t] = 0, \qquad \mathbb{E}_t \|B_t\|_F^2 \le \frac{\sigma^2}{SK}. \tag{29}$$

Proof. Condition first on $\mathcal{F}_t$ and $\mathcal{S}_t$. At equal local steps, cross-client noise products have zero conditional expectation by the oracle assumption. At unequal steps $k < \ell$, $\Pi_i^t \varepsilon_i^{t,k}$ is measurable before step ℓ, whereas $\Pi_j^t \varepsilon_j^{t,\ell}$ has zero conditional expectation. The tower property therefore eliminates every cross term in the squared norm of the sum. Since projectors contract norms, the SK diagonal terms are each at most σ^2 in expectation. Division by S^2K^2 proves the variance bound. The mean-zero assertion follows in the same way. □

Lemma 3 (Local drift with separate noise dependence)**.** *Suppose $x \le 1/8$ and define*

$$D_{t,k} = \frac{1}{N} \sum_i \mathbb{E}_t \left\| W_i^{t,k} - W^t \right\|_F^2, \qquad H_t = \|G_t\|_F^2 + \zeta^2.$$

Then, for $0 \le k \le K$,

$$D_{t,k} \le 5\gamma^2 k^2 H_t + 3\gamma^2 k \sigma^2, \tag{30}$$

$$\mathbb{E}_t \|C_t\|_F^2 \le a_K H_t + e_K \sigma^2. \tag{31}$$

Proof. Separate the gradient and noise sums in the unrolled update:

$$W_i^{t,k} - W^t = -\gamma \sum_{j<k} \Pi_i^t \nabla f_i(W_i^{t,j}) - \gamma \sum_{j<k} \Pi_i^t \varepsilon_i^{t,j}.$$

Although these two sums are generally correlated, the inequality $\|X+Y\|_F^2 \le 2\|X\|_F^2 + 2\|Y\|_F^2$ applies without independence. Within the second sum, the martingale-difference property eliminates all cross terms, so

$$\mathbb{E}_t \left\| \sum_{j<k} \Pi_i^t \varepsilon_i^{t,j} \right\|_F^2 \le k\sigma^2.$$

Cauchy–Schwarz on the first sum therefore gives

$$D_{t,k} \le 2\gamma^2 k \sum_{j<k} \frac{1}{N} \sum_i \mathbb{E}_t \left\| \nabla f_i(W_i^{t,j}) \right\|_F^2 + 2\gamma^2 k \sigma^2. \tag{32}$$

By smoothness and the zero-mean identity for d_i^t,

$$\begin{aligned} \frac{1}{N} \sum_i \mathbb{E}_t \left\| \nabla f_i(W_i^{t,j}) \right\|_F^2 &\le 2L^2 D_{t,j} + 2\frac{1}{N} \sum_i \left\| G_t + d_i^t \right\|_F^2 \\ &= 2L^2 D_{t,j} + 2\|G_t\|_F^2 + 2\frac{1}{N} \sum_i \left\| d_i^t \right\|_F^2 \\ &\le 2L^2 D_{t,j} + 2H_t. \end{aligned}$$

Substitution into equation 32 yields

$$D_{t,k} \le 4\gamma^2 L^2 k \sum_{j<k} D_{t,j} + 4\gamma^2 k^2 H_t + 2\gamma^2 k \sigma^2. \tag{33}$$

Fix k and put $M_{t,k} = \max_{0 \le j \le k} D_{t,j}$. For each $\ell \le k$, apply equation 33 at ℓ and bound $\sum_{j<\ell} D_{t,j} \le \ell M_{t,k}$. Taking the maximum gives

$$M_{t,k} \le 4\gamma^2 L^2 k^2 M_{t,k} + 4\gamma^2 k^2 H_t + 2\gamma^2 k \sigma^2.$$

Since $4\gamma^2L^2k^2 \le 4x^2 \le 1/16$,

$$D_{t,k} \le M_{t,k} \le \frac{4\gamma^2k^2H_t + 2\gamma^2k\sigma^2}{1 - 4\gamma^2L^2k^2} \le 5\gamma^2k^2H_t + 3\gamma^2k\sigma^2.$$

For C_t, Jensen's inequality, contraction, and smoothness imply

$$\begin{aligned}\mathbb{E}_t \left\|C_t\right\|_F^2 &\le \frac{L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}_t\left[\frac{1}{S}\sum_{i\in\mathcal{S}_t}\left\|W_i^{t,k} - W^t\right\|_F^2\right] \\ &= \frac{L^2}{K}\sum_{k=0}^{K-1} D_{t,k}.\end{aligned}$$

For the equality, conditional independence of the subset and virtual trajectories gives

$$\mathbb{E}_t\left[\mathbf{1}\{i\in\mathcal{S}_t\}\left\|W_i^{t,k} - W^t\right\|_F^2\right] = \frac{S}{N}\mathbb{E}_t\left\|W_i^{t,k} - W^t\right\|_F^2.$$

Using $\sum_{k<K} k^2 = K(K-1)(2K-1)/6$ and $\sum_{k<K} k = K(K-1)/2$ proves equation 31. □

PROOF OF THE CONVERGENCE THEOREM

Proof of Theorem 4. All expectations in the first part are conditional on $\mathcal{F}_t$. The update $W^{t+1} = W^t - hU_t$ and smoothness give

$$\mathbb{E}_t f(W^{t+1}) \le f(W^t) - h\left\langle G_t, \mathbb{E}_t U_t\right\rangle_F + \frac{Lh^2}{2}\mathbb{E}_t\left\|U_t\right\|_F^2. \tag{34}$$

By the sampling and noise lemmas, $\mathbb{E}_t U_t = A_tG_t + b_t + \mathbb{E}_t C_t$. Coverage and two applications of $ab \le \mu a^2/4 + b^2/\mu$ yield

$$\begin{aligned}-\left\langle G_t, \mathbb{E}_t U_t\right\rangle_F &\le -\mu\left\|G_t\right\|_F^2 + \left\|G_t\right\|_F\left\|b_t\right\|_F + \left\|G_t\right\|_F\left\|\mathbb{E}_t C_t\right\|_F \\ &\le -\frac{\mu}{2}\left\|G_t\right\|_F^2 + \frac{1}{\mu}\left\|b_t\right\|_F^2 + \frac{1}{\mu}\mathbb{E}_t\left\|C_t\right\|_F^2.\end{aligned} \tag{35}$$

The noise makes no contribution to this first-order term.

For the second moment, $q_{\mathcal{S}_t}$ is measurable with respect to $\sigma(\mathcal{F}_t, \mathcal{S}_t)$ and $\mathbb{E}[B_t \mid \mathcal{F}_t, \mathcal{S}_t] = 0$. Hence

$$\mathbb{E}_t\left\langle q_{\mathcal{S}_t}, B_t\right\rangle_F = 0, \qquad \mathbb{E}_t\left\|q_{\mathcal{S}_t} + B_t\right\|_F^2 = \mathbb{E}_t\left\|q_{\mathcal{S}_t}\right\|_F^2 + \mathbb{E}_t\left\|B_t\right\|_F^2.$$

Consequently, the two-term norm inequality gives

$$\begin{aligned}\mathbb{E}_t\left\|U_t\right\|_F^2 &\le 2\mathbb{E}_t\left\|q_{\mathcal{S}_t} + B_t\right\|_F^2 + 2\mathbb{E}_t\left\|C_t\right\|_F^2 \\ &\le 6\left\|G_t\right\|_F^2 + 4\left\|b_t\right\|_F^2 + 2\omega_S\zeta^2 + \frac{2\sigma^2}{SK} + 2\mathbb{E}_t\left\|C_t\right\|_F^2.\end{aligned} \tag{36}$$

Substituting equation 35 and equation 36 into equation 34, with $x = Lh$, gives

$$\begin{aligned}\mathbb{E}_t f(W^{t+1}) \le f(W^t) &- h\left(\frac{\mu}{2} - 3x\right)\left\|G_t\right\|_F^2 \\ &+ h\left(\frac{1}{\mu} + 2x\right)\left\|b_t\right\|_F^2 + h\left(\frac{1}{\mu} + x\right)\mathbb{E}_t\left\|C_t\right\|_F^2 \\ &+ hx\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right).\end{aligned} \tag{37}$$

Since $x \le \mu/16 \le 1/16$, Lemma 3 applies. Also,

$$\frac{1}{\mu} + 2x \le \frac{2}{\mu}, \qquad \frac{1}{\mu} + x \le \frac{2}{\mu},$$

and the coefficient to be absorbed satisfies

$$3x + \left(\frac{1}{\mu} + x\right)a_K \le 3x + \frac{10x^2}{3\mu} \le \left(\frac{3}{16} + \frac{5}{384}\right)\mu = \frac{77}{384}\mu < \frac{\mu}{4}.$$

Thus equation 37 implies

$$\mathbb{E}_t f(W^{t+1}) \le f(W^t) - \frac{\mu h}{4}\left\|G_t\right\|_F^2 + \frac{2h}{\mu}\left\|b_t\right\|_F^2 + \frac{2h}{\mu}(a_K\zeta^2 + e_K\sigma^2) + hx\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right). \tag{38}$$

Taking total expectations and summing over $t < T$ telescopes the objective:

$$\frac{\mu h}{4}\sum_{t<T}\mathbb{E}\left\|G_t\right\|_F^2 \le f(W^0) - \mathbb{E}f(W^T) + \frac{2h}{\mu}\sum_{t<T}\mathbb{E}\left\|b_t\right\|_F^2 + \frac{2hT}{\mu}(a_K\zeta^2 + e_K\sigma^2) + hxT\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right).$$

Use $f(W^T) \ge f^\star$ and divide by $\mu hT/4$. This proves equation 23; Theorem 3 supplies equation 24. □

A SHARPER BOUND FOR SHARED PROJECTIONS

For shared projections, the entire server direction belongs to the common subspace. Retaining this fact gives a stronger result than setting $b_t = 0$ in Theorem 4. In particular, projected stationarity does not require a lower bound on global-gradient coverage.

Theorem 5 (Shared-projection stationarity with an explicit residual)**.** *Assume the sampling and oracle conditions in Section and Assumptions 5 and 6. All clients use the same $\mathcal{F}_t$-measurable orthogonal projector Π^t throughout round t. Allow deterministic round-dependent stepsizes $\gamma_t > 0$, constant within each round, and define*

$$h_t = K\gamma_t, \quad x_t = Lh_t \le \tfrac{1}{8}, \quad \Lambda_T = \sum_{t<T} h_t,$$
$$a_{K,t} = \tfrac{5}{6}L^2\gamma_t^2(K-1)(2K-1), \quad e_{K,t} = \tfrac{3}{2}L^2\gamma_t^2(K-1),$$
$$\rho_t^2 = \left\|(I-\Pi^t)G_t\right\|_F^2, \quad V_S = \omega_S\zeta^2 + \frac{\sigma^2}{SK}. \tag{39}$$

Set

$$\mathcal{B}_T = \frac{2\Delta_0}{\Lambda_T} + \frac{4}{\Lambda_T}\sum_{t<T} h_t(a_{K,t}\zeta^2 + e_{K,t}\sigma^2) + \frac{2V_S}{\Lambda_T}\sum_{t<T} h_t x_t. \tag{40}$$

Without Assumption 7,

$$\frac{1}{\Lambda_T}\sum_{t<T} h_t\mathbb{E}\left\|\Pi^t G_t\right\|_F^2 \le \mathcal{B}_T, \tag{41}$$

$$\frac{1}{\Lambda_T}\sum_{t<T} h_t\mathbb{E}\left\|G_t\right\|_F^2 \le \mathcal{B}_T + \frac{1}{\Lambda_T}\sum_{t<T} h_t\mathbb{E}\rho_t^2. \tag{42}$$

If, additionally, for a constant $\mu \in (0,1]$ and every t,

$$\mathbb{E}\left\|\Pi^t G_t\right\|_F^2 \ge \mu\mathbb{E}\left\|G_t\right\|_F^2, \tag{43}$$

then the left-hand side of equation 42 is at most $\mathcal{B}_T/\mu$. Thus expected coverage suffices; pathwise coverage in Assumption 7 is a sufficient, stronger condition.

Proof. Fix t, condition on $\mathcal{F}_t$, and write $\Pi = \Pi^t$, $Q_t = \Pi G_t$, $\gamma = \gamma_t$, $h = h_t$, and $x = x_t$. The round-start sampled direction has conditional mean Q_t. Since $\sum_i d_i^t = 0$ and Π is shared, the finite-population identity gives

$$\mathbb{E}_t\left\|q_{\mathcal{S}_t}\right\|_F^2 = \left\|Q_t\right\|_F^2 + \omega_S\frac{1}{N}\sum_i\left\|\Pi d_i^t\right\|_F^2 \le \left\|Q_t\right\|_F^2 + \omega_S\zeta^2. \tag{44}$$

The projected version of the local-drift estimate is

$$\mathbb{E}_t \|C_t\|_F^2 \le a_{K,t}(\|Q_t\|_F^2 + \zeta^2) + e_{K,t}\sigma^2. \tag{45}$$

For completeness, let $D_{t,k} = N^{-1}\sum_i \mathbb{E}_t \left\|W_i^{t,k} - W^t\right\|_F^2$. The unrolled recursion and martingale orthogonality yield

$$D_{t,k} \le 2\gamma^2 k \sum_{j<k} \frac{1}{N}\sum_i \mathbb{E}_t \left\|\Pi\nabla f_i(W_i^{t,j})\right\|_F^2 + 2\gamma^2 k\sigma^2.$$

Smoothness and the identity

$$\frac{1}{N}\sum_i \left\|\Pi\nabla f_i(W^t)\right\|_F^2 = \|Q_t\|_F^2 + \frac{1}{N}\sum_i \left\|\Pi d_i^t\right\|_F^2 \le \|Q_t\|_F^2 + \zeta^2$$

give

$$D_{t,k} \le 4\gamma^2 L^2 k \sum_{j<k} D_{t,j} + 4\gamma^2 k^2(\|Q_t\|_F^2 + \zeta^2) + 2\gamma^2 k\sigma^2.$$

Because $L\gamma K \le 1/8$, the maximum argument in Lemma 3 proves $D_{t,k} \le 5\gamma^2 k^2(\|Q_t\|_F^2 + \zeta^2) + 3\gamma^2 k\sigma^2$. Averaging over the uniformly sampled clients and local steps, as in that lemma, gives equation 45 with the stated $a_{K,t}, e_{K,t}$.

Now $\Pi U_t = U_t$ and $\Pi C_t = C_t$ pathwise, and $\mathbb{E}_t U_t = Q_t + \mathbb{E}_t C_t$. Hence

$$\begin{aligned} -\langle G_t, \mathbb{E}_t U_t\rangle_F &= -\|Q_t\|_F^2 - \langle Q_t, \mathbb{E}_t C_t\rangle_F \\ &\le -\tfrac{3}{4}\|Q_t\|_F^2 + \mathbb{E}_t\|C_t\|_F^2. \end{aligned} \tag{46}$$

Here we used $ab \le a^2/4 + b^2$ and conditional Jensen. Lemma 2, the orthogonality of q_{S_t} and B_t in conditional expectation, and equation 44 imply

$$\mathbb{E}_t\|U_t\|_F^2 \le 2\|Q_t\|_F^2 + 2V_S + 2\mathbb{E}_t\|C_t\|_F^2.$$

Applying smoothness to $W^{t+1} = W^t - hU_t$ therefore gives

$$\begin{aligned} \mathbb{E}_t f(W^{t+1}) &\le f(W^t) - h(3/4 - x)\|Q_t\|_F^2 + h(1+x)\mathbb{E}_t\|C_t\|_F^2 + hxV_S \\ &\le f(W^t) - h[3/4 - x - (1+x)a_{K,t}]\|Q_t\|_F^2 \\ &\quad + h(1+x)(a_{K,t}\zeta^2 + e_{K,t}\sigma^2) + hxV_S. \end{aligned}$$

Since $a_{K,t} \le 5x^2/3$ and $x \le 1/8$,

$$x + (1+x)a_{K,t} \le \tfrac{1}{8} + \tfrac{9}{8}\cdot\tfrac{5}{192} = \tfrac{79}{512} < \tfrac{1}{4}, \qquad 1 + x \le 2.$$

Consequently,

$$\mathbb{E}_t f(W^{t+1}) \le f(W^t) - \frac{h_t}{2}\left\|\Pi^t G_t\right\|_F^2 + 2h_t(a_{K,t}\zeta^2 + e_{K,t}\sigma^2) + h_t x_t V_S. \tag{47}$$

Take total expectations, sum over $t < T$, use $f(W^T) \ge f^\star$, and divide by $\Lambda_T/2$ to prove equation 41. The orthogonal decomposition $\|G_t\|_F^2 = \|\Pi^t G_t\|_F^2 + \rho_t^2$ proves equation 42. Finally, summing equation 43 with the deterministic weights h_t gives the last assertion. □

Meaning of the residual. The first bound controls stationarity within the currently used subspace. The second exposes the missing component of the full gradient exactly. Neither an arbitrary shared projector nor subspace refresh alone forces ρ_t to vanish. The theorem applies to the actual predictable shared selection rule even when a uniform positive coverage constant is unavailable; in that case it is a residual bound, not an unconditional full-stationarity guarantee. For nonconstant schedules, the conclusion is weighted by h_t; it does not assert the same unweighted rate for every cosine schedule.

Corollary 1 (Shared projections under additive expected coverage)**.** *Under Theorem 5, suppose*

$$\mathbb{E}\left\|\Pi^t G_t\right\|_F^2 \ge \mu\mathbb{E}\|G_t\|_F^2 - \chi_t$$

for deterministic $\chi_t \ge 0$ and $\mu \in (0, 1]$. Then

$$\frac{1}{\Lambda_T}\sum_{t<T} h_t\mathbb{E}\|G_t\|_F^2 \le \frac{\mathcal{B}_T}{\mu} + \frac{1}{\mu\Lambda_T}\sum_{t<T} h_t\chi_t.$$

Proof. Sum the assumed inequality with weights h_t and apply equation 41. □

CONSEQUENCES FOR SHARED AND CLIENT-SPECIFIC PROJECTIONS

Define the common remainder

$$\mathcal{R}_T = \frac{4\Delta_0}{\mu h T} + \frac{8}{\mu^2}(a_K\zeta^2 + e_K\sigma^2) + \frac{4x}{\mu}\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right). \tag{48}$$

Corollary 2 (Sharper shared-projection convergence)**.** *Under Theorem 5 and expected coverage equation 43, fix a horizon T and set $\gamma_t = c/(LK\sqrt{T})$ for all $t < T$, where $0 < c \le 1/8$. Then*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\left\|G_t\right\|_F^2 \le \frac{2L\Delta_0}{\mu c\sqrt{T}} + \frac{20c^2\zeta^2}{3\mu T} + \frac{6c^2\sigma^2}{\mu KT} + \frac{2c}{\mu\sqrt{T}}\left(\omega_S\zeta^2 + \frac{\sigma^2}{SK}\right). \tag{49}$$

For fixed problem parameters, K, S, c, μ, this is $O(T^{-1/2})$. For $K = 1$, both drift terms vanish in the unrelaxed bound $\mathcal{B}_T/\mu$; for $S = N$, the client-sampling term vanishes. Without coverage, the same right-hand side with $\mu = 1$ bounds the average projected-gradient norm, and adding $T^{-1}\sum_{t<T}\mathbb{E}\rho_t^2$ bounds the average full-gradient norm.

Proof. Substitute $h_t = c/(L\sqrt{T})$, $x_t = c/\sqrt{T}$, $a_{K,t} \le 5c^2/(3T)$, and $e_{K,t} \le 3c^2/(2KT)$ into equation 40, and use Theorem 5. □

The admissible shared stepsize is independent of μ; the resulting coverage factor enters as $1/\mu$. This improvement uses the common subspace of the entire server direction. It does not extend automatically to client-specific projections.

Corollary 3 (Comparison under common constants)**.** *Consider shared and client-specific runs on the same objective and initialization, with the same γ, K, S and common valid constants L, σ, ζ, μ. Their bounds in Theorem 4 are respectively*

$$\mathcal{R}_T \quad \textit{and} \quad \mathcal{R}_T + \frac{8}{\mu^2 T}\sum_{t<T}\mathbb{E}\left\|b_t^{\mathrm{local}}\right\|_F^2.$$

The second bound is strictly larger if the total bias energy is positive.

Proof. Apply equation 23 to each run and use $b_t^{\mathrm{shared}} = 0$. □

The comparison concerns upper bounds evaluated along different trajectories. It is not an ordering of the actual optimization errors. In particular, client-specific projections can converge when their bias is sufficiently small, even if they are not identical.

Corollary 4 (Client-specific projections with controlled bias)**.** *Under Theorem 4, suppose for some $\theta \in [0, 1)$,*

$$\mathbb{E}\left\|b_t\right\|_F^2 \le \frac{\theta\mu^2}{8}\mathbb{E}\left\|G_t\right\|_F^2 \quad \textit{for every } t. \tag{50}$$

Then $T^{-1}\sum_{t<T}\mathbb{E}\left\|G_t\right\|_F^2 \le \mathcal{R}_T/(1-\theta)$. More generally, with $\gamma = c/(LK\sqrt{T})$ and $0 < c \le \mu/16$, $T^{-1}\sum_{t<T}\mathbb{E}\left\|b_t\right\|_F^2 = O(T^{-1/2})$ is sufficient for the same $O(T^{-1/2})$ stationarity rate.

Proof. Substitute equation 50 into equation 23: its bias term is at most $\theta T^{-1}\sum_{t<T}\mathbb{E}\left\|G_t\right\|_F^2$. Move it to the left. The second statement follows directly from the same bound. □

For example, a pathwise condition $\left\|A_t - A_t^2\right\|_2 N^{-1}\sum_i \left\|d_i^t\right\|_F^2 \le \theta\mu^2\left\|G_t\right\|_F^2/8$ implies equation 50 by Theorem 3. Thus disagreement alone does not force a residual error; its coupling to heterogeneity determines the obstruction.

A STRICT SEPARATION WITH EXACT LOCAL SVD PROJECTIONS

Theorem 6 (Local optimal subspaces can stall away from stationarity)**.** *There exist four* 1*-smooth, strongly convex quadratic client objectives on* $\mathbb{R}^{2\times 2}$ *with bounded gradient heterogeneity for which the following holds under full participation, exact gradients, rank* $r = 1$*, and any fixed number of local steps* $K \geq 1$ *(with each projector fixed within a round). Starting from the same* $W^0 = 0$*:*

1. *Local projected SGD, using each local gradient's exact leading left singular subspace at the start of every round, stays at* $W^t = 0$ *for all* t*. Its squared global-gradient norm is* $1/2$*, and its global objective gap is* $1/4$*. Every local projection captures* $4/5$ *of its own round-start gradient energy. The average local projection also satisfies aggregate global-gradient coverage with* $\mu = 1/2$*.*

2. *Shared projected SGD, using a leading left singular subspace of the round-start global gradient, reaches the global minimizer in two rounds when* $\gamma = 1$*. For any constant* $0 < \gamma \leq 1$*, define* $\nu = 1 - (1-\gamma)^K \in (0, 1]$*. It satisfies*

$$f(W^t) - f^\star \leq \left(1 - \frac{\nu(2-\nu)}{2}\right)^t [f(W^0) - f^\star]. \tag{51}$$

In particular, strict separation also holds for any sufficiently small stepsize satisfying Theorem 4.

Proof. Let

$$C_1 = \mathrm{diag}(2, 1), \quad C_2 = \mathrm{diag}(-2, 1), \quad C_3 = \mathrm{diag}(1, 2), \quad C_4 = \mathrm{diag}(1, -2),$$

and set

$$f_i(W) = \tfrac{1}{2}\|W\|_F^2 + \langle C_i, W\rangle_F .$$

Then $\nabla f_i(W) = W + C_i$ and every Hessian is the identity. Their average is $\bar{C} = \frac{1}{2} I_2$, so

$$f(W) = \tfrac{1}{2}\|W\|_F^2 + \left\langle \tfrac{1}{2} I_2, W\right\rangle_F , \qquad W^\star = -\tfrac{1}{2} I_2,$$

and

$$f(W) - f^\star = \tfrac{1}{2}\|W - W^\star\|_F^2 , \qquad \nabla f(W) = W + \tfrac{1}{2} I_2. \tag{52}$$

The heterogeneity is independent of W:

$$\frac{1}{4}\sum_{i=1}^{4} \|\nabla f_i(W) - \nabla f(W)\|_F^2 = \frac{1}{4}\sum_i \left\|C_i - \bar{C}\right\|_F^2 = 5 - \tfrac{1}{2} = \tfrac{9}{2}.$$

Thus $\zeta^2 = 9/2$ is a global valid bound, and $\sigma = 0$. At $W = 0$, the leading singular values are all 2, strictly larger than the other singular value 1. Hence the rank-one orthogonal projectors are uniquely determined as

$$\Pi_1 = \Pi_2 = \mathrm{diag}(1, 0), \qquad \Pi_3 = \Pi_4 = \mathrm{diag}(0, 1).$$

Their projected gradients are

$$\Pi_1 C_1 = \mathrm{diag}(2, 0), \quad \Pi_2 C_2 = \mathrm{diag}(-2, 0), \quad \Pi_3 C_3 = \mathrm{diag}(0, 2), \quad \Pi_4 C_4 = \mathrm{diag}(0, -2).$$

The average is zero. To see that this persists for any K, unroll the quadratic local dynamics with a fixed projector:

$$W_i^{t,k} = W^t - [1 - (1-\gamma)^k]\Pi_i^t(W^t + C_i).$$

Indeed, if $Y_i^k = \Pi_i^t(W_i^{t,k} + C_i)$, idempotence gives $Y_i^{k+1} = (1-\gamma)Y_i^k$ and $Y_i^0 = \Pi_i^t(W^t + C_i)$. Summing $W_i^{t,k+1} - W_i^{t,k} = -\gamma Y_i^k$ and using $\gamma\sum_{j=0}^{k-1}(1-\gamma)^j = 1 - (1-\gamma)^k$ gives the displayed expression. At $W^t = 0$, the averaged endpoint is therefore $-\nu\,\frac{1}{4}\sum_i \Pi_i C_i = 0$. Thus every server round leaves $W = 0$ unchanged, regardless of the positive local stepsize. By induction, all subsequent local SVDs return the same projectors. Each retained-energy ratio is $4/(4+1) = 4/5$, yet

$$\|\nabla f(0)\|_F^2 = \left\|\tfrac{1}{2} I_2\right\|_F^2 = \tfrac{1}{2}, \qquad f(0) - f^\star = \tfrac{1}{4}.$$

Moreover,

$$A = \tfrac{1}{2} I_2, \quad A - A^2 = \tfrac{1}{4} I_2, \quad AG = \tfrac{1}{4} I_2, \quad b = -\tfrac{1}{4} I_2.$$

Thus $\langle G, AG\rangle_F = \frac{1}{2}\|G\|_F^2$: global-gradient coverage holds even though the bias cancels the covered descent direction exactly. Choose any leading left singular direction of the current global gradient $G = W + \frac{1}{2}I_2$, with projector Π. At the round-start point, linearity and full participation give

$$\frac{1}{4}\sum_i \Pi\nabla f_i(W) = \Pi\nabla f(W) = \Pi G.$$

Since G is a 2×2 matrix, a leading left singular direction retains at least half its squared Frobenius norm: $\|\Pi G\|_F^2 \geq \frac{1}{2}\|G\|_F^2$. The same unrolled local dynamics show that after K steps, $W^+ = W - \nu\Pi G$, where $\nu = 1 - (1-\gamma)^K$. For this quadratic, $G^+ = G - \nu\Pi G$. Orthogonality gives the exact identity

$$\left\|G^+\right\|_F^2 = \|G\|_F^2 - \nu(2-\nu)\|\Pi G\|_F^2 \leq \left(1 - \frac{\nu(2-\nu)}{2}\right)\|G\|_F^2.$$

Using equation 52 proves equation 51. For $\gamma = 1$, $\nu = 1$ for every $K \geq 1$. Choose $\Pi^0 = \mathrm{diag}(1, 0)$ to resolve the initial tie:

$$W^1 = \mathrm{diag}(-1/2, 0), \quad G_1 = \mathrm{diag}(0, 1/2), \quad \Pi^1 = \mathrm{diag}(0, 1), \quad W^2 = -\tfrac{1}{2}I_2 = W^\star.$$

After reaching the minimizer, the shared aggregate remains zero. □

Both runs use the same rank, participation, local stepsize, and number of local steps. Their difference is the choice of the round-wise subspace. The example establishes a failure of exact local SVD selection despite positive local and aggregate coverage. The shared comparator uses the global gradient; constructing an adequate approximation with limited communication is a separate requirement, addressed by the coverage bounds below.

Coverage residual of the implemented projector average

The shared basis in FedLore is obtained from candidate projectors, not from the SVD of an averaged gradient. We analyze this construction directly. Candidate projectors below are denoted by Q_i to distinguish them from the optimization projector Π^t.

Proposition 2 (Residual certificate for projector averaging)**.** *Let $H_i \in \mathbb{R}^{m\times n}$ be candidate gradient matrices and let Q_i be rank-r orthogonal projectors, with $1 \leq r \leq m$. Define*

$$\begin{aligned}
\bar{H} &= \frac{1}{N}\sum_i H_i, \qquad A^{\mathrm{cand}} = \frac{1}{N}\sum_i Q_i, \\
\epsilon_{\mathrm{loc}}^2 &= \frac{1}{N}\sum_i \|(I - Q_i)H_i\|_F^2, \qquad M = \max_i \|H_i\|_2, \\
d_r &= r - \sum_{j=1}^{r} \lambda_j(A^{\mathrm{cand}}),
\end{aligned} \tag{53}$$

where eigenvalues are ordered nonincreasingly. Let $\widehat{\Pi}$ be a rank-r orthogonal projector satisfying

$$\mathrm{tr}(\widehat{\Pi}A^{\mathrm{cand}}) \geq \sum_{j=1}^{r}\lambda_j(A^{\mathrm{cand}}) - \varepsilon_{\mathrm{eig}}, \qquad \varepsilon_{\mathrm{eig}} \geq 0. \tag{54}$$

For any target gradient G and $\delta = \left\|\bar{H} - G\right\|_F$,

$$\left\|(I - \widehat{\Pi})G\right\|_F \leq \epsilon_{\mathrm{loc}} + M\sqrt{d_r + \varepsilon_{\mathrm{eig}}} + \delta, \tag{55}$$

$$\left\|\widehat{\Pi}G\right\|_F^2 \geq \|G\|_F^2 - (\epsilon_{\mathrm{loc}} + M\sqrt{d_r + \varepsilon_{\mathrm{eig}}} + \delta)^2. \tag{56}$$

For exact candidate SVDs, $\epsilon_{\mathrm{loc}}^2$ is the mean discarded squared singular-value energy. For an exact leading eigenspace of A^{cand}, one can take $\varepsilon_{\mathrm{eig}} = 0$.

Proof. Since every Q_i has rank r, $\mathrm{tr}(A^{\mathrm{cand}}) = r$. For every rank-$r$ orthogonal projector Π,

$$\frac{1}{N}\sum_i \|(I-\Pi)Q_i\|_F^2 = \frac{1}{N}\sum_i \mathrm{tr}(Q_i(I-\Pi)Q_i) \\ = r - \mathrm{tr}(\Pi A^{\mathrm{cand}}).$$

The eigenvalue variational principle therefore gives

$$d_r = \min_{\Pi} \frac{1}{N}\sum_i \|(I-\Pi)Q_i\|_F^2 \geq 0,$$

and equation 54 bounds this average for $\widehat{\Pi}$ by $d_r + \varepsilon_{\mathrm{eig}}$. Decompose $H_i = Q_iH_i + (I-Q_i)H_i$. The triangle inequality and Jensen give

$$\left\|(I-\widehat{\Pi})\bar{H}\right\|_F \leq \frac{1}{N}\sum_i \left\|(I-\widehat{\Pi})Q_iH_i\right\|_F + \frac{1}{N}\sum_i \|(I-Q_i)H_i\|_F \\ \leq M\left(\frac{1}{N}\sum_i \left\|(I-\widehat{\Pi})Q_i\right\|_F^2\right)^{1/2} + \epsilon_{\mathrm{loc}} \\ \leq M\sqrt{d_r+\varepsilon_{\mathrm{eig}}} + \epsilon_{\mathrm{loc}}.$$

Here $\|XY\|_F \leq \|X\|_F \|Y\|_2$ controls each product. Contraction and $G = \bar{H} + (G-\bar{H})$ prove equation 55. Pythagoras proves equation 56. No independence between the H_i, their projectors, and G is needed. □

Application to FedLore. Under full participation and for $t \geq 1$, let $H_i^t = g_i^{t-1,K-1}$ be the last *positive* local gradient estimate from the preceding round, and let $Q_i^t = \widehat{P}_i^{t-1}(\widehat{P}_i^{t-1})^\top$ be its candidate projector, where $\widehat{P}_i^{t-1}$ denotes the uploaded candidate basis (called P_i^{t-1} in the main algorithm), not the basis used for local optimization in the preceding round. A negative gradient produces the same singular subspace. Then A_t^{cand} and its leading projector Π^t are $\mathcal{F}_t$-measurable. Apply Proposition 2 to these matrices with $G = G_t$. Denote its three-term right-hand side by

$$R_t^{\mathrm{sel}} = \epsilon_{\mathrm{loc},t} + M_t\sqrt{d_{r,t}+\varepsilon_{\mathrm{eig},t}} + \delta_t, \qquad \delta_t = \left\|N^{-1}\sum_i H_i^t - G_t\right\|_F. \tag{57}$$

The quantities for $t=0$ are unnecessary: its contribution can be bounded directly by $\left\|(I-\Pi^0)G_0\right\|_F^2$. The certificate separates local truncation, disagreement of candidate subspaces (and eigensolver error), and the mismatch between the candidate mean gradient and the current global gradient. It provides a sufficient condition, not a necessary one; the bound may be loose under cancellation.

Corollary 5 (Full-gradient bound for the implemented selection rule)**.** *Under the shared projected-SGD dynamics of Theorem 5, full participation, and the preceding selection rule,*

$$\frac{1}{\Lambda_T}\sum_{t<T} h_t\mathbb{E}\|G_t\|_F^2 \leq \mathcal{B}_T + \frac{h_0}{\Lambda_T}\mathbb{E}\rho_0^2 + \frac{1}{\Lambda_T}\sum_{t=1}^{T-1} h_t\mathbb{E}(R_t^{\mathrm{sel}})^2. \tag{58}$$

The initial projector may be deterministic or random. With $\gamma_t = c/(LK\sqrt{T})$, $0 < c \leq 1/8$, fixed K, N, and finite $\mathbb{E}\rho_0^2$, the right-hand side is $O(T^{-1/2})$ whenever $T^{-1}\sum_{t=1}^{T-1}\mathbb{E}(R_t^{\mathrm{sel}})^2 = O(T^{-1/2})$. Alternatively, if $\mathbb{E}(R_t^{\mathrm{sel}})^2 \leq (1-\mu)\mathbb{E}\|G_t\|_F^2$ for $t \geq 1$ and the initialization satisfies expected coverage with the same $\mu > 0$, the bound $\mathcal{B}_T/\mu$ holds.

Proof. The certificate gives $\rho_t^2 \leq (R_t^{\mathrm{sel}})^2$ for $t \geq 1$. Substitute into equation 42. The rate follows from equation 40; the initialization term is $O(T^{-1})$ for constant-within-run stepsizes. The last assertion follows from the orthogonal decomposition and equation 43. □

Proposition 3 (Mismatch of reused last-step gradients)**.** *Under full participation and the oracle conditions, set $X_i^t = W_i^{t-1,K-1}$ and $H_i^t = g_i^{t-1,K-1}$ for $t \geq 1$. Assume additionally that the*

candidate noises are conditionally uncorrelated in the ambient space at that local step: for $i \neq j$, $\mathbb{E}[\left\langle \varepsilon_i^{t-1,K-1}, \varepsilon_j^{t-1,K-1} \right\rangle_F \mid \mathcal{H}_{t-1,K-1}] = 0$. Then the mismatch in equation 57 satisfies

$$\mathbb{E}\delta_t^2 \leq \frac{2\sigma^2}{N} + \frac{2L^2}{N} \sum_i \mathbb{E} \left\| X_i^t - W^t \right\|_F^2 . \tag{59}$$

Proof. Write $\bar{\varepsilon} = N^{-1} \sum_i [H_i^t - \nabla f_i(X_i^t)]$. The last-step oracle conditions imply $\mathbb{E} \left\| \bar{\varepsilon} \right\|_F^2 \leq \sigma^2/N$ by the additional ambient-space uncorrelatedness assumption. Since

$$\frac{1}{N} \sum_i H_i^t - G_t = \bar{\varepsilon} + \frac{1}{N} \sum_i [\nabla f_i(X_i^t) - \nabla f_i(W^t)],$$

the two-term squared-norm inequality, Jensen, and smoothness give the bound. The candidate noise can be correlated with W^t; no cross-term cancellation between these two summands is used. □

Scope of this construction bound. For Proposition 3, ambient-space conditional uncorrelatedness of the candidate noises is an additional hypothesis. With only the original projected-noise condition, Jensen gives the valid replacement $2\sigma^2$ for $2\sigma^2/N$. Fixed candidate minibatch noise need not vanish with T, and the candidate truncation and disagreement terms need not vanish either. Thus equation 58 does not establish unconditional $O(T^{-1/2})$ full stationarity of the implemented rule. It states explicitly which residuals must be controlled. For partial participation, constructing candidates from only the previous active subset requires an additional sampling-error analysis; the full-participation certificate is not silently used as a population-gradient estimator bound in that setting.

Proposition 4 (Agreement alone does not imply coverage)**.** *There exist two 1-smooth strongly convex quadratic objectives with bounded heterogeneity for which exact local leading singular projectors agree uniquely, but their average selects a shared projector with zero coverage at a nonstationary point.*

Proof. Take $f_i(W) = \frac{1}{2} \left\| W \right\|_F^2 + \langle C_i, W \rangle_F$, where $C_1 = \mathrm{diag}(2, 1)$ and $C_2 = \mathrm{diag}(-2, 1)$. At $W = 0$, both unique leading rank-one left projectors are $Q_1 = Q_2 = \mathrm{diag}(1, 0)$, whereas $G = \mathrm{diag}(0, 1)$. Consequently $\widehat{\Pi} = Q_1$, $b = 0$, and $\widehat{\Pi} G = 0$ although $\left\| G \right\|_F^2 = 1$. The heterogeneity is $N^{-1} \sum_i \left\| C_i - G \right\|_F^2 = 4$ at every W. With $K = 1$, full participation, exact gradients, and $\Pi^0 = Q_1$, the two local projected updates cancel. The server stays at $W = 0$, and the last-step candidate rule selects the same projector in every round. Here $d_r = 0$, $\delta = 0$, and $\epsilon_{\mathrm{loc}} = 1$, so equation 55 is attained with equality. □

Coverage of an approximate shared subspace

This subsection treats a different construction: the leading singular subspace of a global-gradient estimate. Its results are benchmarks and sufficient conditions for that construction, not guarantees for the candidate-projector average. The latter is covered separately by Proposition 2.

Proposition 5 (Shared subspace selection)**.** *Fix a point W with global gradient G. Under full participation, exact gradients, and $K = 1$, any shared rank-r orthogonal projector yields*

$$f(W - \gamma \Pi G) \leq f(W) - \gamma(1 - L\gamma/2) \left\| \Pi G \right\|_F^2 . \tag{60}$$

For $0 < \gamma < 2/L$, a projector onto the leading r left singular vectors of G maximizes the certified decrease on the right-hand side. For $G \neq 0$, its retained fraction is

$$\mu_\star(G) = \frac{\sum_{j=1}^r \sigma_j(G)^2}{\left\| G \right\|_F^2} \geq \frac{\min\{r, q\}}{q}, \qquad q = \mathrm{rank}(G), \tag{61}$$

where singular values are padded by zeros to length m. In particular, $\mu_\star(G) \geq r/m$ for $1 \leq r \leq m$.

Proof. Smoothness, $\langle G, \Pi G \rangle_F = \left\| \Pi G \right\|_F^2$, and the shared aggregation identity prove equation 60. To maximize captured energy, write $GG^\top = \sum_{j=1}^m \lambda_j u_j u_j^\top$ with $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. For an

orthogonal rank-r projector, $a_j = u_j^\top \Pi u_j$ satisfies $0 \le a_j \le 1$ and $\sum_j a_j = r$. Therefore,

$$\|\Pi G\|_F^2 = \operatorname{tr}(\Pi G G^\top) = \sum_j \lambda_j a_j \le \sum_{j=1}^{r} \lambda_j.$$

Equality is attained by the leading eigenspace. Exactly q eigenvalues are positive. If $r < q$, the mean of the largest r is at least the mean of those q values; if $r \ge q$, all energy is retained. This proves equation 61. The optimum is for the smoothness bound; it need not minimize the actual post-update loss of a nonquadratic objective. □

Proposition 6 (Coverage from an approximate global-gradient estimate)**.** *Let $G \neq 0$, let H satisfy $\|H - G\|_F \le \epsilon \|G\|_F$, and let $\widehat{\Pi}$ be a leading rank-r left singular projector of H. If $\Pi_\star$ is a leading rank-r projector of G, then*

$$\frac{\left\|\widehat{\Pi} G\right\|_F^2}{\|G\|_F^2} \ge \left[\max\{\sqrt{\mu_\star(G)} - 2\epsilon, 0\}\right]^2. \tag{62}$$

In particular, uniform bounds $\mu_\star(G_t) \ge \underline{\mu}_\star > 0$ and $\epsilon_t \le \bar{\epsilon} < \sqrt{\underline{\mu}_\star}/2$ imply Assumption 7 for the shared method with $\mu = (\sqrt{\underline{\mu}_\star} - 2\bar{\epsilon})^2$.

Proof. Triangle inequalities, contraction, and optimality on H give

$$\begin{aligned}
\left\|\widehat{\Pi} G\right\|_F &\ge \left\|\widehat{\Pi} H\right\|_F - \|H - G\|_F \\
&\ge \|\Pi_\star H\|_F - \|H - G\|_F \\
&\ge \|\Pi_\star G\|_F - 2\|H - G\|_F \\
&\ge (\sqrt{\mu_\star(G)} - 2\epsilon)\|G\|_F .
\end{aligned}$$

The norm is nonnegative, so take the positive part before squaring. At $G = 0$, the coverage inequality is automatically true. □

Proposition 7 (Additive coverage error)**.** *Replace Assumption 7 by*

$$\langle G_t, A_t G_t \rangle_F \ge \mu \|G_t\|_F^2 - \chi_t, \qquad \chi_t \ge 0, \tag{63}$$

where χ_t is $\mathcal{F}_t$-measurable and integrable. All other assumptions and the stepsize restriction remain unchanged. Then

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|G_t\|_F^2 \le \mathcal{R}_T + \frac{8}{\mu^2 T}\sum_{t<T} \mathbb{E}\|b_t\|_F^2 + \frac{4}{\mu T}\sum_{t<T} \mathbb{E}\chi_t. \tag{64}$$

In particular, let H_t be $\mathcal{F}_t$-measurable with $\mathbb{E}\|H_t - G_t\|_F^2 < \infty$, and let the shared projector $\widehat{\Pi}_t$ maximize $\|\Pi H_t\|_F^2$ over rank-r orthogonal projectors. If $\mu_\star(G_t) \ge \underline{\mu}_\star > 0$ whenever $G_t \neq 0$, then equation 64 holds with

$$\mu = \frac{\underline{\mu}_\star}{2}, \qquad \chi_t = 4\|H_t - G_t\|_F^2, \qquad b_t = 0. \tag{65}$$

Proof. Under equation 63, the right-hand side of equation 35 acquires only the additive term χ_t. Consequently, the one-round inequality equation 38 acquires $h\chi_t$. Summing and dividing by $\mu h T/4$ proves equation 64.

Let $\Pi_{\star,t}$ be a leading rank-r projector of G_t. The triangle inequality and optimality on H_t imply

$$\|\Pi_{\star,t} G_t\|_F \le \|\Pi_{\star,t} H_t\|_F + \|H_t - G_t\|_F \le \left\|\widehat{\Pi}_t G_t\right\|_F + 2\|H_t - G_t\|_F .$$

Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\left\|\widehat{\Pi}_t G_t\right\|_F^2 \ge \tfrac{1}{2}\|\Pi_{\star,t} G_t\|_F^2 - 4\|H_t - G_t\|_F^2 \ge \tfrac{1}{2}\underline{\mu}_\star \|G_t\|_F^2 - 4\|H_t - G_t\|_F^2 .$$

This is equation 63 with the stated constants. At $G_t = 0$ the inequality is automatic. □

Unlike a uniform relative-error condition, the additive formulation does not require the estimator error to shrink in proportion to a vanishing true gradient. With $\gamma = c/(LK\sqrt{T})$ and $0 < c \le \mu/16$, an average squared estimation error of order $T^{-1/2}$ suffices to retain the same stationarity rate. A fixed error level instead contributes its explicit average penalty in equation 64.

Corollary 6 (Stale global-gradient estimates)**.** *Under Proposition 7, suppose*

$$H_t = \nabla f(W^{t-\tau_t}) + E_t^{\mathrm{est}}, \qquad 0 \le \tau_t \le t.$$

The delay and both terms are $\mathcal{F}_t$-measurable. Provided the following right-hand side is integrable, one may take

$$\chi_t = 8\left\|E_t^{\mathrm{est}}\right\|_F^2 + 8L^2\tau_t \sum_{j=t-\tau_t}^{t-1} \left\|W^{j+1} - W^j\right\|_F^2. \tag{66}$$

The sum is zero when $\tau_t = 0$.

Proof. By smoothness of f,

$$\begin{aligned}\|H_t - G_t\|_F^2 &\le 2\left\|E_t^{\mathrm{est}}\right\|_F^2 + 2L^2\left\|W^{t-\tau_t} - W^t\right\|_F^2 \\ &\le 2\left\|E_t^{\mathrm{est}}\right\|_F^2 + 2L^2\tau_t \sum_{j=t-\tau_t}^{t-1} \left\|W^{j+1} - W^j\right\|_F^2.\end{aligned}$$

Multiply by four and use equation 65. □

EXACT SYNCHRONIZATION AND OPTIMIZER SCOPE

Proposition 8 (Exact coordinate aggregation with common decay)**.** *Fix a shared basis P during a round. Suppose all clients start at the same W and use updates*

$$W_{i,k+1} = a_k W_{i,k} + PD_{i,k}, \qquad W_{i,0} = W,$$

where a_k are common known scalars and $D_{i,k} \in \mathbb{R}^{r\times n}$ may be arbitrary, including coordinatewise adaptive updates. Define

$$a = \prod_{k=0}^{K-1} a_k, \qquad Z_i = \sum_{k=0}^{K-1} \left(\prod_{\ell=k+1}^{K-1} a_\ell\right) D_{i,k}.$$

An empty product is one. For any participating set $\mathcal{S}$,

$$W_{i,K} = aW + PZ_i, \qquad \frac{1}{S}\sum_{i\in\mathcal{S}} W_{i,K} = aW + P\left(\frac{1}{S}\sum_{i\in\mathcal{S}} Z_i\right). \tag{67}$$

Proof. Induction in k unrolls the common scalar recursion and gives the endpoint factorization. Linearity of the mean gives the second identity. □

For zero weight decay, $a_k = 1$. With stepsize η_k and scale α, the Adam implementation uses $D_{i,k} = \eta_k \alpha N_{i,k+1}$ and must accumulate these *weighted* coordinates if the learning rate varies within a round. Constant learning rate reduces to the factorization in the main algorithm. With common decoupled weight decay λ, $a_k = 1 - \eta_k\lambda$. The increment then contains the generally full-rank term $(a-1)W$, but a client holding the synchronized cache W reconstructs it from a without receiving a dense matrix.

Under full participation and synchronized caches, each projected matrix requires rn coordinate scalars and mr candidate-basis scalars uploaded per client, and rn aggregated-coordinate scalars and mr next-basis scalars downloaded per client per round, excluding initialization and unprojected parameters. This is $r(m+n)$ scalars in each direction. The two coordinate moment tensors occupy $2rn$ scalars; this is not a claim about total peak training memory. Under intermittent participation, a returning client additionally needs the missed increments and their bases, or a fresh model. The convergence sampling model does not itself guarantee this full-participation communication cost.

Adaptive updates. Proposition 8 holds for the Adam implementation because it uses only a fixed basis and linear aggregation of endpoint coordinates. The bias identity concerns linear projected gradients, and the stationarity theorems concern projected SGD without weight decay. In general, averaging independently normalized Adam directions does not equal normalizing an averaged gradient. Sharing therefore does not imply that all adaptive-optimizer bias or local drift vanishes. Projector averaging is invariant to the representation of each candidate subspace; coordinatewise Adam need not be invariant under an arbitrary rotation of the shared basis. These are distinct properties.

Diagnostics. At selected checkpoints, measure the retained-energy ratio $\|\Pi^t G_t\|_F^2 / \|G_t\|_F^2$ when $G_t \neq 0$, the residual ρ_t^2, and, for client-specific projections, $\|b_t\|_F^2$ and $\langle G_t, A_t G_t + b_t \rangle_F$. The direction of the bias matters, so projector disagreement alone is insufficient evidence of harmful coupling. For the implemented selection rule, also report local discarded energy, $d_{r,t}$, and candidate-gradient staleness where feasible. Diagnostic global-gradient estimation has its own cost and need not be part of the training protocol. The geometric certificate is valid for candidate gradients from either implementation. Its stationarity consequence requires the projected-SGD updates specified in Corollary 5.

C4 Partitioning Protocol

Domain-based grouping. Since C4 does not provide classification labels, we construct heterogeneous client datasets using the source information associated with each document. For every training document, we extract the hostname from its URL, convert it to lowercase, and remove a leading `www.` prefix. Documents with the same normalized hostname form a source group. Different subdomains remain separate groups, and documents with missing or invalid hostnames are assigned to a dedicated unknown-source group. These groups serve only as partitioning metadata; they are not prediction targets and are not added to the model inputs. Source identity provides a proxy for differences in document content and writing style, rather than an explicit semantic topic label.

C4 Dirichlet allocation. Let $\mathcal{D}_{\text{train}}$ denote the selected C4 training corpus before client allocation, and let $\{\mathcal{D}_c\}_{c=1}^{C}$ be its source groups. We partition these documents among $N = 20$ clients. For each group c, we independently sample

$$\boldsymbol{p}_c \sim \text{Dirichlet}\left(\alpha_{\text{C4}} \mathbf{1}_N\right), \qquad \alpha_{\text{C4}} = 0.5, \tag{68}$$

where $p_{c,i}$ specifies the target fraction of group c assigned to client i. Smaller concentration parameters produce more uneven allocations of each source group across clients.

To convert these proportions into integer document counts, let $M_c = |\mathcal{D}_c|$ and initialize

$$n_{c,i} = \lfloor M_c p_{c,i} \rfloor . \tag{69}$$

The remaining $M_c - \sum_{i=1}^{N} n_{c,i}$ documents are assigned, one per client, to those with the largest fractional remainders $M_c p_{c,i} - n_{c,i}$. Ties are resolved by client index. We randomly shuffle the documents within each group and allocate consecutive blocks of sizes $n_{c,1}, \ldots, n_{c,N}$ to the clients. The resulting client datasets satisfy

$$\mathcal{D}_{\text{train}} = \bigcup_{i=1}^{N} \mathcal{D}_i, \qquad \mathcal{D}_i \cap \mathcal{D}_j = \varnothing \quad \text{for } i \neq j. \tag{70}$$

Thus, every training document belongs to exactly one client.

Client sizes and local sampling. This procedure allows both source composition and dataset size to vary across clients. We do not rebalance client datasets after allocation, because doing so would alter the sampled distribution. If any client receives no documents, we redraw the allocation using the next values from the same seeded random-number generator. All clients participate in each communication round and perform $K = 50$ local steps with batch size 16. Each client samples from its own shuffled data stream, reshuffling and restarting the stream when it is exhausted. Consequently, clients perform the same number of local updates even when their dataset sizes differ.

Tokenization and sequence construction. Client assignment is performed at the document level, before tokenization or sequence packing. Within each client, documents are tokenized using the same tokenizer, separated by an end-of-sequence token, and packed into sequences of length 256. A final incomplete block is padded, with padding positions excluded from the language-modeling loss. Packing is performed separately for each client, so no training sequence contains documents from different clients. The source metadata used for partitioning are excluded from the training text.

Evaluation and reproducibility. The validation corpus is kept separate from the training partition and is identical across methods. We evaluate all global models on this common validation corpus, computing perplexity as the exponential of the mean negative log-likelihood over non-padding target tokens. For each seed in $\{42, 43, 44, 45, 46\}$, we generate one client partition and reuse it for every compared method. The selected training documents, tokenizer, document assignments, and validation corpus are therefore matched across methods within each run. We use separate random-number streams for partitioning and model training so that differences in training code do not change the client allocation.

BASELINE ALGORITHM

Algorithm 2: Local GaLore in Federated Training

Require: Global weights W^0, learning rate η, scaling factor α, rank r, decay rates β_1, β_2, numerical constant ϵ, communication rounds T, local steps K, number of clients N, weight decay $\lambda = 0$

Ensure: Trained global model W^T

for $t = 0$ to $T - 1$ **do**
  Server broadcasts current global model W^t to all clients
  for each client $i \in \{1, \dots, N\}$ in parallel **do**
    Compute a round-start gradient using a separate construction mini-batch (independent of the subsequent local optimization draws): $G_i^t \leftarrow -\nabla_W f_i(W^t; \xi_i^t) \in \mathbb{R}^{m \times n}$
    Compute truncated SVD: $U, S, V \leftarrow \mathrm{SVD}(G_i^t)$
    Construct local projector: $P_i^t \leftarrow U[:, 1:r] \in \mathbb{R}^{m \times r}$
    Initialize $M_i^{t,0} \leftarrow 0$, $V_i^{t,0} \leftarrow 0$, $W_i^{t,0} \leftarrow W^t$
    for $k = 0$ to $K - 1$ **do**
      Sample a fresh mini-batch $\xi_i^{t,k}$ and compute gradient: $G_i^{t,k} \leftarrow -\nabla_W f_i(W_i^{t,k}; \xi_i^{t,k}) \in \mathbb{R}^{m \times n}$
      Project gradient to the local low-rank subspace: $R_i^{t,k} \leftarrow (P_i^t)^\top G_i^{t,k} \in \mathbb{R}^{r \times n}$
      Adam update in the low-rank space:
      $M_i^{t,k+1} \leftarrow \beta_1 M_i^{t,k} + (1 - \beta_1) R_i^{t,k}$
      $V_i^{t,k+1} \leftarrow \beta_2 V_i^{t,k} + (1 - \beta_2)(R_i^{t,k} \odot R_i^{t,k})$
      $\widehat{M}_i^{t,k+1} \leftarrow M_i^{t,k+1} / (1 - \beta_1^{k+1})$
      $\widehat{V}_i^{t,k+1} \leftarrow V_i^{t,k+1} / (1 - \beta_2^{k+1})$
      $N_i^{t,k+1} \leftarrow \widehat{M}_i^{t,k+1} / (\sqrt{\widehat{V}_i^{t,k+1}} + \epsilon)$
      Reconstruct the ambient-space low-rank direction: $\widehat{G}_i^{t,k} \leftarrow \alpha P_i^t N_i^{t,k+1} \in \mathbb{R}^{m \times n}$
      Update local weights: $W_i^{t,k+1} \leftarrow W_i^{t,k} + \eta \widehat{G}_i^{t,k} - \lambda \eta W_i^{t,k}$
    end for
    Accumulate updates: $\Delta W_i^t \leftarrow W_i^{t,K} - W_i^{t,0}$
    Client i uploads (ΔW_i^t) to the server
  end for
  Aggregate updates in the ambient space:
  $W^{t+1} \leftarrow W^t + \frac{1}{N} \sum_{i=1}^N \Delta W_i^t$
end for

Algorithm 3: `FedLore` Training Procedure

Require: Global weights W^0; learning rate η; scale α; rank r; Adam coefficients β_1, β_2; $\epsilon > 0$; communication rounds T; local steps K; client set $\mathcal{N}$ with $|\mathcal{N}| = N$; weight decay $\lambda = 0$.

Ensure: Trained global model W^T.

Initialize a shared orthonormal basis $P^0 \in \mathbb{R}^{m \times r}$
Server broadcasts (W^0, P^0) to all clients
Each client initializes global-model cache: $\widetilde{W}_i \leftarrow W^0$
for $t = 0$ to $T - 1$ **do**
  for each client $i \in \mathcal{N}$ **in parallel do**
    `//Low-Rank Model Synchronization`
    if $t > 0$ **then**
      Receive (P^t, Δ^{t-1}) from the server
      $\widetilde{W}_i \leftarrow \widetilde{W}_i + \eta\alpha P^{t-1}\Delta^{t-1}$
    end if
    $W_i^{t,0} \leftarrow \widetilde{W}_i$, $M_i^{t,0} \leftarrow 0$, $V_i^{t,0} \leftarrow 0$, $\Delta Z_i^t \leftarrow 0$
    `//Local Projected Adam Updates`
    for $k = 0$ to $K - 1$ **do**
      $G_i^{t,k} \leftarrow -\nabla_W f_i(W_i^{t,k}; \xi_i^{t,k}) \in \mathbb{R}^{m \times n}$; $R_i^{t,k} \leftarrow (P^t)^\top G_i^{t,k} \in \mathbb{R}^{r \times n}$
      $M_i^{t,k+1} \leftarrow \beta_1 M_i^{t,k} + (1 - \beta_1) R_i^{t,k}$; $V_i^{t,k+1} \leftarrow \beta_2 V_i^{t,k} + (1 - \beta_2)(R_i^{t,k} \odot R_i^{t,k})$
      $\widehat{M}_i^{t,k+1} \leftarrow M_i^{t,k+1}/(1 - \beta_1^{k+1})$; $\widehat{V}_i^{t,k+1} \leftarrow V_i^{t,k+1}/(1 - \beta_2^{k+1})$
      $N_i^{t,k+1} \leftarrow \widehat{M}_i^{t,k+1}/(\sqrt{\widehat{V}_i^{t,k+1}} + \epsilon)$; $W_i^{t,k+1} \leftarrow W_i^{t,k} + \eta\alpha P^t N_i^{t,k+1}$
      $\Delta Z_i^t \leftarrow \Delta Z_i^t + N_i^{t,k+1}$
    end for
    `//Local Candidate Subspace`
    Reuse the last local gradient $G_i^{t,K-1}$; $P_i^t \leftarrow \mathrm{TopLeftSVD}_r(G_i^{t,K-1}) \in \mathbb{R}^{m \times r}$
    Upload $(\Delta Z_i^t, P_i^t)$ to the server; Retain P^t for next-round model synchronization
  end for
  `//Server-Side Update Aggregation`
  $\Delta^t \leftarrow \frac{1}{N}\sum_{i \in \mathcal{N}} \Delta Z_i^t$; $W^{t+1} \leftarrow W^t + \eta\alpha P^t \Delta^t$
  `//Shared-Subspace Construction`
  $B^t \leftarrow \frac{1}{\sqrt{N}}[P_1^t, \ldots, P_N^t] \in \mathbb{R}^{m \times Nr}$; $P^{t+1} \leftarrow \mathrm{TopLeftSVD}_r(B^t) \in \mathbb{R}^{m \times r}$
end for

Algorithm 4: `FedLore-Random` Training Procedure

Require: Global weights $W^0 \in \mathbb{R}^{m\times n}$; learning rate η; scale α; rank $1 \le r \le m$; Adam coefficients β_1, β_2; $\epsilon > 0$; communication rounds T; local steps K; client set $\mathcal{N}$ with $|\mathcal{N}| = N$; weight decay $\lambda = 0$.
Ensure: Trained global model W^T.

Server samples a random seed s^0 and sets $P^0 \leftarrow \mathrm{SharedBasis}(s^0, m, r)$
Server broadcasts (W^0, s^0) to all clients
Each client initializes $\widetilde{W}_i \leftarrow W^0$ and reconstructs $P^0 \leftarrow \mathrm{SharedBasis}(s^0, m, r)$
for $t = 0$ to $T-1$ **do**
  if $t > 0$ **then**
    Server samples a fresh random seed s^t, independently of the training randomness
    Server sets $P^t \leftarrow \mathrm{SharedBasis}(s^t, m, r)$ and broadcasts (s^t, Δ^{t-1})
  end if
  for each client $i \in \mathcal{N}$ **in parallel do**
    `// Low-Rank Model Synchronization`
    if $t \ge 0$ **then**
      $\widetilde{W}_i \leftarrow \widetilde{W}_i + \eta\alpha P^{t-1}\Delta^{t-1}$
      Reconstruct the shared random basis: $P^t \leftarrow \mathrm{SharedBasis}(s^t, m, r)$
    end if
    $W_i^{t,0} \leftarrow \widetilde{W}_i$, $M_i^{t,0} \leftarrow 0$, $V_i^{t,0} \leftarrow 0$, $\Delta Z_i^t \leftarrow 0$
    `// Local Projected Adam Updates`
    for $k = 0$ to $K-1$ **do**
      Sample a fresh mini-batch $\xi_i^{t,k}$
      $G_i^{t,k} \leftarrow -\nabla_W f_i(W_i^{t,k}; \xi_i^{t,k})$; $R_i^{t,k} \leftarrow (P^t)^\top G_i^{t,k}$
      $M_i^{t,k+1} \leftarrow \beta_1 M_i^{t,k} + (1-\beta_1)R_i^{t,k}$; $V_i^{t,k+1} \leftarrow \beta_2 V_i^{t,k} + (1-\beta_2)(R_i^{t,k} \odot R_i^{t,k})$
      $\widehat{M}_i^{t,k+1} \leftarrow M_i^{t,k+1}/(1-\beta_1^{k+1})$; $\widehat{V}_i^{t,k+1} \leftarrow V_i^{t,k+1}/(1-\beta_2^{k+1})$
      $N_i^{t,k+1} \leftarrow \widehat{M}_i^{t,k+1}/(\sqrt{\widehat{V}_i^{t,k+1}} + \epsilon)$; $W_i^{t,k+1} \leftarrow W_i^{t,k} + \eta\alpha P^t N_i^{t,k+1}$
      $\Delta Z_i^t \leftarrow \Delta Z_i^t + N_i^{t,k+1}$
    end for
    Upload ΔZ_i^t to the server; retain P^t for next-round model synchronization
  end for
  `// Server-Side Update Aggregation`
  $\Delta^t \leftarrow \frac{1}{N}\sum_{i\in\mathcal{N}} \Delta Z_i^t$; $W^{t+1} \leftarrow W^t + \eta\alpha P^t \Delta^t$
end for