RETHINKING AGGREGATION ERROR IN FEDERATED LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-Rank Adaptation (LoRA) is the dominant approach for parameter-efficient federated fine-tuning, but independently averaging the two factor matrices across clients does not recover the average of their products. This *aggregation inconsistency* has motivated a family of mitigation methods (freezing one factor, alternating the frozen factor, correcting the residual at the server, or averaging in product space), all built on the unstated assumption that the error is harmful. **We revisit this assumption.** We show that the aggregation error equals the weighted cross-factor covariance, is bounded by the product of factor dispersions, and decomposes into three geometric components dominated by subspace misalignment ($\gtrsim 84\%$). To test whether the error causally harms accuracy, we introduce a *controlled instrument*: a per-client regularizer of strength λ, in two forms, used not as a method but to vary error along a single axis, and we cross-validate its conclusion along further axes that involve no regularization at all: local steps, LoRA rank, client count, and exact error-free aggregation. Across two model scales (RoBERTa-base, DeBERTa-v3-large), three seeds, and more than three orders of magnitude in λ, **error varies by more than $30\times$ (RoBERTa-base) while accuracy among non-freezing methods varies by less than 1.5%**; error grows $190\times$ with local steps and $2.5\times$ with client count while accuracy rises with the former and stays flat with the latter; and zero-error methods range from -3.2% to $+1.5\%$ relative to matched FedAvg baselines, so the error level predicts neither the sign nor the size of accuracy differences; the accompanying mechanism does. **In the evaluated settings (encoder models, NLU classification), the practical cost of aggregation inconsistency lies not in the error itself but in the constraints imposed to eliminate it**; we state explicit conditions under which this conclusion may fail.

1 INTRODUCTION

Fine-tuning data for foundation models is often distributed across hospitals, banks, or mobile devices where centralization is infeasible. Federated learning (FL) addresses this (McMahan et al., 2017; Kairouz et al., 2021), but full fine-tuning of a 125M–7B model in FL is prohibitive. Low-Rank Adaptation (LoRA) (Hu et al., 2022) makes federated fine-tuning practical: by training $\Delta\mathbf{W} = \mathbf{AB}$ ($\mathbf{A} \in \mathbb{R}^{d\times r}$, $\mathbf{B} \in \mathbb{R}^{r\times d}$, $r \ll d$) it reduces communication to $O(dr)$ per round (Zhang et al., 2024; Cho et al., 2024).

LoRA's bilinear structure, however, creates an aggregation inconsistency unique to federation: factor-wise averaging gives $\bar{\mathbf{A}}\bar{\mathbf{B}} \neq \sum_k \omega_k \mathbf{A}_k\mathbf{B}_k$ in general (Wang et al., 2024; Sun et al., 2024). A growing body of work treats this gap as a problem to be solved: FFA-LoRA freezes $\mathbf{A}$ (Sun et al., 2024), RoLoRA alternates the frozen factor (Chen et al., 2025), FedEx-LoRA absorbs the residual into the frozen weights (Singhal et al., 2025), and FlexLoRA and FLoRA aggregate in product space (Bai et al., 2024; Wang et al., 2024), accepting reduced expressiveness or added complexity as the price.

But is aggregation error actually the problem? Rather than proposing yet another mitigation, we ask whether the error is harmful enough to warrant the constraints these methods impose, and find that, in the settings we evaluate, it is not.

(C1) Covariance framework. We show the error is *exactly* the weighted cross-factor covariance $\mathbf{C}_\omega(\mathbf{A}, \mathbf{B}) = \sum_k \omega_k(\mathbf{A}_k - \bar{\mathbf{A}})(\mathbf{B}_k - \bar{\mathbf{B}})$ (Proposition 1), bound its Frobenius norm by the dispersion

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.

product (Proposition 2), and decompose the covariance into subspace misalignment, scale divergence, and cross-coupling. These identities are elementary; their value is that they organize both the taxonomy of existing methods and the instrument we build.

(C2) A controlled instrument. To isolate the effect of aggregation error on accuracy, we introduce a per-client regularizer in two forms (a dispersion-product (CP) penalty derived from Proposition 2, and an output-space (OS) penalty that targets the error directly), used as a *scientific instrument* rather than a method. Its strength λ dials error downward from FedAvg (λ=0). Because any regularizer also changes the objective, we cross-check its conclusion on four axes that involve no regularization: local steps, rank, client count, and exact error-free aggregation.

(C3) Aggregation error is benign in the evaluated settings. On RoBERTa-base and DeBERTa-v3-large, *error varies by* $>30\times$ *while accuracy varies by* $<1.5\%$ *among non-freezing methods*. FedAvg and a rank-r post-hoc correction (LS-Corr) achieve the highest accuracies (84.5%, 84.9%) at the highest error (0.036), whereas factor-freezing methods that drive error to zero lose 1–3$\%$ on RoBERTa. With the round budget fixed, FedAvg's error grows $\sim 190\times$ with local steps while accuracy rises monotonically; error grows $2.5\times$ with client count while accuracy stays flat; and an exact error-free aggregator (FlexLoRA) changes accuracy by $+1.5$pp relative to a same-pipeline FedAvg baseline, a difference attributable to its re-factorization, not to the absence of error.

(C4) Taxonomy and design implications. The framework sorts existing methods into four families (Table 5). Factor-freezing zeros one dispersion (a sufficient but unnecessarily strong condition); a rank-r post-hoc correction cannot represent error in a $d\times d$ space; FedProx penalizes all drift uniformly; exact aggregators (FedEx-LoRA, FLoRA, FlexLoRA) remove the error at the cost of a full-rank residual or a per-round re-factorization. The theory predicts *where* the error is large (growing with local steps and client count, shrinking with rank), and the experiments confirm these predictions; whether the error is *harmful* is not a theorem but a controlled empirical question, and we treat it as one. The instrument achieves 33–96$\%$ error reduction at no significant accuracy cost: error reduction and expressiveness are not in tension, but the reduction buys no accuracy.

2 Related work

Federated LoRA and responses to aggregation error. Federated LoRA (Hu et al., 2022) is widely adopted for communication-efficient FL (Zhang et al., 2024; Cho et al., 2024; Bai et al., 2024; Zhuang et al., 2023) under non-IID data (Zhao et al., 2018; Hsu et al., 2019; Li et al., 2022). The factor- vs. product-wise aggregation inconsistency was identified in several lines of work (Wang et al., 2024; Sun et al., 2024; Singhal et al., 2025; Kim & Choi, 2025), and the literature's responses fall into four families. (i) *Manage it*: FedAvg leaves the error untouched; general drift-control methods (FedProx (Li et al., 2020), SCAFFOLD (Karimireddy et al., 2020), FedDyn (Acar et al., 2021), MOON (Li et al., 2021), adaptive server optimizers (Reddi et al., 2021; Wang et al., 2020; Yang et al., 2021)) ignore LoRA's bilinear structure; low-rank server-side corrections such as LoRA-FAIR's aggregation refinement (Bian et al., 2025) cannot represent a full-rank residual (r=8 $\ll$ d=768). (ii) *Eliminate it by constraint*: FFA-LoRA (Sun et al., 2024) and RoLoRA (Chen et al., 2025) freeze one factor (permanently or alternately), a sufficient but unnecessarily strong condition. (iii) *Eliminate it by exact aggregation*: FedEx-LoRA (Singhal et al., 2025) keeps factor-wise averaging but absorbs the residual $\sum_k \omega_k \mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}\bar{\mathbf{B}}$ into the frozen weights $\mathbf{W}_0$, making the global update exact; FLoRA (Wang et al., 2024) stacks client factors so that the sum of products is formed exactly; FlexLoRA (Bai et al., 2024) averages the products and re-factorizes by truncated SVD; and FedRot-LoRA (Zhang et al., 2026) aligns client factors by orthogonal rotations before averaging. (iv) *Sidestep it by not aggregating one factor*: FedSA-LoRA (Guo et al., 2025) shares only $\mathbf{A}$ and keeps each client's $\mathbf{B}$ local, a personalized design with no single global model. Our analysis concerns methods that form a global model (families i–iii); we discuss family (iv) in Appendix F. All of this work *assumes* the error is harmful. To our knowledge, ours is the first to express the error as a cross-factor covariance and the first controlled test of its effect.

Rotational vs. semantic mismatch. FedRot-LoRA targets *gauge* misalignment: since $(\mathbf{AR})(\mathbf{R}^\top\mathbf{B}) = \mathbf{AB}$, semantically identical products can sit in rotated factor bases and interfere under factor averaging, a removable artifact of factorization non-uniqueness. The subspace component we identify (Sec. 3.3) is, in our cross-task setting, dominated by genuine semantic hetero-

geneity, which no rotation can align. Our decomposition does not separate the two mechanisms; the works are complementary, and our finding concerns the error that remains after any such alignment.

LoRA theory and variants. Bilinear factorization is studied in matrix sensing (Bhojanapalli et al., 2016; Du et al., 2018) and implicit regularization (Arora et al., 2019; Gunasekar et al., 2017); federated bilinear convergence in Sun et al. (2024). LoRA variants (Zhang et al., 2023; Dettmers et al., 2023; Liu et al., 2024; Kopiczko et al., 2024; Hayou et al., 2024; Renduchintala et al., 2024), intrinsic dimensionality (Aghajanyan et al., 2021), and learning dynamics (Biderman et al., 2024) are complementary: our decomposition explains *why* factor-freezing loses accuracy (expressiveness) and *when* error control matters (cross-task heterogeneity).

3 Theoretical framework

This section gives an exact expression for the aggregation error, bounds it by a computable quantity, and decomposes it into geometric components. All three results are exact and require only $\omega_k \geq 0$, $\sum_k \omega_k = 1$; they characterize *what* the error is, not whether it is harmful (Sections 4–5).

3.1 Setup and notation

Consider K clients. In each of T communication rounds, a subset $\mathcal{S}^t \subseteq [K]$ of size ρK (participation rate ρ) is sampled; each participating client initializes from the broadcast global factors, runs P local optimizer steps on its data $\mathcal{D}_k$, and uploads its factors $\mathbf{A}_k \in \mathbb{R}^{d\times r}$, $\mathbf{B}_k \in \mathbb{R}^{r\times d}$, which the server averages with weights $\omega_k = |\mathcal{D}_k| / \sum_{j\in\mathcal{S}^t} |\mathcal{D}_j|$. All sums below run over $\mathcal{S}^t$; we suppress round and projection indices.

Definition 1 (Weighted means, covariance, and dispersions)**.**

$$\bar{\mathbf{A}} = \textstyle\sum_k \omega_k \mathbf{A}_k, \quad \bar{\mathbf{B}} = \sum_k \omega_k \mathbf{B}_k, \tag{1}$$

$$\mathbf{C}_\omega(\mathbf{A},\mathbf{B}) = \textstyle\sum_k \omega_k (\mathbf{A}_k - \bar{\mathbf{A}})(\mathbf{B}_k - \bar{\mathbf{B}}), \tag{2}$$

$$\sigma_\omega(\mathbf{A}) = \sqrt{\textstyle\sum_k \omega_k \left\|\mathbf{A}_k - \bar{\mathbf{A}}\right\|_F^2}, \quad \sigma_\omega(\mathbf{B}) = \sqrt{\textstyle\sum_k \omega_k \left\|\mathbf{B}_k - \bar{\mathbf{B}}\right\|_F^2}. \tag{3}$$

$\mathbf{C}_\omega$ is the matrix analogue of a weighted covariance: it is large when $\mathbf{A}_k$ and $\mathbf{B}_k$ deviate from their means in correlated directions and tends to vanish when they deviate independently.

3.2 Aggregation error is cross-factor covariance

Proposition 1 (Aggregation error identity)**.** $\sum_k \omega_k \mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}\bar{\mathbf{B}} = \mathbf{C}_\omega(\mathbf{A},\mathbf{B})$.

The proof (Appendix A) expands $\mathbf{A}_k = \bar{\mathbf{A}} + \delta\mathbf{A}_k$ and uses $\sum_k \omega_k \delta\mathbf{A}_k = 0$. The identity reveals that aggregation error arises from *correlated* deviations: if $\mathbf{A}_k$ drifts but $\mathbf{B}_k$ stays near $\bar{\mathbf{B}}$, the error is small. Any configuration in the union $\{\delta\mathbf{A}_k{=}0\} \cup \{\delta\mathbf{B}_k{=}0\}$ contributes zero error; factor-freezing forces one deviation to zero for every client, a sufficient but unnecessarily restrictive condition.

Proposition 2 (Dispersion product bound)**.** $\|\mathbf{C}_\omega(\mathbf{A},\mathbf{B})\|_F \leq \sigma_\omega(\mathbf{A}) \cdot \sigma_\omega(\mathbf{B})$.

The proof (Appendix A) uses submultiplicativity and weighted Cauchy–Schwarz. The bound is a *product*: controlling either factor's dispersion suffices. FFA-LoRA sets $\sigma_\omega(\mathbf{A}){=}0$; RoLoRA does so on alternating rounds (Corollary 4). This product structure is what a regularizer should target: a multiplicative penalty $\|\delta\mathbf{A}\|_F^2 \, \|\delta\mathbf{B}\|_F^2$ follows the bound, whereas an additive one $\|\delta\mathbf{A}\|_F^2 + \|\delta\mathbf{B}\|_F^2$ (FedProx) also penalizes single-factor drift, which Proposition 1 shows to be error-free when shared by all clients.

3.3 Decomposition of covariance sources

To understand *where* the error comes from, we decompose factor deviations via thin QR $\bar{\mathbf{A}} = \mathbf{U}\mathbf{R}$ ($\mathbf{U} \in \mathbb{R}^{d\times r}$) and $\bar{\mathbf{B}}^\top = \mathbf{V}\mathbf{S}$ ($\mathbf{V} \in \mathbb{R}^{d\times r}$), assuming $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ have rank r:

$$\delta\mathbf{A}_k = \underbrace{\mathbf{U}\mathbf{U}^\top\delta\mathbf{A}_k}_{\delta\mathbf{A}_k^{\parallel}} + \underbrace{(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\delta\mathbf{A}_k}_{\delta\mathbf{A}_k^{\perp}}, \quad \delta\mathbf{B}_k = \underbrace{\delta\mathbf{B}_k\mathbf{V}\mathbf{V}^\top}_{\delta\mathbf{B}_k^{\parallel}} + \underbrace{\delta\mathbf{B}_k(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)}_{\delta\mathbf{B}_k^{\perp}}, \tag{4}$$

yielding:

$$\mathbf{C}_\omega = \underbrace{\textstyle\sum_k \omega_k \delta\mathbf{A}_k^{\perp}\delta\mathbf{B}_k^{\perp}}_{\mathbf{C}^{\text{sub}}} + \underbrace{\textstyle\sum_k \omega_k \delta\mathbf{A}_k^{\parallel}\delta\mathbf{B}_k^{\parallel}}_{\mathbf{C}^{\text{scale}}} + \underbrace{\textstyle\sum_k \omega_k (\delta\mathbf{A}_k^{\perp}\delta\mathbf{B}_k^{\parallel} + \delta\mathbf{A}_k^{\parallel}\delta\mathbf{B}_k^{\perp})}_{\mathbf{C}^{\text{cross}}}. \tag{5}$$

Interpretation. $\mathbf{C}^{\text{sub}}$ collects deviations *orthogonal* to $\mathrm{col}(\bar{\mathbf{A}})$ and $\mathrm{row}(\bar{\mathbf{B}})$: an SST-2 client learning sentiment features and an MNLI client learning entailment features push their deviations into near-orthogonal subspaces, and these components dominate empirically ($\gtrsim 84\%$ of the summed component norms at every logged checkpoint; round 300 in Table 9). $\mathbf{C}^{\text{scale}}$ collects deviations that stay *inside* the global spaces (re-scalings and re-mixings of the r global directions, e.g., a 393K-sample MNLI client vs. a 2.5K-sample RTE client) and stays small ($\lesssim 2\%$) because the LoRA scale α/r normalizes update magnitudes. $\mathbf{C}^{\text{cross}}$ is their interaction. The dominance of $\mathbf{C}^{\text{sub}}$ is the key to why the error can be harmless: it reflects clients exploring *different* task subspaces, not a systematic bias in a shared direction. Removing $\mathbf{C}^{\text{sub}}$ by freezing a factor removes exactly the cross-task adaptation capacity that produced it (Appendix I).

3.4 Taxonomy of existing methods

The framework gives a unified lens for the global-model families of Section 2 (Table 5, Appendix A): each method either restricts the update space, weakens gradient utilization, re-factorizes the global product, or controls error post-hoc. None combines full-rank factor updates with per-step, structure-aware control of the error itself, which an *instrument* for varying it requires.

4 An instrument to vary aggregation error

Existing methods conflate two effects: they reduce aggregation error while also restricting the update space, weakening gradient utilization, or altering capacity. To test whether the error itself harms accuracy, we vary its magnitude while holding other design choices fixed, via a per-client regularizer of strength λ that dials error from FedAvg ($\lambda{=}0$) toward zero ($\lambda{\to}\infty$). *This is not a deployable method*; we label rows using it as **FedDecoRA** with base strength λ_0.

4.1 Construction: dispersion-product penalty

Proposition 2 gives $\|\mathbf{C}_\omega\|_F \le \sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B})$, so the dispersion product is the natural quantity to penalize. The per-client regularizer is the dispersion-product (CP) form:

$$\mathcal{L}_k^{\text{CP}}(\mathbf{A}_k, \mathbf{B}_k) = \left\|\mathbf{A}_k - \bar{\mathbf{A}}^t\right\|_F^2 \cdot \left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2, \tag{6}$$

where $(\bar{\mathbf{A}}^t, \bar{\mathbf{B}}^t)$ are the factors broadcast at the start of round t, giving local objective $\mathcal{L}_k + \lambda^t \mathcal{L}_k^{\text{CP}}$. The gradient $\nabla_{\mathbf{A}_k}\mathcal{L}_k^{\text{CP}} = 2(\mathbf{A}_k - \bar{\mathbf{A}}^t)\left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2$ scales the restoring force on $\mathbf{A}_k$ by $\mathbf{B}_k$'s drift, so asymmetric factor drift is allowed (a large $\delta\mathbf{A}$ is compatible with a small $\delta\mathbf{B}$). Proposition 2 measures dispersion about the *post-round* means, which a client cannot know during local training, so Eq. (6) anchors at the broadcast factors; the per-client product is a computable surrogate for the dispersion product, not an upper bound on $\|\mathbf{C}_\omega\|_F^2$ (Appendix G.1). The instrument's validity therefore rests on the empirical response of the error to λ: error decreases with λ on RoBERTa-base (Table 10); on DeBERTa-v3-large the single-seed sweep's error is nearly flat (Table 2), with small non-monotonicities consistent with seed-level noise.

4.2 Output-space form: an alternative parameterization

A second parameterization targets the error directly rather than the dispersion product that bounds it:

$$\mathcal{L}_k^{\text{OS}}(\mathbf{A}_k, \mathbf{B}_k) = \left\|\mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t\right\|_F^2, \tag{7}$$

with gradient $2(\mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t)\mathbf{B}_k^\top$. It penalizes the product residual that Proposition 1 identifies as the error, and it yields the same qualitative conclusion as the CP form. By the parallel-axis identity ($\mathbf{M}_k = \mathbf{A}_k\mathbf{B}_k$, $\boldsymbol{\mu} = \sum_k \omega_k \mathbf{M}_k$), $\sum_k \omega_k \left\|\mathbf{M}_k - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t\right\|_F^2 = \sum_k \omega_k \|\mathbf{M}_k - \boldsymbol{\mu}\|_F^2 +$

$\left\|\boldsymbol{\mu} - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t\right\|_F^2$, so the aggregated OS objective also contains an across-client product-variance term (and the last term includes the global product update). Neither form bounds the error; both are surrogates whose effect on it we verify empirically. We use the CP form for the multi-seed RoBERTa-base evaluation (Table 1) and the OS form for the DeBERTa-v3-large λ sweep (Table 2); Appendix D.1 maps every table to its form.

4.3 Adaptive schedule, complexity, and what else λ changes

The server updates λ from observed dispersions,

$$\lambda^{t+1} = \lambda_0 \cdot \min\left(1,\ \hat{\sigma}_\omega^t(\mathbf{A})\hat{\sigma}_\omega^t(\mathbf{B})/\epsilon_{\text{target}}\right), \tag{8}$$

capped at λ_0, where $\hat{\sigma}_\omega^t$ are the dispersions of the uploaded factors and ϵ_{target} a target dispersion product (pseudocode in Appendix C). Communication is identical to FedAvg+LoRA ($2dr$ per projection); compute adds $\sim$6% wall-clock at P=200. Raising λ also changes the effective smoothness $L + O(\lambda R^2)$, the local trajectory and its stationary points, and (for the OS form) the product-variance term above, so the instrument is not a perfectly clean intervention; Section 5 confirms its conclusion on four regularization-free axes.

4.4 Supporting analysis: convergence under per-step control

This analysis relies on additional assumptions and only gives intuition for the per-step design; the empirical claim does *not* rest on it. We vectorize $\theta = \mathrm{vec}(\mathbf{A}, \mathbf{B})$, write $\bar{\theta}^t$ for the aggregated iterate, η for the local step size, $\mathcal{L} = \sum_k \omega_k \mathcal{L}_k$, $\mathcal{L}^* = \min_\theta \mathcal{L}(\theta)$ over the parameter space a method can reach, and $\Delta_0 = \mathcal{L}(\bar{\theta}^0) - \mathcal{L}^*$. We assume **(A1)** each $\mathcal{L}_k$ is L-smooth (the regularized local objective $F_k = \mathcal{L}_k + \lambda\mathcal{L}_k^{\text{CP}}$ is then $L+O(\lambda R^2)$-smooth by (A4), absorbed into L); **(A2)** stochastic gradients are unbiased with variance $\le \sigma^2$; **(A3)** $\sum_k \omega_k \left\|\nabla\mathcal{L}_k - \nabla\mathcal{L}\right\|_F^2 \le G^2$; **(A4)** $\left\|\mathbf{A}_k\right\|_F, \left\|\mathbf{B}_k\right\|_F \le R$ throughout training.

Theorem 3 (Prevention vs. correction; informal)**.** *Under (A1)–(A4), the near-stationarity step and balanced-drift condition (H) of Lemma 5, and $\eta \le \min(1/(2LP), 1/\sqrt{T})$, the average squared gradient over T rounds satisfies*

$$\frac{1}{T}\sum_t \mathbb{E}\left\|\nabla\mathcal{L}(\bar{\theta}^t)\right\|_F^2 \ \le\ \frac{2\Delta_0}{\eta PT} + \frac{L\eta\sigma^2}{K} + L^2\eta^2P^2G^2 + \mathcal{E}_{agg}, \tag{9}$$

where (a) for unconstrained FedAvg+LoRA the aggregation term is $\mathcal{E}_{agg} = O(L\,P^4\eta^4G^4)$ (each factor drifts $O(P\eta G)$, so $\left\|\mathbf{C}_\omega\right\|_F = O(P^2\eta^2G^2)$ by Proposition 2), whereas (b) for the CP instrument with $\lambda^t \ge \lambda_{\min}$ it is $\mathcal{E}_{agg} = LC_1G^{4/3}/\lambda_{\min}^{4/3}$ for a constant C_1, independent of P in its leading term. Factor-freezing sets $\mathcal{E}_{agg} = 0$ but restricts $\mathcal{L}^*$.

The $\lambda^{-4/3}$ rate rests on a heuristic near-stationarity step and a balanced-drift condition (Lemma 5; full statement and limits in Appendix B). The headline finding relies only on Propositions 1–2 and the controlled experiments that follow.

5 Experiments

We ask: (Q1) Does aggregation error harm accuracy, as prior work assumes? (Q2) Does the answer survive on axes that involve no regularization (exact error-free aggregation, local steps P, client count K, and rank r)? (Q3) Can error be controlled without the expressiveness loss of factor-freezing?

5.1 Experimental setup

Model and LoRA. RoBERTa-base (Liu et al., 2019) (d=768, 12 layers); LoRA on query/value projections (24 modules), $\mathbf{A} \in \mathbb{R}^{768\times8}$, $\mathbf{B} \in \mathbb{R}^{8\times768}$, α/r=2. Backbone frozen. **Federation.** Cross-task on 6 GLUE (Wang et al., 2019) tasks (SST-2/MNLI/QNLI/RTE/QQP/MRPC), K=40 clients with non-overlapping per-task partitions (strong subspace heterogeneity). Classification heads are task-specific and kept fully local; only the LoRA factors are communicated and averaged (Appendix D).

Table 1: Cross-task federated GLUE (RoBERTa-base, K=40, T=300, P=200, r=8). Mean ± std over 3 seeds. **Bold** = best; underline = second best. The *FedDecoRA* row uses the CP form (Eq. 6).

	SST-2	MNLI	QNLI	QQP	RTE	MRPC	Overall	Agg. Err.
Factor-freezing (zero error, reduced expressiveness):								
FFA-LoRA	89.9±.4	79.8±.4	84.1±.6	84.8±.2	72.5±1.9	76.6±3.4	81.3±.3	0.000
RoLoRA	91.8±.1	83.5±.1	87.6±.1	86.7±.4	74.7±1.5	77.4±2.4	83.6±.3	0.000
General-purpose drift control:								
FedProx	92.2±.2	84.0±.2	87.9±.3	87.3±.2	73.2±1.6	**79.5**±1.4	84.0±.2	0.001
Unconstrained / post-hoc correction:								
FedAvg	92.6±.2	84.5±.1	**88.6**±.5	**87.6**±.1	75.1±.3	78.3±1.1	84.5±.2	0.036
LS-Corr	**92.8**±.2	**84.7**±.1	88.5±.9	**87.6**±.2	**76.3**±.3	79.4±2.1	**84.9**±.3	0.036
Covariance-targeted control (instrument, CP form):								
FedDecoRA	92.2±.3	84.2±.2	88.0±.3	87.1±.2	75.6±.4	79.4±1.7	84.4±.3	0.022

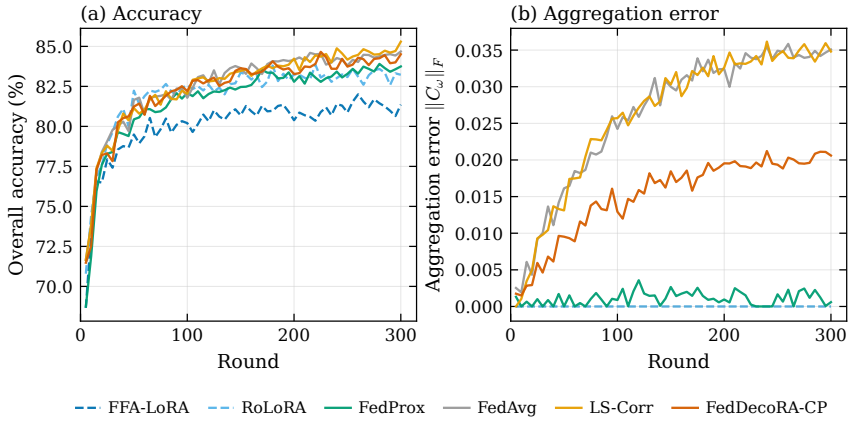


Figure 1: Training dynamics, 300 rounds, seed 42. **(a)** FedDecoRA tracks FedAvg in accuracy; factor-freezing methods plateau lower. **(b)** FedDecoRA keeps error markedly lower throughout (38% on the 3-seed mean, Table 1); FFA-LoRA/RoLoRA have error $\equiv 0$.

Training. AdamW, η=5×10^{-4}, P=200, batch 16, T=300, 30% participation, fp32; same client selection sequence across methods. **Baselines.** FedAvg+LoRA, FFA-LoRA (Sun et al., 2024) ($\mathbf{A}$ frozen), RoLoRA (Chen et al., 2025) (alternating freeze), LS-Corr (a rank-r least-squares residual correction, $\bar{\mathbf{B}} \leftarrow \bar{\mathbf{B}} + \Delta\mathbf{B}$ with $\bar{\mathbf{A}}\Delta\mathbf{B} \approx \mathbf{C}_\omega$; Appendix D), FedProx (Li et al., 2020) (μ=0.01); the exact aggregator FlexLoRA (Bai et al., 2024) is evaluated in a matched-pipeline comparison (Sec. 5.4). **FedDecoRA (instrument):** λ_0=0.05, ϵ_{target}=0.01. **Metrics.** Overall accuracy is the unweighted mean of per-task accuracies; aggregation error $\|\mathbf{C}_\omega\|_F$ is averaged over all 24 projections. Means ± std over seeds $\{42, 123, 456\}$ unless stated. Full details and a table-to-configuration map: Appendix D.

5.2 Main results (Q1: Does error harm accuracy?)

Table 1 presents the main results. Four findings emerge.

(F1) Factor-freezing is counterproductive at r=8. FFA-LoRA and RoLoRA achieve $\|\mathbf{C}_\omega\|_F = 0$ yet lose 3.2% and 0.9% vs. FedAvg (Fig. 2(a); FFA-LoRA drops 4.7% on MNLI and 2.6% on RTE).

(F2) Method ranking correlates with error, but the correlation is structural. Across 15 method×seed pairs, accuracy correlates positively with error (Pearson r=+0.78, p<0.001): unconstrained methods sit at high error and high accuracy, freezing methods at zero error and lower accuracy. This reflects the comparison design, in which the constraints chosen to reduce error are themselves harmful, not evidence that error helps (Appendix H.5).

(F3) Additive drift control is blunt. FedProx reduces error to 0.001 at −0.5% accuracy. Its penalty $\|\delta\mathbf{A}\|_F^2 + \|\delta\mathbf{B}\|_F^2$ vanishes only at the single point $\delta\mathbf{A} = \delta\mathbf{B} = 0$, so all drift is suppressed uniformly,

Table 2: Cross-task federated GLUE on DeBERTa-v3-large (d=1024, K=40, T=200, P=200; OS form). FedAvg, FFA-LoRA, and FedDecoRA at $\lambda_0 \in \{0.005, 0.01\}$ aggregate over 3 seeds; the remaining λ_0 values use seed 42. Final accuracy varies by 0.33% across the single-seed sweep (85.33–85.66%); multi-seed FedDecoRA at λ_0=0.01 matches FedAvg within 0.01%.[1]

Method	λ_0	Final acc.	Best acc.	$\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B})$	Agg. Err.
FedAvg (3 seeds)	–	86.57±1.19	87.08±0.98	0.0095	0.0298
FFA-LoRA (3 seeds)	–	**87.06±1.01**	**87.13±0.94**	0.0000	0.0169
FedDecoRA instrument, multi-seed (3 seeds):					
FedDecoRA	0.005	86.22±0.55	–	**0.0039**	0.0229
FedDecoRA	0.01	86.56±0.70	–	0.0063	0.0279
FedDecoRA instrument, λ_0 sweep (seed 42):					
λ_0=0.001	0.001	85.60%	85.70%	0.0089	0.0296
λ_0=0.05	0.05	85.63%	85.98%	0.0075	0.0289
λ_0=0.2	0.2	85.49%	85.90%	0.0078	0.0297
λ_0=0.5	0.5	85.54%	85.59%	0.0080	0.0297
λ_0=1.0	1.0	85.33%	85.40%	0.0084	0.0298
λ_0=2.0	2.0	85.66%	85.81%	0.0086	0.0296

whereas the CP penalty $\|\delta\mathbf{A}\|_F^2 \|\delta\mathbf{B}\|_F^2$ vanishes on the entire union $\{\delta\mathbf{A}{=}0\} \cup \{\delta\mathbf{B}{=}0\}$; when every client adapts the same single factor, such configurations contribute zero error by Proposition 1. FedProx thus over-penalizes error-free single-factor adaptation (Appendix I).

(F4) The instrument decouples error from expressiveness. At λ_0=0.05, it achieves 38% error reduction at no measurable accuracy cost (84.4±.3% vs. FedAvg's 84.5±.2%; mean difference −0.04%). A TOST equivalence test (Δ=0.5%, α=0.05, Welch's t, df=4) gives $\max p$=0.038, establishing equivalence rather than failure-to-reject. The λ knob varies error without inducing accuracy change (Fig. 1).

5.3 DeBERTa-v3-large: λ sweep at scale (Q1 expanded)

Are the RoBERTa findings an artifact of one model scale? We sweep λ_0 over more than three orders of magnitude on DeBERTa-v3-large (He et al., 2023) (d=1024, 24 layers) with FedAvg and FFA-LoRA references; the full sweep is single-seed for compute reasons, with 3-seed results for $\lambda_0 \in \{0.005, 0.01\}$ and both references.

Across both models, error spans 0.000 to 0.036 while final accuracy varies by under 1.5% among non-freezing methods. On DeBERTa, accuracy varies by 0.33% across the single-seed sweep, *multi-seed FedDecoRA at λ_0=0.01 (86.56±0.70%) matches multi-seed FedAvg (86.57±1.19%)*, and the multi-seed FFA-LoRA gap (+0.49%) is within seed-level standard deviation: factor-freezing becomes competitive, but not strictly winning, at scale.

5.4 Exact error-free aggregation: FlexLoRA (Q2)

The cleanest non-regularization test is a method that removes the error *by construction* without changing the local objective. FlexLoRA averages the products $\mathbf{A}_k\mathbf{B}_k$ and re-factorizes the mean by truncated SVD, so the factor-averaging residual of Proposition 1 never arises (only the truncation residual remains). We compare it against FedAvg under an identical pipeline (Table 3; bf16, η=10^{-4}, 40% participation, so absolute values are not comparable to Table 1; the comparison is strictly within the matched pair, with paired client selection per seed).

Three findings. **(i) Truncation is negligible.** The residual $\left\|\bar{\mathbf{M}} - \mathrm{SVD}_r(\bar{\mathbf{M}})\right\|_F$ stays ≤ 0.001 for all 300 rounds, so re-factorization is essentially lossless at r=8 here. (Unlike LS-Corr, which must represent $\mathbf{C}_\omega$ inside $\mathrm{col}(\bar{\mathbf{A}})$, the SVD selects the optimal rank-r subspace.) **(ii) FlexLoRA gains** +1.49±0.27**pp**, sign-consistent across seeds. **(iii) The gain is not attributable to error removal.** FlexLoRA changes two things at once, the error (to zero) and the parameterization: the per-round

[1]The FFA-LoRA Agg. Err. 0.0169 is a snapshot-estimator artifact: the theoretical $\mathbf{C}_\omega$ of Proposition 1 is exactly 0 for FFA-LoRA (Appendix H.3).

Table 3: FlexLoRA vs. FedAvg under an identical pipeline (RoBERTa-base, T=300, P=200, K=40, r=8; 3 seeds). Per-seed paired gaps: +1.28, +1.79, +1.41pp (mean +1.49±0.27). †Snapshot-estimator artifact (Appendix H.3); the theoretical $\mathbf{C}_\omega$ is exactly 0.

Method	Acc. (%)	Agg. Err.	SVD trunc. residual	$\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B})$
FedAvg (matched pipeline)	81.49±0.19	0.0319	–	0.0077
FlexLoRA	**82.99±0.42**	0 (0.0019†)	≤ 0.001	0.599

Table 4: Accuracy (%) and aggregation error vs. P on RoBERTa-base; all rows T=100, seed 42. FedAvg's error grows ∼190× while accuracy increases monotonically; the instrument keeps error at or below FedAvg's at every P.

	FedAvg+LoRA		FedDecoRA (instrument)				
P	Acc.	Error	Acc.	Error	λ_0	Err. ↓	ΔAcc.
5	68.9	0.0005	67.7	0.0004	0.05	−20%	−1.2
50	80.0	0.011	80.2	0.010	0.05	−6%	+0.2
200	81.8	0.037	82.2	0.025	0.05	−33%	+0.4
500	82.4	0.096	82.3	0.004	0.01	−96%	−0.1

SVD re-balances ($\|\mathbf{a}_i\|=\|\mathbf{b}_i\|=\sqrt{s_i}$) and rotates the factors, and bilinear gradient descent is not factorization-invariant (Du et al., 2018; Arora et al., 2019). The dispersion product records this intervention (0.599 vs. 0.0077): large yet decorrelated deviations at zero error. Placed beside Table 1, zero-error methods span −3.2% (FFA-LoRA) and −0.9% (RoLoRA) to +1.5% (FlexLoRA) relative to their matched FedAvg baselines, while LS-Corr gains +0.4% at *unchanged* error: the error level predicts neither the sign nor the size of these differences; the accompanying mechanism does.

5.5 Scaling with local steps P and client count K (Q2)

Local steps. We vary $P \in \{5, 50, 200, 500\}$ at a fixed round budget T=100 on RoBERTa-base (Table 4, Fig. 2(b)). FedAvg's error grows ∼190× from P=5 to P=500 while its accuracy *rises monotonically* (68.9 → 80.0 → 81.8 → 82.4); the instrument's error reduction grows with P (−33% at P=200, −96% at P=500) at matching accuracy, consistent with Theorem 3. λ_0 must be scaled manually with P ($\lambda_0 \propto 1/P$); Eq. (8) is a within-round cap, not automatic P-scaling.

Client count. The framework locates K's influence in the dispersions: at a fixed task mix, more clients thin per-client shards and raise deviations, while the covariance structure stays task-driven. A K-sweep (Table 6, Appendix D; fixed 6-task mix, T=100, seed 42, matched pipeline of Table 3) confirms this: error grows monotonically with K (0.0129 → 0.0165 → 0.0322) as dispersions rise (0.0019 → 0.0111), while accuracy stays flat (79.6–80.5%).

5.6 Ablations: covariance decomposition, λ_0, rank

Full tables are in Appendix H.1. **(i)** $\left\|\mathbf{C}^{\text{sub}}\right\|_F$ accounts for $\gtrsim$84% of the summed component norms (scale $\lesssim$2%, cross ∼14%); LS-Corr leaves it essentially untouched. **(ii)** At P=200, $\lambda_0 \in \{0.01, 0.05\}$ reduce error by 6–33% at zero accuracy cost; λ_0=0.10 halves error at −0.5%. **(iii)** Doubling r to 16 cuts error ∼15× *without any error control* and narrows FFA-LoRA's gap from −3.3% to −0.9%: the rank-r constraint, not the error, is the binding bottleneck.

6 Discussion

Consilience across six axes. The central claim rests on agreement across structurally independent ways of varying error: (a) CP-form λ (Table 1); (b) OS-form λ (Table 2); (c) local steps P (error grows ∼190×, Table 4); (d) rank r (error shrinks ∼15×, Table 11); (e) client count K (Table 6); and (f) exact error-free aggregation (Table 3). Axes (c)–(f) involve no regularization, so the conclusion is not an artifact of the instrument's modified objective. Operationally, "benign" means: within the evaluated settings, varying $\|\mathbf{C}_\omega\|_F$ by >30× along a controlled axis changes accuracy by <1.5%

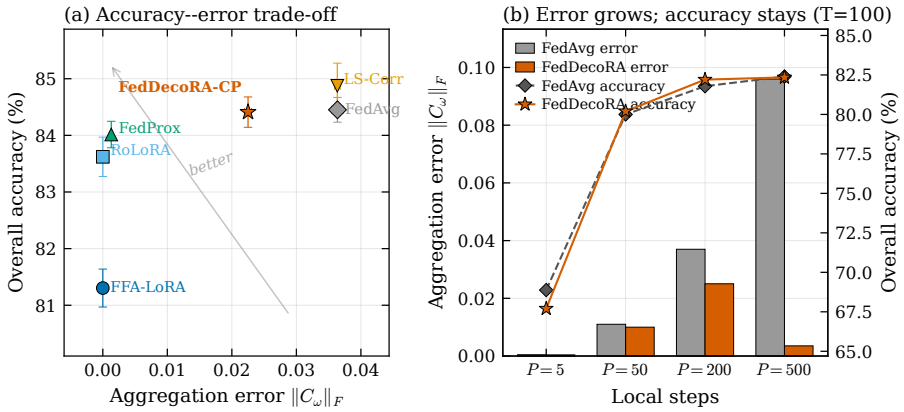


Figure 2: **(a)** Accuracy vs. aggregation error across methods (3-seed means, RoBERTa-base, Table 1). Factor-freezing methods achieve zero error at substantial accuracy cost; FedDecoRA lies on the Pareto frontier. **(b)** Error vs. local steps P at $T{=}100$ (Table 4): FedAvg's error grows $\sim$190$\times$; the instrument keeps it low.

among non-freezing methods, with the instrument-vs-FedAvg comparison TOST-equivalent to zero ($\Delta{=}0.5\%$, $\alpha{=}0.05$).

Why can FedAvg tolerate the error? Convergence bounds cannot answer this: an upper bound shows that small error suffices for the standard rate, not that large error is harmless. Two mechanisms are plausible (Appendix I): the per-round perturbation is $\sim 10^{-4}$ of the accumulated model and is partially corrected by subsequent descent, and covariance terms across projections may cancel. LS-Corr's $+0.4\%$ at unchanged error is within seed noise (Appendix I).

Scope, and where the conclusion may fail. Our evidence covers two encoder models and NLU classification on GLUE. We expect the conclusion to be tested most sharply in decoder-only LLMs, generative objectives, personalization-heavy FL where client optima genuinely diverge, and cross-device deployments. Two falsification conditions follow from the framework: very small r/d, where both the error and the freezing cost are largest (the $r{=}8{\to}16$ ablation shows both shrink as r/d grows); and any setting in which $\mathbf{C}^{\text{scale}}$ rather than $\mathbf{C}^{\text{sub}}$ dominates, which would undercut the subspace interpretation.

Limitations. (1) The DeBERTa λ sweep and the K-sweep are single-seed except where noted, and the FlexLoRA and K-sweep comparisons use a separate matched pipeline. (2) Encoder models and NLU classification only; no decoder or generative evaluation. (3) The instrument is not a deployable method. (4) The convergence analysis is supporting intuition with a heuristic step. (5) Among exact aggregators only FlexLoRA is evaluated.

7 Conclusion

We revisit the assumption that aggregation inconsistency in federated LoRA is harmful. The error is the weighted cross-factor covariance, bounded by the dispersion product and dominated by subspace misalignment. A controlled instrument varies it along one axis, and four regularization-free axes confirm the verdict: in the evaluated settings, error varies by $>30\times$ while accuracy among non-freezing methods varies by $<1.5\%$, and zero-error methods range from -3.2% to $+1.5\%$ relative to matched baselines, so the error level predicts neither the sign nor the size of accuracy differences. The cost of aggregation inconsistency lies in the constraints imposed to eliminate it, not in the error itself; whether this extends beyond encoder models and NLU classification is an open question, for which we state explicit falsification conditions.

AI USE STATEMENT

The authors conducted the entire research process themselves, including determining the research questions, deciding which hypotheses to pursue and which experiments to conduct, and interpreting the results. In this process, we used LLM tools as assistants to support theoretical analysis, coordinate experiments, and detect errors. The authors verified all proofs, implementations, and references and take full responsibility for the content.

REPRODUCIBILITY STATEMENT

Propositions 1–2 are proved in Appendix A; the assumptions and proof sketch of the supporting convergence analysis are in Appendix B, with its heuristic step stated explicitly. Appendix D gives the complete experimental setup (data partitions, model and LoRA configuration, optimizer, precision, participation, the local classification-head protocol, and baselines), and Appendix D.1 maps every table and figure to its instrument form, seeds, and (T, P, K) configuration, including the separate matched pipeline used for Tables 3–6. Pseudocode is in Appendix C; the aggregation-error estimator is defined in Appendix H.3. Anonymized source code with per-table run commands is provided as supplementary material.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations (ICLR)*, 2021.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory (COLT)*, pp. 530–582, 2016.

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 (ICCV)*, pp. 3737–3746, 2025.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 (ICLR)*, 2025.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning (ICML)*, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023.

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, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14 (1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, 2020.

Bumjun Kim and Wan Choi. Two-stage wireless federated LoRA fine-tuning with sparsified orthogonal updates. *arXiv preprint arXiv:2505.00333*, 2025.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki Markus Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations (ICLR)*, 2024.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10713–10722, 2021.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-IID data silos: An experimental study. In *IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978, 2022.

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

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations (ICLR)*, 2020.

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 *International Conference on Machine Learning (ICML)*, 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.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Adithya Renduchintala, Tugrul Konuk, and Oleksii Kuchaiev. Tied-LoRA: Enhancing parameter efficiency of LoRA with weight tying. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336, 2025.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. FedRot-LoRA: Mitigating rotational misalignment in federated LoRA. *arXiv preprint arXiv:2602.23638*, 2026.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the FederatedGPT: Federated instruction tuning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919, 2024.

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

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*, 2018.

Weiming Zhuang, Chen Chen, Jingtao Li, Chaochao Chen, Yaochu Jin, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.

A PROOFS OF THE EXACT RESULTS

Proof of Proposition 1. Write $\mathbf{A}_k = \bar{\mathbf{A}} + \delta\mathbf{A}_k$, $\mathbf{B}_k = \bar{\mathbf{B}} + \delta\mathbf{B}_k$ with $\sum_k \omega_k \delta\mathbf{A}_k = \sum_k \omega_k \delta\mathbf{B}_k = 0$. Then:

$$\begin{aligned}\textstyle\sum_k \omega_k \mathbf{A}_k\mathbf{B}_k &= \textstyle\sum_k \omega_k (\bar{\mathbf{A}} + \delta\mathbf{A}_k)(\bar{\mathbf{B}} + \delta\mathbf{B}_k) \\ &= \bar{\mathbf{A}}\bar{\mathbf{B}} + \bar{\mathbf{A}} \underbrace{\textstyle\sum_k \omega_k \delta\mathbf{B}_k}_{=\mathbf{0}} + \underbrace{\textstyle\sum_k \omega_k \delta\mathbf{A}_k}_{=\mathbf{0}} \bar{\mathbf{B}} + \textstyle\sum_k \omega_k \delta\mathbf{A}_k \delta\mathbf{B}_k.\end{aligned}$$

Subtracting $\bar{\mathbf{A}}\bar{\mathbf{B}}$ gives $\mathbf{C}_\omega = \sum_k \omega_k \delta\mathbf{A}_k \delta\mathbf{B}_k$, where $\mathbf{C}_\omega \in \mathbb{R}^{d\times d}$. □

Proof of Proposition 2. By the triangle inequality and submultiplicativity of the Frobenius norm: $\|\mathbf{C}_\omega\|_F = \|\sum_k \omega_k \delta\mathbf{A}_k \delta\mathbf{B}_k\|_F \le \sum_k \omega_k \|\delta\mathbf{A}_k\|_F \|\delta\mathbf{B}_k\|_F$. Applying weighted Cauchy–Schwarz with $x_k = \|\delta\mathbf{A}_k\|_F$, $y_k = \|\delta\mathbf{B}_k\|_F$:

$$\textstyle\sum_k \omega_k x_k y_k \le \sqrt{\sum_k \omega_k x_k^2} \cdot \sqrt{\sum_k \omega_k y_k^2} = \sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B}).$$ □

Corollary 4 (Prior methods as special cases)**.** *If $\mathbf{A}_k = \mathbf{A}$ for all k, then $\sigma_\omega(\mathbf{A}) = 0$ and $\mathbf{C}_\omega = 0$, recovering FFA-LoRA. If $\mathbf{A}_k = \mathbf{A}$ on even rounds and $\mathbf{B}_k = \mathbf{B}$ on odd rounds, each round has $\mathbf{C}_\omega = 0$, recovering RoLoRA. Both follow immediately from Proposition 2.*

Table 5: Structural comparison of federated LoRA methods through the covariance lens. The instrument used to vary error magnitude (Sec. 4) is shown in the bottom row.

Method	Update space	Grad. util.	Capacity	Error control
FFA-LoRA	rank-r (fixed)	50%	all to $\mathbf{B}$	Exact
RoLoRA	rank-r (alt.)	50%	50/50 fixed	Exact/round
FedProx	full rank	100%	adaptive	Indirect (additive)
LS-Corr (rank-r)	full rank	100%	adaptive	Post-hoc
FedEx-LoRA, FLoRA	full rank	100%	adaptive	Exact
FlexLoRA	full rank (SVD re-fact.)	100%	adaptive	Exact (product)
FedDecoRA (instrument)	**full rank**	**100%**	**adaptive**	**Quantity, per-step**

A.1 PROOF OF THEOREM 3

Sketch. Part (a): In unconstrained FedAvg+LoRA, both factors drift during P local steps. Each factor moves $O(\eta G)$ per step, accumulating $O(P\eta G)$ deviation from the broadcast parameters. By Proposition 2, the aggregation error is bounded by $\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B}) = O(P\eta G)^2 = O(P^2\eta^2 G^2)$, so its square, which enters the bound, is $O(P^4\eta^4 G^4)$. Note that factor-freezing (FFA-LoRA, RoLoRA) eliminates this term by setting $\sigma_\omega(\mathbf{A}) = 0$, but at the cost of restricting the achievable optimum $\mathcal{L}^*$.

Part (b): With FedDecoRA, the regularizer gradient $2\lambda(\mathbf{A}_k - \bar{\mathbf{A}}^t)\left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2$ acts as a restoring force at every step. At a near-stationary point of the regularized objective, the dispersion product satisfies $\|\delta\mathbf{A}_k\|_F \|\delta\mathbf{B}_k\|_F \lesssim (G/\lambda)^{2/3}$ (balancing task gradient against regularizer gradient in both stationarity conditions; Lemma 5). Since this holds per-step, under the balanced-drift condition (H) of Lemma 5, the squared aggregation error is bounded by $O(\lambda^{-4/3})$, independent of P in its leading term. The full proof follows the framework of Karimireddy et al. (2020) with bilinear covariance terms.

B SUPPORTING CONVERGENCE ANALYSIS (CP FORM)

Status of this appendix. The results here are *supporting analysis*: they rely on assumptions (A1)–(A4) and on a heuristic near-stationarity step (Lemma 5), and no empirical claim in the paper depends on them. They are included to explain why per-step control yields an aggregation term whose leading part is independent of P (under condition (H) below), which motivated the instrument's design. We provide a convergence bound for the dispersion-product (CP) form of the instrument

applied as the local regularizer (Eq. 6, $\mathcal{L}_k^{\text{CP}} = \left\|\mathbf{A}_k - \bar{\mathbf{A}}^t\right\|_F^2 \cdot \left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2$ with broadcast anchors), analogous to standard federated learning rates (Li et al., 2020; Karimireddy et al., 2020; Yang et al., 2021). We analyze the CP form because its penalty is a direct surrogate for the dispersion product of Proposition 2; the OS form (Eq. 7) carries an additional across-client product-variance term (Appendix G.1). Neither aggregated penalty upper-bounds $\|\mathbf{C}_\omega\|_F^2$ in general, because the penalties are anchored at the broadcast factors while $\mathbf{C}_\omega$ is defined about the post-round means; the lemma below therefore carries an explicit balanced-drift condition. The CP form is the empirical instrument of Table 1, the OS form that of Table 2; a formal analysis for the OS form is left as future work.

The gradient used in this analysis is therefore $\nabla_{\mathbf{A}_k}\mathcal{L}_k^{\text{CP}} = 2(\mathbf{A}_k - \bar{\mathbf{A}}^t)\left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2$, with the symmetric expression for $\mathbf{B}_k$. For clarity we vectorize LoRA factors and write $\theta_k = \text{vec}(\mathbf{A}_k, \mathbf{B}_k) \in \mathbb{R}^{2dr}$, with global mean $\bar{\theta}^t = \sum_k \omega_k \theta_k^t$. The local objective is $F_k(\theta) = \mathcal{L}_k(\theta) + \lambda \cdot \mathcal{L}_k^{\text{CP}}(\theta)$ and the global objective is $F(\theta) = \sum_k \omega_k F_k(\theta)$.

B.1 Assumptions

We adopt standard assumptions from federated non-convex analysis.

Definition 2 (Assumptions). ***(A1) L-smoothness:*** *Each $\mathcal{L}_k$ is L-smooth: $\|\nabla\mathcal{L}_k(\theta) - \nabla\mathcal{L}_k(\theta')\|_F \le L\|\theta - \theta'\|_F$. Under (A4), $F_k = \mathcal{L}_k + \lambda\mathcal{L}_k^{CP}$ is L_{eff}-smooth on the feasible set with $L_{eff} = L + O(\lambda R^2)$; we write L for L_{eff} below.*

(A2) Unbiased stochastic gradients with bounded variance: *The stochastic gradient $g_k(\theta) = \nabla F_k(\theta; \xi_k)$ satisfies $\mathbb{E}[g_k(\theta)] = \nabla F_k(\theta)$ and $\mathbb{E}\|g_k(\theta) - \nabla F_k(\theta)\|_F^2 \le \sigma^2$.*

(A3) Bounded heterogeneity: *$\|\nabla\mathcal{L}_k(\theta) - \nabla\mathcal{L}(\theta)\|_F^2 \le G^2$ for all k and θ, where $\mathcal{L} = \sum_k \omega_k \mathcal{L}_k$ (a per-client form of the usual averaged bound).*

(A4) Bounded LoRA factors: *There exists $R > 0$ such that $\|\mathbf{A}_k\|_F, \|\mathbf{B}_k\|_F \le R$ throughout training.*

Assumption (A4) is standard for LoRA (Hu et al., 2022; Hayou et al., 2024); it is not implied by the CP regularizer (which vanishes whenever one factor sits at its anchor), so we assume it. The smoothness constant L depends on the pretrained model and rank r; for RoBERTa-base with r=8, empirical estimation gives $L \approx 2$–5.

B.2 Key lemma: aggregation error bound

The core technical step is bounding the aggregation error in terms of the regularizer strength.

Lemma 5 (Aggregation error under FedDecoRA, heuristic form). *Under (A1)–(A4), assuming local SGD converges sufficiently close to a stationary point of F_k within P steps, and under the balanced-drift condition **(H)** $\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B}) \le c_H \max_k \left\|\mathbf{A}_k - \bar{\mathbf{A}}^t\right\|_F \left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F$ for a constant c_H (clients do not drift in complementary factors), for any round t with regularization strength $\lambda^t \ge \lambda_{\min} > 0$:*

$$\mathbb{E}\left\|\mathbf{C}_\omega^t\right\|_F^2 \lesssim \frac{C_1 G^{4/3}}{\lambda_{\min}^{4/3}} + \frac{C_2 \eta^2 P \sigma^2}{\lambda_{\min}}, \tag{10}$$

where C_1, C_2 are constants depending on L and c_H, and $\lesssim$ denotes inequality up to logarithmic factors. Crucially, the bound is independent of P in its leading term.

Informal derivation. At a (near-)local stationary point of F_k, the gradient conditions give

$$\nabla_{\mathbf{A}_k}\mathcal{L}_k + 2\lambda^t(\mathbf{A}_k - \bar{\mathbf{A}}^t)\left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F^2 \approx 0, \qquad \nabla_{\mathbf{B}_k}\mathcal{L}_k + 2\lambda^t\left\|\mathbf{A}_k - \bar{\mathbf{A}}^t\right\|_F^2(\mathbf{B}_k - \bar{\mathbf{B}}^t) \approx 0. \tag{11}$$

Let $a_k = \left\|\mathbf{A}_k - \bar{\mathbf{A}}^t\right\|_F$, $b_k = \left\|\mathbf{B}_k - \bar{\mathbf{B}}^t\right\|_F$, and $s_k = a_k b_k$. Taking norms of both stationarity conditions and applying (A3):

$$a_k \cdot b_k^2 \lesssim \frac{G + \|\nabla\mathcal{L}\|_F}{2\lambda^t}, \qquad b_k \cdot a_k^2 \lesssim \frac{G + \|\nabla\mathcal{L}\|_F}{2\lambda^t}. \tag{12}$$

Multiplying these and taking the cube root yields $s_k = a_k b_k \lesssim (G/\lambda^t)^{2/3}$, so

$$\mathbb{E}[s_k^2] \;\lesssim\; \frac{C_1' G^{4/3}}{(\lambda^t)^{4/3}} + \frac{\eta^2 P \sigma^2}{\lambda^t}, \tag{13}$$

where the second term reflects finite-time SGD error of order $\eta\sqrt{P}\sigma$ (Stich, 2019) (a stationarity residual ε enters the bound as $\varepsilon^{4/3}/\lambda^{4/3}$ by the same argument; we state it in this simpler form). Proposition 2 and (H) give $\|\mathbf{C}_\omega\|_F \le \sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B}) \le c_H \max_k s_k$. Taking expectations and using $\lambda^t \ge \lambda_{\min}$ yields Eq. 10. Condition (H) is where the broadcast anchor matters: if one client moves only $\mathbf{A}$ and another only $\mathbf{B}$, every $s_k = 0$ while $\mathbf{C}_\omega = -(\bar{\mathbf{A}}^{t+1} - \bar{\mathbf{A}}^t)(\bar{\mathbf{B}}^{t+1} - \bar{\mathbf{B}}^t) \neq 0$; (H) excludes such complementary drift and is one reason the empirical claim does not rest on this analysis. □

On the rigor of Lemma 5. Three subtleties merit explicit discussion. First, non-convex SGD is not guaranteed to reach an exact stationary point in finite steps; the "stationary point" step above is a first-order approximation, justified empirically by the fact that local training converges to near-stationary points in our experiments (loss plateaus within P=200 steps). Second, the leading term is $O(\lambda^{-4/3})$, obtained by combining both stationarity conditions; bounding only the $\mathbf{A}_k$ condition would give a different exponent. Third, the norm inequalities in the stationarity step bound gradient norms from above and are stated up to constants; we make no tightness claim, and a fully rigorous proof using a quadratic surrogate for $\mathcal{L}_k^{\text{CP}}$ remains future work. Empirically, Table 4 is consistent with the prediction that the instrument's error does not grow with P (its error stays within 0.004–0.025 across $P \in \{50, 200, 500\}$ while FedAvg's grows $\sim$190×); monotonicity in λ holds over the range $\lambda_0 \le 0.1$ (Table 10).

This lemma is the formal statement of Theorem 3(b): under (H), the error is bounded, in its leading term, by a quantity depending on $\lambda^{-4/3}$ rather than growing with P.

B.3 Main convergence theorem

We now establish the convergence of FedDecoRA on the global objective.

Theorem 6 (FedDecoRA convergence)**.** *Under (A1)–(A4) and the conditions of Lemma 5, let* $\eta \le \min\left(\frac{1}{2LP}, \frac{1}{\sqrt{T}}\right)$ *and* $\lambda^t = \lambda_{\min} > 0$ *for all* t*. Let* $\mathcal{L}^* = \min_\theta \mathcal{L}(\theta)$*. Then after* T *rounds, FedDecoRA satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla \mathcal{L}(\bar{\theta}^t)\right\|_F^2 \;\le\; \underbrace{\frac{2(\mathcal{L}(\bar{\theta}^0) - \mathcal{L}^*)}{\eta P T}}_{\textit{initialization}} + \underbrace{\frac{L\eta\sigma^2}{K}}_{\textit{stochastic noise}} + \underbrace{L^2\eta^2 P^2 G^2}_{\textit{client drift}} + \underbrace{\frac{LC_1 G^{4/3}}{\lambda_{\min}^{4/3}}}_{\textit{aggregation error}}. \tag{14}$$

With $\eta = 1/\sqrt{T}$ *and appropriately chosen* $\lambda_{\min}$*, the right-hand side is* $O(1/\sqrt{T}) + O(\lambda_{\min}^{-4/3})$.

Proof sketch. Using (A1) (smoothness) and the descent lemma for $\mathcal{L}$ at $\bar{\theta}^t$ (we expand $\mathcal{L}$ rather than the round-dependent F, whose anchors change every round, so that the recursion telescopes):

$$\mathcal{L}(\bar{\theta}^{t+1}) \le \mathcal{L}(\bar{\theta}^t) + \langle \nabla \mathcal{L}(\bar{\theta}^t), \bar{\theta}^{t+1} - \bar{\theta}^t \rangle + \frac{L}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|_F^2. \tag{15}$$

The server update decomposes as

$$\bar{\theta}^{t+1} - \bar{\theta}^t = -\eta P\, \mathbb{E}[g_k(\theta_k^t)] + \underbrace{\Delta\theta_{\text{agg}}^t}_{\text{aggregation error}} + \underbrace{\text{client drift}}_{\text{within-round}} + \underbrace{\text{stochastic noise}}_{\text{mini-batch}}, \tag{16}$$

where $\Delta\theta_{\text{agg}}^t \in \mathbb{R}^{2dr}$ is the factor-space update component induced by the product-space residual $\mathbf{C}_\omega^t \in \mathbb{R}^{d\times d}$.[2] The client drift term is standard (Li et al., 2020; Yang et al., 2021) and yields

[2]The product-space residual $\mathbf{C}_\omega^t \in \mathbb{R}^{d\times d}$ does not directly inhabit the factor-space update space $\mathbb{R}^{2dr}$; it enters through the LoRA reparameterization $\Delta\mathbf{W} = \mathbf{AB}$ via a chain-rule expansion that depends on the specific projection structure. Our analysis uses $\left\|\mathbf{C}_\omega^t\right\|_F$ as a scalar magnitude proxy, which is sufficient to establish the rate Eq. 14 but does not capture the precise factor-space contribution; a chain-rule expansion that resolves the proxy into the true factor-space gradient is deferred to future work.

the $L^2\eta^2P^2G^2$ term. The stochastic noise term gives $L\eta\sigma^2/K$ under partial participation. The aggregation error term, unique to federated LoRA, is bounded by Lemma 5 via the magnitude proxy $\|\mathbf{C}_\omega^t\|_F$.

Telescoping over $t = 0, \ldots, T-1$ and dividing by T gives Eq. 14. □

Note on effective smoothness. As noted in (A1), the local objective $F_k = \mathcal{L}_k + \lambda\mathcal{L}_k^{\text{CP}}$ has an effective smoothness constant $L_{\text{eff}} = L + O(\lambda R^2)$, since the Hessian of $\mathcal{L}_k^{\text{CP}} = \|\delta\mathbf{A}_k\|_F^2\,\|\delta\mathbf{B}_k\|_F^2$ contains terms of order $\|\mathbf{B}_k\|_F^2$ and $\|\mathbf{A}_k\|_F^2$, which are bounded by R^2 under (A4). For small λ (our setting: λ=0.05, $R^2 \lesssim 1$), $L_{\text{eff}} \approx L$ and the theorem holds as stated. For large λ, the constants C_1 in Eq. 14 should be interpreted with L_{eff} in place of L. We omit this refinement in the main statement for clarity.

B.4 Interpretation and comparison with baselines (CP form)

The convergence bound derived above applies to the CP-form instrument. We compare its rate and structural properties to existing federated LoRA baselines to clarify what the analysis says about the broader landscape; this is not a method-paper comparison, since the empirical claim of the paper rests on the controlled experiments (Sec. 5), not on a head-to-head method ranking.

Dependence on P. Theorem 6 exhibits the usual $O(\eta^2P^2G^2)$ client-drift term, matching FedAvg (Li et al., 2020) and improving on round-level methods. Critically, the leading aggregation error term $LC_1G^{4/3}/\lambda_{\min}^{4/3}$ is *independent of* P; this is the key benefit of per-step control.

Comparison with FFA-LoRA. For FFA-LoRA (Assumption (A4) with $\delta\mathbf{A}_k \equiv 0$), the aggregation error vanishes, so Eq. (14) loses the last term. However, the effective parameter dimension is halved (only $\mathbf{B}$ is updated), which alters the trajectory term $\mathcal{L}(\bar{\theta}^0) - \mathcal{L}^*$: the achievable optimum $\mathcal{L}^*$ over the restricted parameter space may be strictly worse than the full-parameter optimum. In practice, this manifests as the -3.2% accuracy gap we observe on RoBERTa-base (Table 1); on DeBERTa-large the gap shrinks to within seed variance ($+0.49\%$ multi-seed, Table 2), consistent with the rank-r/d relationship discussed in Sec. 6.

Comparison with RoLoRA. For RoLoRA, the error vanishes at round boundaries while the within-round drift term is unchanged (only the free factor drifts). The long-run convergence rate is similar, but the effective search space is still constrained to alternating rank-r subspaces.

Comparison with rank-r post-hoc correction (LS-Corr). LS-Corr applies a post-hoc correction $\Delta\mathbf{B}$ satisfying $\bar{\mathbf{A}}\Delta\mathbf{B} \approx \mathbf{C}_\omega$ in least squares. Since $\bar{\mathbf{A}} \in \mathbb{R}^{d\times r}$, the correction $\bar{\mathbf{A}}\Delta\mathbf{B}$ has rank at most r and column space $\text{col}(\bar{\mathbf{A}})$, so it cannot represent a generic $d \times d$ residual. The correction error $\left\|\mathbf{C}_\omega - \bar{\mathbf{A}}\Delta\mathbf{B}\right\|_F$ can be large when $d \gg r$ (our setting: d=768, r=8), leading to effective aggregation error equal to FedAvg's (as we observe: 0.036 in both, Table 1). Exact aggregators such as FedEx-LoRA (residual absorbed into $\mathbf{W}_0$) and FlexLoRA (product-space averaging) instead drive the aggregation term to zero while keeping the full parameter space, as FFA-LoRA does not.

Rate. Setting $\eta = c/\sqrt{T}$ and $\lambda_{\min} = c'T^{3/8}$ (so that $\lambda_{\min}^{-4/3} = O(T^{-1/2})$), the RHS of Eq. 14 becomes $O(1/\sqrt{T})$, matching the standard FL rate (Yang et al., 2021). The constant depends on G, R, σ but *not on P in the leading aggregation error term* (under (H)), providing a theoretical justification for the empirical observation that FedDecoRA scales favorably with P (Table 4).

B.5 Adaptive λ schedule

For the adaptive schedule (Eq. 8), λ^t varies with the observed dispersion. A simple modification of Theorem 6 handles this.

Corollary 7 (Adaptive λ). *If the adaptive schedule maintains $\lambda^t \geq \lambda_{\min} > 0$ for all t, then Theorem 6 holds with the aggregation error term replaced by the time-averaged bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \frac{LC_1 G^{4/3}}{(\lambda^t)^{4/3}}. \tag{17}$$

In particular, if $\lambda^t \to \lambda_\infty > 0$ in the limit (as observed empirically), the effective constant is $LC_1 G^{4/3}/\lambda_\infty^{4/3}$.

In the RoBERTa-base experiments, λ^t stabilized near $\lambda_0 = 0.05$ because the observed dispersion product consistently exceeded ϵ_{target} (on DeBERTa the dispersion product is below ϵ_{target}, so $\lambda^t < \lambda_0$ there); the corollary applies in both regimes.

B.6 Summary

The supporting analysis is consistent with the empirical findings:

1. FedDecoRA achieves the standard $O(1/\sqrt{T})$ non-convex FL rate for a T-dependent choice $\lambda_{\min} \propto T^{3/8}$, matching FedAvg and factor-freezing baselines.
2. Unlike FFA-LoRA and RoLoRA, the leading aggregation error term depends only on $\lambda_{\min}^{-4/3}$, with no P dependence in this term.
3. The trajectory term depends on the expressiveness of the achievable parameter space; factor-freezing methods trade a vanishing aggregation term for a worse $\mathcal{L}^*$.

C Algorithm pseudocode

The full pseudocode for the FedDecoRA instrument is below, in its general form. We instantiate $\mathcal{L}_k^{\text{reg}}$ as either the dispersion-product (CP) form (Eq. 6, used for the multi-seed RoBERTa-base evaluation in Table 1) or the output-space (OS) form (Eq. 7, used for the DeBERTa-large λ sweep in Table 2); the rest of the algorithm is identical in both cases.

Algorithm 1 FedDecoRA instrument (general form: $\mathcal{L}_k^{\text{reg}}$ is either CP, Eq. 6, or OS, Eq. 7). The optimizer step shown is plain SGD for clarity; in our experiments AdamW is used instead.

Require: Global factors $\bar{\mathbf{A}}^0, \bar{\mathbf{B}}^0$; strength λ_0; target ϵ_{target}; steps P; per-step penalty $\mathcal{L}_k^{\text{reg}}$ (CP or OS)

for $t = 0, \ldots, T-1$ **do**
  Sample $\mathcal{S}^t \subseteq [K]$ uniformly at random with $|\mathcal{S}^t| = \rho K$ (participation rate ρ); broadcast $(\bar{\mathbf{A}}^t, \bar{\mathbf{B}}^t, \lambda^t)$
  for each $k \in \mathcal{S}^t$ **in parallel do**
    $\mathbf{A}_k \leftarrow \bar{\mathbf{A}}^t,\ \mathbf{B}_k \leftarrow \bar{\mathbf{B}}^t$
    for $p = 1, \ldots, P$ **do**
      Forward pass on minibatch $\mathcal{B}_p$
      Compute $\mathcal{L}_k^{\text{reg}}$ using anchors $\bar{\mathbf{A}}^t, \bar{\mathbf{B}}^t$ (the round's start-of-round broadcast, as in Eqs. 6–7)
      $\mathbf{A}_k \leftarrow \mathbf{A}_k - \eta\big[\nabla_{\mathbf{A}_k}\mathcal{L}_k + \lambda^t \nabla_{\mathbf{A}_k}\mathcal{L}_k^{\text{reg}}\big]$ (AdamW in practice)
      $\mathbf{B}_k \leftarrow \mathbf{B}_k - \eta\big[\nabla_{\mathbf{B}_k}\mathcal{L}_k + \lambda^t \nabla_{\mathbf{B}_k}\mathcal{L}_k^{\text{reg}}\big]$ (AdamW in practice)
    end for
    Upload $(\mathbf{A}_k, \mathbf{B}_k)$
  end for
  $\bar{\mathbf{A}}^{t+1} \leftarrow \sum_k \omega_k \mathbf{A}_k, \quad \bar{\mathbf{B}}^{t+1} \leftarrow \sum_k \omega_k \mathbf{B}_k$
  Update λ^{t+1} via Eq. (8)
end for

D EXPERIMENTAL DETAILS

Client assignment. 40 clients: SST-2 (8 clients, 67K samples), MNLI (8, 393K), QNLI (8, 105K), RTE (5, 2.5K), QQP (6, 364K), MRPC (5, 3.7K). Non-overlapping partitions; 10% held out for validation.

Classification-head protocol. Task-specific classification heads are kept fully local: they are never transmitted or aggregated, and only the LoRA factors $\mathbf{A}, \mathbf{B}$ are communicated and averaged at the server. Each client trains its own task's head jointly with its LoRA factors and is evaluated with that local head on its held-out validation split; overall accuracy is the unweighted mean of per-task accuracies. The same protocol is applied identically across all methods, so it does not differentially affect any comparison. Heads use the same optimizer and learning rate as the LoRA factors and the standard random initialization of the model's classification module.

Model. RoBERTa-base, LoRA on all query/value projections (24 modules). $\mathbf{A} \in \mathbb{R}^{768\times 8}$, $\mathbf{B} \in \mathbb{R}^{8\times 768}$, $\alpha/r = 16/8 = 2$. Total LoRA: $\sim$295K. Classifier heads: $\sim$590K (task-specific).

Model (DeBERTa-v3-large). DeBERTa-v3-large (d=1024, 24 layers, 435M parameters), LoRA on all query/value projections (48 modules). $\mathbf{A} \in \mathbb{R}^{1024\times 8}$, $\mathbf{B} \in \mathbb{R}^{8\times 1024}$, $\alpha/r = 16/8 = 2$. Backbone frozen. Mixed precision (bfloat16) used to fit within 12 GB VRAM.

Training. AdamW, $\eta = 5 \times 10^{-4}$, no weight decay on LoRA, full precision (fp32). P=200 steps/round, 30% participation (12/40 clients). Same random selection sequence across methods.

Training (DeBERTa-v3-large). AdamW, $\eta = 3 \times 10^{-4}$, no weight decay on LoRA, P=200, T=200, 30% participation, batch size 8 (smaller than RoBERTa-base due to memory).

Instrument hyperparameters. For RoBERTa-base, $\lambda_0 = 0.05$ and $\epsilon_{\text{target}} = 0.01$ (default). For the DeBERTa-v3-large λ_0 sweep (Table 2), λ_0 varies as the swept dimension while $\epsilon_{\text{target}} = 0.01$ is held fixed across all λ_0 values.

Baselines. FFA-LoRA: $\mathbf{A}$ Kaiming init, frozen; only $\mathbf{B}$ trained/aggregated. RoLoRA: even rounds freeze $\mathbf{A}$, odd freeze $\mathbf{B}$. LS-Corr: after factor-wise averaging, solve $\min_{\Delta\mathbf{B}} \left\|\bar{\mathbf{A}}\Delta\mathbf{B} - \mathbf{C}_\omega\right\|_F$ in least squares and set $\bar{\mathbf{B}} \leftarrow \bar{\mathbf{B}} + \Delta\mathbf{B}$ (a rank-r post-hoc correction; it does not absorb the residual into $\mathbf{W}_0$ as FedEx-LoRA does). FlexLoRA: the server averages the products $\bar{\mathbf{M}} = \sum_k \omega_k \mathbf{A}_k\mathbf{B}_k$, computes the rank-$r$ truncated SVD $\bar{\mathbf{M}} \approx \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^\top$, and broadcasts the balanced re-factorization $\bar{\mathbf{A}} = \mathbf{U}_r\mathbf{S}_r^{1/2}$, $\bar{\mathbf{B}} = \mathbf{S}_r^{1/2}\mathbf{V}_r^\top$; the truncation residual $\left\|\bar{\mathbf{M}} - \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^\top\right\|_F$ is logged every round.

Matched pipeline for Tables 3 and 6. These two experiments were run after the main experiments in a separate training configuration: bf16 autocast, $\eta = 10^{-4}$, and 40% participation (16/40 clients at K=40); all other settings (model, LoRA rank, P=200, data partitions, local heads, optimizer) are as above. Because learning rate strongly affects the absolute dispersion and error scale, absolute numbers from this pipeline are not comparable to Tables 1 and 4; every comparison drawn from them is within the matched pipeline. For Table 3, FlexLoRA and FedAvg share the client-selection sequence for each seed (paired comparison). For the K-sweep, the 6-task client assignment scales proportionally: K=20: 4/4/4/3/3/2, K=40: 8/8/8/5/6/5, K=80: 16/16/16/10/12/10 clients for SST-2/MNLI/QNLI/RTE/QQP/MRPC, with per-client shards thinning accordingly.

Compute. 2$\times$ NVIDIA RTX 3080 Ti (12 GB VRAM each). Each 300-round RoBERTa-base experiment takes $\sim$10–12 hours; each 200-round DeBERTa-v3-large experiment takes $\sim$16 hours.

D.1 TABLE-TO-CONFIGURATION MAP

Table 7 records, for every main table and figure, the instrument form, seeds, and (T, P, K) configuration used to produce it. The exact run commands and result-file paths for each row are provided in the supplementary README.

Table 6: Client-count sweep (FedAvg, RoBERTa-base, T=100, P=200, seed 42; matched pipeline of Table 3). Error and dispersion grow with K; accuracy does not move.

K	Acc. (%)	Agg. Err.	$\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B})$
20	79.63	0.0129	0.0019
40	80.48	0.0165	0.0048
80	80.14	0.0322	0.0111

Table 7: Instrument form, seeds, and configuration per table/figure.

Artifact	Form	Seeds	Configuration
Table 1, Table 8, Fig. 1	CP (Eq. 6)	{42,123,456}; Fig. 42	T=300, P=200, K=40, 30%, fp32
Table 2 (DeBERTa-v3-large)	OS (Eq. 7)	multi-seed rows {42,123,456}; sweep 42	T=200, P=200, K=40, 30%, bf16
Table 4, Fig. 2(b)	CP	42	T=100, $P \in \{5, 50, 200, 500\}$, K=40, 30%, fp32
Table 10	CP	42	T=100, P=200, K=40, 30%, fp32
Tables 9, 11	CP	42	T=300, P=200, K=40, 30%, fp32
Fig. 2(a)	CP	3-seed means	as Table 1
Table 3	– (FedAvg / FlexLoRA)	{42,123,456}, paired	T=300, P=200, K=40, 40%, bf16, η=10^{-4}
Table 6	– (FedAvg)	42	T=100, P=200, $K \in \{20, 40, 80\}$, 40%, bf16, η=10^{-4}
Appendix E	CP	42	T=300, P=200, K=50, SST-2 Dirichlet

E Non-IID Within-Task Results

We additionally evaluate on SST-2 with Dirichlet α=0.1 (50 clients, severe label skew). Results: FedAvg 92.2%, FedDecoRA 92.3%, both with error $\sim$0.001. The error is 30$\times$ smaller than cross-task (0.036), indicating that *task heterogeneity, not label skew, is the primary driver of aggregation error*. This validates the covariance decomposition finding: $\mathbf{C}^{\text{sub}}$ (subspace misalignment between tasks) dominates in cross-task FL, and vanishes when all clients share a single task.

F On FedSA-LoRA and personalized designs

FedSA-LoRA (Guo et al., 2025) shares and averages only the $\mathbf{A}$ matrices while each client keeps its own $\mathbf{B}$ local. No single global product $\bar{\mathbf{A}}\bar{\mathbf{B}}$ is formed, so the aggregation error of Proposition 1 is not defined for it: the design does not manage or eliminate the error but sidesteps the aggregation of $\mathbf{B}$ altogether. It is therefore a personalized-FL scheme rather than a global-model baseline, and we do not include it in Tables 1 or 3: in our cross-task setting a local $\mathbf{B}$ obtains a task-specific personalization benefit that is unrelated to aggregation, and a direct comparison would conflate that benefit with the aggregation effects this paper isolates. Our results apply only to methods that form a global model (families i–iii of Section 2).

G Extended Theoretical Discussion

This section provides detailed intuitions for the exact results, complementing the concise statements in Section 3.

Interpretation of Proposition 1. The identity $\mathbf{E} = \mathbf{C}_\omega(\mathbf{A}, \mathbf{B})$ reveals that aggregation error is *not* an inherent consequence of updating both factors simultaneously. Rather, it arises only when

client deviations $\delta\mathbf{A}_k$ and $\delta\mathbf{B}_k$ are *correlated* across clients. Consider two scenarios: (i) Client k moves $\mathbf{A}_k$ far from $\bar{\mathbf{A}}$ but keeps $\mathbf{B}_k$ near $\bar{\mathbf{B}}$; then $\delta\mathbf{A}_k\delta\mathbf{B}_k \approx \mathbf{0}$ regardless of $\|\delta\mathbf{A}_k\|$, so the error is small. (ii) Client k moves both $\mathbf{A}_k$ and $\mathbf{B}_k$ in a correlated fashion; then $\delta\mathbf{A}_k\delta\mathbf{B}_k$ is large, and the covariance accumulates across clients. This distinction is critical: factor-freezing methods (FFA-LoRA, RoLoRA) address scenario (ii) by forcing one deviation to zero, but this also prevents beneficial adaptation in scenario (i). A method that merely decorrelates the deviations without constraining their magnitudes can control the error while preserving expressiveness.

Design implications of Proposition 2. The bound $\|\mathbf{C}_\omega\|_F \leq \sigma_\omega(\mathbf{A}) \cdot \sigma_\omega(\mathbf{B})$ is a *product* of dispersions. This has three consequences for method design. First, controlling either factor's dispersion alone suffices: if $\sigma_\omega(\mathbf{A}) = 0$ (FFA-LoRA), the bound is zero regardless of $\sigma_\omega(\mathbf{B})$. Second, the bound is *symmetric*: doubling $\sigma(\mathbf{A})$ while halving $\sigma(\mathbf{B})$ leaves the bound unchanged. This motivates a regularizer that allows asymmetric factor drift, letting adaptation capacity flow to whichever factor needs it. Third, the product structure suggests a *multiplicative* penalty $\|\delta\mathbf{A}\|^2 \cdot \|\delta\mathbf{B}\|^2$ rather than an additive one $\|\delta\mathbf{A}\|^2 + \|\delta\mathbf{B}\|^2$: the multiplicative form directly targets the quantity that bounds the error, whereas the additive form over-penalizes benign configurations where one factor is near zero.

Geometric interpretation of the covariance decomposition. The three-component decomposition (Eq. 5) admits a geometric interpretation. $\mathbf{C}^{\text{sub}}$ captures *subspace misalignment*: an SST-2 client learning sentiment features and an MNLI client learning entailment features will have orthogonal deviations $\delta\mathbf{A}_k^{\perp}$, generating large $\mathbf{C}^{\text{sub}}$. This is the dominant component ($\gtrsim 84\%$ in our experiments) because cross-task FL creates fundamentally different optimization landscapes. $\mathbf{C}^{\text{scale}}$ captures *scale divergence*: deviations that stay inside the global column and row spaces (re-scalings and re-mixings of the r global directions), e.g., a client with 393K MNLI samples producing larger gradient updates than one with 2.5K RTE samples. This is small ($\lesssim 2\%$) because LoRA factors are normalized by the α/r scaling. $\mathbf{C}^{\text{cross}}$ captures the interaction between subspace and scale effects.

Convergence: per-step vs. round-level control. The key contrast in Theorem 3 is between the $O(P^4\eta^4G^4)$ aggregation term for unconstrained FedAvg+LoRA (the error itself grows as $O(P^2\eta^2G^2)$) and the $O(\lambda^{-4/3})$ term for the instrument. Round-level methods allow the unfrozen factor to accumulate P steps of drift before any correction occurs at aggregation; since covariance is a product of deviations, the error grows quadratically with P. FedDecoRA's per-step regularizer counteracts this accumulation: at every gradient step, it applies a restoring force on each factor scaled by the other factor's squared deviation. This makes the leading term of the bound independent of P, a crucial property for communication-efficient FL, where P may be 200–500 to reduce the number of communication rounds.

G.1 CP FORM: EMPIRICAL VALIDATION AND SAME-ORDER TERM COMPARISON WITH OS

This appendix presents (a) the empirical validation of the CP-form instrument introduced in §4.1 (Eq. 6) on RoBERTa-base across three seeds, (b) a detailed comparison with the OS-form instrument used in the DeBERTa-large λ sweep (Eq. 7) showing the two forms yield the same empirical conclusion, and (c) the formal accounting of the same-order product-variance term that arises when aggregating the OS form and that the CP form does not carry (although, as shown there, neither aggregated penalty upper-bounds $\|\mathbf{C}_\omega\|_F^2$).

Empirical validation: CP form on RoBERTa-base. The CP instrument (Eq. 6) is the construction used for the FedDecoRA row of Table 1, evaluated across three seeds (42, 123, 456). Table 8 summarizes the headline comparison against the FedAvg baseline: the CP form reduces aggregation error by 38% (from 0.0364 to 0.0225) at no measurable accuracy cost (final accuracy $84.41\pm.27\%$ vs. FedAvg's $84.45\pm.22\%$, mean difference -0.04%, well within seed-level standard deviation). A two one-sided test (TOST) for equivalence with margin $\Delta{=}0.5\%$ rejects non-equivalence at $\alpha{=}0.05$, establishing equivalence rather than mere failure-to-reject (full TOST details in §5.2, F4).

Cross-form consistency: CP and OS yield the same empirical conclusion. The CP form (Eq. 6, used here on RoBERTa-base) and the OS form (Eq. 7, used in Table 2 on DeBERTa-large) target different statistics of the aggregation error and have different gradient structures: the CP form

Table 8: CP-form instrument on RoBERTa-base (the construction reported in Table 1's FedDecoRA row), K=40, T=300, P=200, r=8, mean $\pm$ std over 3 seeds (42, 123, 456). The CP instrument achieves 38% error reduction with mean accuracy difference of only -0.04% vs. FedAvg, the central empirical finding of the paper at the smaller model scale.

Method	Final Accuracy	Aggregation Error
FedAvg	$84.45 \pm 0.22\%$	0.0364 ± 0.0009
Instrument (CP form), λ_0=0.05	$84.41 \pm 0.27\%$	$\mathbf{0.0225 \pm 0.0005}$
Mean difference	-0.04%	-38.2%

targets a per-client surrogate of the dispersion product $\sigma_\omega(\mathbf{A})\sigma_\omega(\mathbf{B})$ (which upper-bounds the error), while the OS form targets the product residual $\left\|\mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t\right\|_F$ directly. Despite these structural differences, both forms yield the same empirical conclusion: error reduction produces no measurable accuracy improvement (33–96% reduction with the CP form on RoBERTa-base, Table 4; up to 23% with the OS form on DeBERTa, Table 2). The agreement across two regularizer forms strengthens the claim that the empirical finding is a property of the federated LoRA landscape rather than an artifact of any particular regularization choice.

Same-order product-variance term in the OS form (formal account). Aggregating the per-client OS penalty (used for the DeBERTa λ sweep) does not cleanly recover $\|\mathbf{C}_\omega\|_F^2$. By the parallel-axis identity for Frobenius norms (let $\mathbf{M}_k = \mathbf{A}_k\mathbf{B}_k$, $\boldsymbol{\mu} = \sum_k \omega_k \mathbf{M}_k$, $\mathbf{G} = \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t$ the broadcast product):

$$\sum_k \omega_k \left\|\mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t\right\|_F^2 = \underbrace{\sum_k \omega_k \|\mathbf{M}_k - \boldsymbol{\mu}\|_F^2}_{\text{across-client product variance}} + \underbrace{\|\boldsymbol{\mu} - \mathbf{G}\|_F^2}_{\boldsymbol{\mu}-\mathbf{G} = \mathbf{C}_\omega + (\bar{\mathbf{A}}^{t+1}\bar{\mathbf{B}}^{t+1} - \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t)} . \tag{18}$$

The across-client product-variance term is of the same order as $\|\mathbf{C}_\omega\|_F^2$ in general (not a lower-order shift) and can dominate, and the second term also contains the global product update. The CP form (Eq. 6) does not carry the variance term, but, as discussed in Appendix B, its aggregated penalty is likewise not an upper bound on $\|\mathbf{C}_\omega\|_F^2$ under the broadcast anchor (clients drifting in complementary factors incur zero penalty but nonzero error). Both forms are therefore surrogates whose effect on the error is verified empirically. Constructing a regularizer whose formal guarantee targets the aggregation error directly (i.e., that separates the variance and global-update contributions from the error term) remains an open question, with three possible directions: (i) showing the product-variance term vanishes asymptotically or is bounded by quantities the adaptive schedule controls; (ii) reframing the regularizer as a product-space proximal operator rather than a $\|\mathbf{C}_\omega\|_F^2$ minimizer; or (iii) introducing an explicit anchor that absorbs the variance term. We leave this to future work; the CP-form analysis in Appendix B provides supporting intuition under an explicit balanced-drift condition.

Empirical robustness across forms and scales. The OS instrument's final-accuracy variation across the six single-seed λ_0 values in Table 2 is 0.33% over more than three orders of magnitude on DeBERTa-large; the CP instrument matches FedAvg within 0.04% across three seeds on RoBERTa-base. This dual robustness across two structurally distinct parameterizations and across two model scales is the strongest empirical support for the central claim: even if the OS same-order term contributes nontrivially to the realized minimum, the conclusion that varying error along the λ axis does not strongly drive accuracy is supported by both forms.

H Extended Experimental Analysis

H.1 Training dynamics and ablation tables

The three ablation tables referenced in Section 5.6 follow (training dynamics: Fig. 1). All experiments use RoBERTa-base on the cross-task setting (seed 42).

Table 9: Covariance decomposition at round 300. Subspace misalignment ($\mathbf{C}^{\text{sub}}$) accounts for $\gtrsim$84% of the summed component norms across methods. FedDecoRA reduces all components proportionally; LS-Corr's rank-r correction cannot represent the full-rank $\mathbf{C}^{\text{sub}}$. The rightmost column reports $\sum_i \left\|\mathbf{C}^{(i)}\right\|_F$, an upper bound on $\|\mathbf{C}_\omega\|_F$ by the triangle inequality; Table 1's Agg. Err. column reports $\|\mathbf{C}_\omega\|_F$ directly, and the two differ (e.g., 0.043 vs. 0.036 for FedAvg) because the components are not mutually orthogonal, so their norms do not add. The percentage column is $\left\|\mathbf{C}^{\text{sub}}\right\|_F$ relative to the summed component norms.

	$\left\|\mathbf{C}^{\text{sub}}\right\|_F$	(%)	$\left\|\mathbf{C}^{\text{scale}}\right\|_F$	$\|\mathbf{C}^{\text{cross}}\|_F$	$\sum_i \left\|\mathbf{C}^{(i)}\right\|_F$
FFA-LoRA / RoLoRA	0.000		0.000	0.000	0.000
FedAvg	0.036	84%	0.001	0.006	0.043
LS-Corr	0.037	84%	0.001	0.007	0.044
FedDecoRA	**0.018**	86%	**0.000**	**0.003**	**0.021**

Table 10: λ_0 sensitivity for the FedDecoRA instrument (T=100, P=200). The λ_0=0.05 setting reduces error by 33% while slightly improving accuracy; larger λ_0 over-regularizes. Practical scaling: $\lambda_0 \propto 1/P$.

λ_0	Acc. (%)	Agg. Err.	Err. $\downarrow$	$\sigma_\omega(\mathbf{A})\cdot\sigma_\omega(\mathbf{B})$
0 (FedAvg)	81.8	0.037		0.216
0.01	82.0	0.035	-6%	0.183
0.05	**82.2**	0.025	-33%	0.130
0.10	81.3	0.019	-49%	0.110

H.2 Detailed baseline descriptions

FedAvg+LoRA (McMahan et al., 2017). The standard federated averaging applied to LoRA factors. Both $\mathbf{A}$ and $\mathbf{B}$ are trained locally and aggregated via weighted averaging each round. No mechanism addresses the bilinear aggregation inconsistency. This serves as the unconstrained baseline.

FFA-LoRA (Sun et al., 2024). $\mathbf{A}$ is initialized with Kaiming uniform and permanently frozen throughout all T rounds. Only $\mathbf{B}$ is trained and aggregated. This ensures $\delta\mathbf{A}_k = \mathbf{0}$ for all clients, so $\mathbf{C}_\omega = 0$ by Proposition 1. However, this confines all adaptation to $\text{col}(\mathbf{A}_{\text{init}})$, a fixed rank-$r$ subspace determined at random initialization. The model cannot escape this subspace regardless of the data, fundamentally limiting its representational capacity.

RoLoRA (Chen et al., 2025). Even-numbered rounds freeze $\mathbf{A}$ and aggregate only $\mathbf{B}$; odd-numbered rounds freeze $\mathbf{B}$ and aggregate only $\mathbf{A}$. This eliminates per-round aggregation error while allowing both factors to eventually update. However, each round's gradient for the frozen factor is computed during the backward pass but *never applied*, discarding 50% of gradient information. The effective update space alternates between two rank-r subspaces.

LS-Corr (rank-r least-squares residual correction). After standard factor-wise averaging, the server computes the aggregation residual $\mathbf{C}_\omega = \sum_k \omega_k \mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}\bar{\mathbf{B}}$ and solves $\min_{\Delta\mathbf{B}} \left\|\bar{\mathbf{A}}\Delta\mathbf{B} - \mathbf{C}_\omega\right\|_F$, updating $\bar{\mathbf{B}} \leftarrow \bar{\mathbf{B}} + \Delta\mathbf{B}$. This is the low-rank post-hoc correction pattern (cf. the aggregation refinement of LoRA-FAIR (Bian et al., 2025), which adds a regularizer). The correction $\bar{\mathbf{A}}\Delta\mathbf{B}$ has rank at most r=8 and column space $\text{col}(\bar{\mathbf{A}})$, so it cannot represent a generic 768×768 residual, which our experiments confirm: LS-Corr's error (0.036) equals FedAvg's. We stress that FedEx-LoRA (Singhal et al., 2025) is a different method: it absorbs the full residual into the frozen weights $\mathbf{W}_0$ and therefore achieves exact aggregation; it belongs with FlexLoRA and FLoRA in the exact-aggregation family and is not evaluated here.

FlexLoRA (Bai et al., 2024). The server averages the client products $\mathbf{A}_k\mathbf{B}_k$ rather than the factors and re-factorizes the mean by rank-r truncated SVD (Appendix D). The factor-averaging residual

Table 11: LoRA rank ablation (T=300). At r=16, aggregation error decreases $\sim$15$\times$ naturally (ratio column, computed before rounding); FFA-LoRA's accuracy gap narrows from -3.3% to -0.9% as the frozen-subspace constraint becomes less binding.

	r=8		r=16		Change	
Method	Acc.	Err.	Acc.	Err.	ΔAcc.	Err. ratio
FFA-LoRA	81.4	0.000	83.4	0.000	$+2.0$	
RoLoRA	83.2	0.000	83.6	0.000	$+0.4$	
FedAvg	84.7	0.036	84.3	0.002	-0.4	0.06$\times$
LS-Corr	85.3	0.037	84.5	0.002	-0.8	0.06$\times$
FedDecoRA	84.6	0.018	84.2	0.002	-0.4	0.13$\times$

of Proposition 1 never arises; the only approximation is the truncation residual, which is ≤ 0.001 throughout training in our setting. Unlike LS-Corr, which must represent $\mathbf{C}_\omega$ inside the fixed column space $\mathrm{col}(\bar{\mathbf{A}})$, the SVD selects the optimal rank-r subspace of the averaged product; FedEx-LoRA likewise avoids this limitation by keeping the residual full-rank in $\mathbf{W}_0$. The re-factorization re-balances and rotates the factors every round, which is the mechanism to which we attribute its accuracy difference from FedAvg (Sec. 5.4).

FedProx (Li et al., 2020). Adds a proximal penalty $\frac{\mu}{2}\|\theta_k - \bar{\theta}\|^2$ (μ=0.01) to each client's local objective, suppressing all parameter drift toward the global model. This is not LoRA-specific and does not distinguish between drift components. Included to compare additive regularization (treats all drift equally) against FedDecoRA's multiplicative regularization (targets only cross-factor covariance).

H.3 Aggregation error metric: snapshot estimator vs. theoretical $\mathbf{C}_\omega$

Proposition 1 states $\sum_k \omega_k \mathbf{A}_k\mathbf{B}_k - \bar{\mathbf{A}}\bar{\mathbf{B}} = \mathbf{C}_\omega(\mathbf{A}, \mathbf{B})$, which vanishes, in particular, when $\delta\mathbf{A}_k \equiv 0$ (FFA-LoRA) or $\delta\mathbf{B}_k \equiv 0$ (RoLoRA). The Agg. Err. column in our tables reports an empirical *snapshot estimator* of this quantity:

$$\widehat{\mathrm{AggErr}} = \frac{1}{|\mathrm{proj}|}\sum_{\mathrm{proj}} \left\| \sum_k \omega_k \mathbf{A}_k^{(t)}\mathbf{B}_k^{(t)} - \bar{\mathbf{A}}^{(t)}\bar{\mathbf{B}}^{(t)} \right\|_F, \tag{19}$$

computed over snapshots of client and server states at evaluation rounds. For methods where one factor is held truly invariant (FFA-LoRA: $\mathbf{A}_k^{(t)} \equiv \bar{\mathbf{A}}^{(t)}$ at every step), the snapshot estimator equals zero up to floating-point precision; we report 0.000 in Table 1 and the surrounding text.

The DeBERTa-large Table 2 value of 0.0169 for FFA-LoRA arises from a small B-only deviation that is a snapshot artifact: client states are recorded after the local update at the round's eval point, while the global state $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ used in the metric corresponds to the start-of-round broadcast. Because $\mathbf{A}_k^{(t)} = \bar{\mathbf{A}}^{(t)}$ exactly throughout the round (FFA freezes A), $\delta\mathbf{A}_k \equiv 0$ and the theoretical $\mathbf{C}_\omega = 0$; the residual 0.0169 comes from the alignment of client-recorded vs. server-recorded B states across the round boundary. This is an estimator artifact, not a violation of Proposition 1. Because the start-of-round anchor is used for every DeBERTa row, each reported value in Table 2 includes the global product update $\bar{\mathbf{A}}^{t+1}\bar{\mathbf{B}}^{t+1} - \bar{\mathbf{A}}^{t}\bar{\mathbf{B}}^{t}$ in addition to $\mathbf{C}_\omega$; the column is therefore comparable within Table 2 but not to the RoBERTa tables. The same mechanism produces the 0.0019 reported for FlexLoRA in Table 3, whose theoretical $\mathbf{C}_\omega$ is likewise exactly zero.

Effect on relative comparison. The snapshot estimator inflates reported error in proportion to within-round update magnitude. Because FedDecoRA suppresses client drift via the λ-penalty, its within-round update magnitudes are smaller than FedAvg's; its reported $\widehat{\mathrm{AggErr}}$ may therefore underestimate the true $\|\mathbf{C}_\omega\|_F$ less than FedAvg's reported value does, and may overestimate it more than FFA's. Computing the metric using post-update aligned global means $(\bar{\mathbf{A}}^{(t+1)}, \bar{\mathbf{B}}^{(t+1)})$ rather than start-of-round states would remove this confound; we did not implement that alternative due to logging overhead, but expect the directional ordering of methods (factor-freezing $\leq$ FedDecoRA $\leq$ FedAvg/LS-Corr) to be preserved. We retain the reported values for transparency rather than rounding

to zero, since the small magnitude does not affect any conclusion: the headline finding ($30\times$ error variation $\Leftrightarrow$ $<1.5\%$ accuracy variation among non-freezing methods) holds under either estimator.

H.4 Per-task accuracy at P=500 and the T-unified comparison

Table 4 compares all values of P at a fixed round budget T=100. An earlier version of this analysis mixed a T=300, 3-seed row at P=200 (the Table 1 mean, 84.5%) with T=100 rows at the other P, which produced an apparent accuracy drop from P=200 to P=500 and an interpretation in terms of small-task overfitting. With T held fixed, that drop disappears: FedAvg's accuracy increases monotonically with P ($68.9 \to 80.0 \to 81.8 \to 82.4$) while its error grows $\sim$190$\times$, and we retract the overfitting interpretation. We therefore interpret the P axis only through Table 4.

H.5 Error–accuracy correlation details

We computed the Pearson and Spearman correlation between aggregation error $\|\mathbf{C}_\omega\|_F$ and overall accuracy across all method$\times$seed combinations:

Comparison	Pearson r (p-value)	Spearman ρ (p-value)
Final round, 15 points (5 methods* $\times$ 3 seeds)	$+0.78$ ($p < 0.001$)	$+0.71$ ($p < 0.01$)

*FedAvg, LS-Corr, FedDecoRA, FFA-LoRA, RoLoRA. The positive correlation arises because the unconstrained methods (FedAvg, LS-Corr) achieve both higher error and higher accuracy, while the factor-freezing methods (FFA-LoRA, RoLoRA) reduce error but sacrifice accuracy. This is *not* evidence that error *helps*; rather, it shows that the constraints imposed to reduce error are more harmful than the error itself.

Within-method seed-level correlations: structural interpretation, not statistical proof. A natural concern about the r=$+0.78$ between-method correlation is whether it reflects an ecological inference fallacy: if the underlying within-method (i.e., seed-level, with method held fixed) relationship between error and accuracy is positive, then aggregating across methods could produce a misleading positive correlation. To examine this, we compute the within-method seed-level Pearson correlation for each of the five methods using the same RoBERTa-base data:

Method	n	Pearson r	p-value
FedAvg	3	-0.31	0.80
FedDecoRA (CP form)	3	-0.80	0.41
LS-Corr	3	$+0.92$	0.25
FFA-LoRA	3	-0.69	0.51
RoLoRA	3	-0.82	0.39
Mean across methods		-0.34	(all $p > 0.2$)

Statistical interpretation. With n=3 per method, the within-method correlations carry no statistical power: even $|r|$=0.99 would not reach α=0.05 at this sample size. The within-method estimates are therefore *uninformative* about whether seed-level error variation drives accuracy. We do not claim these data rule out the within-method effect; rather, they show that we cannot detect such an effect at the available sample size. The structural interpretation of the between-method r=$+0.78$ does not rest on the within-method correlations being significantly negative: it rests on the observation that within any single method, the error magnitude is essentially fixed (seed variation in $\|\mathbf{C}_\omega\|_F$ is on the order of 0.001 on a 0.036 scale), so the cross-method ranking of error magnitudes is determined by structural method differences, not seed-level realizations. Under this interpretation, the between-method correlation reflects which constraint each method imposes on the parameter space, not a causal effect of error on accuracy.

H.6 Extended λ_0 sensitivity at P=500

At P=500, the regularizer strength must be reduced because each local training run accumulates more regularization over more steps. We tested three values:

λ_0	Acc. (%)	Agg. Error	Error ↓	$\sigma_\omega(\mathbf{A})\cdot\sigma_\omega(\mathbf{B})$
0 (FedAvg)	82.4	0.096		0.492
0.005	82.3	0.004	-96%	0.020
0.01	82.3	0.004	-96%	0.020
0.02	82.1	0.003	-97%	0.019
0.05	81.8	0.028	-71%	0.147

Three observations: (1) All $\lambda_0 \leq 0.02$ achieve $> 96\%$ error reduction with $< 0.3\%$ accuracy loss. (2) λ_0=0.05 (optimal for P=200) over-regularizes at P=500, costing 0.6% accuracy and paradoxically achieving *less* error reduction (71% vs. 96%) with a *larger* dispersion product (0.147 vs. 0.020): the over-strong penalty destabilizes local optimization rather than anchoring the factors. (3) The practical guideline $\lambda_0 \propto 1/P$ (i.e., $\lambda_0 = 0.05 \times 200/500 = 0.02$) falls within the optimal range. The adaptive schedule (Eq. 8) with $\epsilon_{\text{target}} = 0.01$ caps the per-round penalty magnitude relative to observed dispersions, preventing within-round over-regularization once λ_0 has been chosen; selecting λ_0 for each P regime still requires the manual $\lambda_0 \propto 1/P$ rule.

I Extended Discussion

Rank dominates error as a control knob. Doubling r from 8 to 16 shrinks the empirical error $\sim 15\times$ without any explicit error control, and FFA-LoRA's gap to FedAvg narrows from -3.3% to -0.9% (Table 11, seed 42): FFA-LoRA's behavior depends primarily on whether rank suffices to absorb the expressiveness cost, with error magnitude secondary. Rank and λ probe different things (rank changes subspace overlap and parameter count, λ varies error at fixed rank), so for the causal question the instrument is the cleaner axis.

Why does FedAvg tolerate high error? We hypothesize two mechanisms. First, *iterative correction*: each round's aggregation error $\mathbf{C}_\omega^t$ is small relative to the accumulated model $\mathbf{W}_0 + \bar{\mathbf{A}}^t\bar{\mathbf{B}}^t$ (error/model ratio $\sim 10^{-4}$), and subsequent rounds of gradient descent partially correct the perturbation (Stich, 2019; Lin et al., 2020). Second, *partial cancellation*: in our setting with 24 LoRA projections, the covariance terms $\mathbf{C}_\omega^{(l)}$ for different projections l may have different signs and partially cancel in their effect on the overall loss. Formalizing these hypotheses, perhaps via a perturbation analysis of the loss landscape, is an important direction for future work.

Additive vs. multiplicative regularization: a detailed comparison. The key difference between FedProx's additive penalty $\|\delta\mathbf{A}_k\|_F^2 + \|\delta\mathbf{B}_k\|_F^2$ and FedDecoRA's multiplicative $\|\delta\mathbf{A}_k\|_F^2 \cdot \|\delta\mathbf{B}_k\|_F^2$ can be understood geometrically. FedProx's iso-penalty contours in the $(\|\delta\mathbf{A}\|, \|\delta\mathbf{B}\|)$ plane are circles: all deviations of equal magnitude are penalized equally, regardless of correlation. FedDecoRA's iso-penalty contours are hyperbolas: large deviation in one factor is permitted as long as the other factor stays close to zero. Equivalently, in terms of zero sets: the FedProx penalty vanishes only at the single point $\delta\mathbf{A} = \delta\mathbf{B} = 0$, whereas the CP penalty vanishes on the entire union $\{\delta\mathbf{A} = 0\} \cup \{\delta\mathbf{B} = 0\}$; when every client adapts the same single factor, such configurations contribute zero aggregation error by Proposition 1. FedProx is blind to LoRA's bilinear structure and over-penalizes error-free single-factor adaptation, while the CP form targets exactly the correlated component that creates $\mathbf{C}_\omega$. This lets a client fully adapt one factor (e.g., $\mathbf{B}$ for a task-specific direction) while keeping the other near the global mean, the asymmetric behavior suggested by Proposition 2. This structural difference predicts that FedProx should reduce error at an accuracy cost while the instrument should not.

Table 1 is consistent with the prediction: FedProx reaches error 0.001 at -0.5% accuracy, whereas the instrument achieves a 38% error reduction at -0.04%. We note that FedProx is evaluated at a single μ; a μ sweep targeting intermediate error levels would sharpen this comparison and is left to future work.

Reassessing the rank-r correction (LS-Corr). LS-Corr achieves slightly higher accuracy than FedAvg ($84.9\pm.3\%$ vs. $84.5\pm.2\%$) despite not reducing error (0.036 in both). Since LS-Corr is exactly FedAvg plus the post-hoc correction $\bar{\mathbf{B}} \leftarrow \bar{\mathbf{B}} + \Delta\mathbf{B}$, disabling the correction recovers FedAvg identically, so this $+0.4\%$ is the corresponding ablation; at n=3 seeds the difference is within seed-level variation. One plausible but untested explanation is that the least-squares

correction $\Delta\mathbf{B} = (\bar{\mathbf{A}}^\top \bar{\mathbf{A}})^{-1}\bar{\mathbf{A}}^\top \mathbf{C}_\omega$, which projects the residual onto $\mathrm{col}(\bar{\mathbf{A}})$, acts as a mild implicit regularizer (Arora et al., 2019). Whatever its origin, the effect is unrelated to error reduction, which LS-Corr does not achieve at $r{=}8$.