# UNDERSTANDING HEAD GEOMETRY AND DYNAMICS IN FEDERATED REGRESSION THROUGH A NATURAL SOLUTION SELECTION RULE: AN UNCONSTRAINED FEATURE MODEL ANALYSIS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In federated averaging, local objectives can admit multiple optimal heads, making the aggregate depend on which heads clients return. We study this ambiguity in federated multivariate regression with private backbones and a shared linear head, using an unconstrained feature model (UFM) that treats training-sample features as free variables. We introduce a natural selection rule: each client returns the optimal head closest to the broadcast head. We show that global minimization with a vanishing proximal penalty on the head realizes this rule. When the clients' optimal Gram matrices and the initial shared Gram matrix are positive definite, the shared Gram matrix follows a closed recursion and converges to the unique Bures–Wasserstein barycenter of the clients' optimal Gram matrices. Even with this alignment, the limit generally differs from the centralized optimal Gram matrix. We decompose this gap into three positive-semidefinite terms arising from differences in client target means, covariance heterogeneity, and averaging the aligned heads. A correction based on a one-time exchange of target means and covariances recovers the centralized optimal Gram matrix in one round under exact local optimization and the same selection rule. We verify these results numerically in the UFM and test its predictions on five tabular and five image regression datasets using deep networks with feature regularization and long local training. In these experiments, ordinary training approaches the predicted barycenter, while a weak proximal penalty improves endpoint agreement and yields trajectories that closely follow the predicted Gram dynamics. The correction moves the final Gram matrices close to the centralized UFM prediction.

## 1 INTRODUCTION

In federated averaging, clients update the shared model parameters using their local data at each communication round and send the updated parameters to the server. The server then averages these parameters and broadcasts the result (McMahan et al., 2017). Effective aggregation relies on alignment between client models in parameter space, but local objectives alone do not guarantee such alignment. Prior work has characterized the limiting behavior of federated averaging in strongly convex settings, where local minimizers are unique (Charles & Konečný, 2021; Mangold et al., 2025). In overparameterized linear regression with fixed features, gradient descent selects the local minimizer closest to the broadcast model, and repeated averaging can recover the centralized solution (Zhu et al., 2026).

In contrast, local training of deep neural networks need not return the minimizer closest to the broadcast model. Different parameter settings can represent the same function (Wang et al., 2020a), so clients may return mutually incompatible solutions. For example, permuting hidden units can cause the same unit index to correspond to different features across clients. Averaging such solutions can produce an arbitrarily poor model (McMahan et al., 2017). Indeed, nonzero local updates can largely cancel upon averaging (Jhunjhunwala et al., 2023), and averaging-based optimization methods can stagnate at nonoptimal points (Pathak & Wainwright, 2020; Teng et al., 2019). To address mismatches in parameter representations, prior work has proposed matching methods across



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.

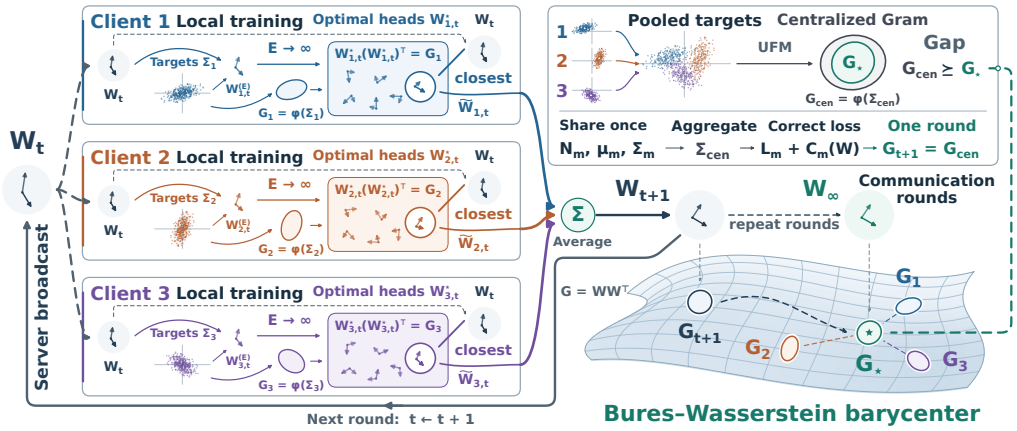


Figure 1: **Federated averaging of a shared head under the selection rule.** In the UFM, local optimality fixes the Gram matrix of a client's head but not the head itself. Our selection rule chooses the optimal head closest to the broadcast head. Averaging these heads over communication rounds drives the Gram matrix of the shared head to the BW barycenter of the clients' optimal Gram matrices. Inset: the gap to centralized training and its one-round correction from shared target moments.

clients before averaging (Wang et al., 2020a; Singh & Jaggi, 2020). These studies, however, did not characterize the limiting behavior of repeated averaging or its relation to centralized training.

To study the limiting behavior of federated averaging when local minimizers are nonunique, we focus on the unconstrained feature model (UFM) (Mixon et al., 2022; Fang et al., 2021; Zhu et al., 2021; E & Wojtowytsch, 2022; Lu & Steinerberger, 2022). This model treats the backbone outputs on the training samples as free optimization variables. Originally introduced to explain neural collapse (Papyan et al., 2020) in classification, it was later extended to regression (Andriopoulos et al., 2024; Ma et al., 2025). For regularized multivariate regression, Andriopoulos et al. (2024) derived an explicit expression for the Gram matrix of an optimal linear head in terms of the target covariance. The head itself, however, is determined only up to an orthogonal transformation.

Building on this characterization, we introduce and analyze a UFM for federated multivariate regression with private backbones and a shared linear head (Liang et al., 2020; Jang et al., 2023). Each client's features are optimized independently, and only the head is averaged (Figure 1). The orthogonal freedom, harmless for an individual model, now complicates aggregation because the server averages the heads rather than their Gram matrices. To resolve the ambiguity in the client returns, we select, for each client, the optimal head closest to the broadcast head in Frobenius norm. This uses the broadcast as a common reference while preserving optimality for each client's original objective. We further show that this selection arises as a vanishing-regularization limit. Specifically, we add a quadratic proximal penalty on the head, of the form used in FedProx (Li et al., 2020), and show that the heads of global minimizers of the penalized local objective converge to the selected head as the penalty weight tends to zero.

Under this selection rule, we identify the limiting Gram matrix of the shared head as the Bures–Wasserstein (BW) barycenter of the clients' optimal Gram matrices. When these matrices are positive definite, we prove convergence to this unique barycenter from any positive-definite initial Gram matrix. The key connection is that optimal head alignment in Frobenius norm corresponds to BW distance between Gram matrices (Bhatia et al., 2019): averaging the selected heads induces the barycenter iteration of Álvarez Esteban et al. (2016).

Even with exact local optimization and aligned heads, this barycenter generally differs from the optimal Gram matrix of the centralized UFM trained on the pooled data, i.e., the combined data from all clients. We express the centralized Gram matrix as the barycenter plus three positive-semidefinite terms, corresponding to differences in client target means, the nonlinear dependence of optimal Gram matrices on target covariances, and averaging heads rather than their Gram matrices. We further construct a correction to the local objectives that recovers the centralized Gram matrix

in a single communication round under the same selection rule. The correction requires a one-time exchange of the clients' target means and covariances (Figure 1, inset).

We first verify these results numerically in the UFM. To test whether they extend to learned features, we train deep networks with the feature and head penalties of the UFM on five tabular and five image regression datasets. With a weak proximal regularizer, the Gram matrix of the shared head closely follows the predicted communication dynamics and ends near the BW barycenter, and the correction moves this endpoint close to the centralized Gram matrix. Without the regularizer, the endpoint also lies near the barycenter but less precisely, and an intermediate proximal weight agrees best. Aligning the heads to the broadcast also implements the selection rule, whereas averaging unaligned heads raises the training objective. Removing the between-client mean term leaves a gap consistent with the decomposition. The agreement persists for other architectures, larger training sets and full-model averaging, and weakens with less local training, as our analysis of finite local steps in the UFM also shows (Appendix G). On the image datasets, proximal training keeps test error comparable, and the correction lowers it slightly.

## 2 Related work

**Unconstrained feature models.** The UFM and the layer-peeled model explain neural collapse through free-feature optimization (Mixon et al., 2022; Fang et al., 2021). Prior work characterizes their global minimizers, landscapes and dynamics under cross-entropy and squared losses (Zhu et al., 2021; Zhou et al., 2022; Han et al., 2022). Extensions address class imbalance (Fang et al., 2021; Thrampoulidis et al., 2022; Hong & Ling, 2024), deeper models (Tirer & Bruna, 2022; Súkeník et al., 2023; 2024) and constrained features (Tirer et al., 2023). Related analyses examine neural collapse in shallow ReLU (Hong & Ling, 2026) and mean-field networks (Wu & Mondelli, 2025). The UFM has also been extended to multi-label classification (Li et al., 2024), supervised contrastive learning (Behnia & Thrampoulidis, 2024), ordinal regression (Ma et al., 2025) and multivariate regression (Andriopoulos et al., 2024; 2025). In federated classification, related analyses study feature collapse and degradation under aggregation (Shi et al., 2024; 2023; Zhu et al., 2025), and other methods prescribe the classifier geometry (Li et al., 2023; Xiao et al., 2024). We build on the characterization of multivariate regression (Andriopoulos et al., 2024) to analyze repeated averaging of selected optimal heads across clients.

**Federated optimization and selection.** Prior analyses identify objective inconsistency and biased fixed points (Wang et al., 2020b; Pathak & Wainwright, 2020). Surrogate objectives describe local updates for quadratic losses (Charles & Konečný, 2021), and bias expansions quantify the limiting deviation from centralized training in strongly convex settings (Mangold et al., 2025). Multiple local updates also enable shared representation learning in multitask linear regression (Collins et al., 2022). With fixed features, local gradient descent in overparameterized linear regression selects the minimizer closest to its broadcast initialization (Zhu et al., 2026). Matching methods explicitly align models before averaging (Wang et al., 2020a; Singh & Jaggi, 2020), while FedProx penalizes deviations from the broadcast model (Li et al., 2020). In our UFM, the heads of global minimizers of the penalized objective converge to the selected head as the proximal weight vanishes. For classifier calibration, CCVR uses uploaded feature moments (Luo et al., 2021); our correction instead uses target means and covariances to recover the centralized Gram matrix.

**Optimal transport and matrix barycenters.** For positive-definite matrices, the BW distance is the Procrustes distance between their factors (Bhatia et al., 2019). Berardini et al. (2026) use BW barycenters to align frozen-encoder feature distributions in one-shot federated learning. Under our selection rule, repeated head averaging instead induces the convergent BW barycenter iteration of Álvarez Esteban et al. (2016) on the shared-head Gram matrix.

## 3 Formulation

**Federated multivariate regression.** We consider a federated multivariate regression problem with $M \geq 2$ clients. Client $m \in \{1, \dots, M\}$ owns $\mathcal{D}_m = \{(x_{m,i}, y_{m,i})\}_{i=1}^{N_m}$ with inputs $x_{m,i}$ and targets

$y_{m,i} \in \mathbb{R}^C$ ($C \geq 2$, not assumed centered), collected columnwise as $Y_m = [y_{m,1}, \ldots, y_{m,N_m}] \in \mathbb{R}^{C \times N_m}$. Write $X_m = [x_{m,1}, \ldots, x_{m,N_m}]$ for the inputs, and let $N = \sum_m N_m$ and $p_m = N_m/N$.

**Federated averaging (FedAvg) with private backbones and a shared head.** A class of federated methods keeps each client's feature backbone local and averages only a shared head (Liang et al., 2020; Jang et al., 2023). In our model, the shared head is the last linear layer $(W, b)$, with $W \in \mathbb{R}^{C \times P}$ and $b \in \mathbb{R}^C$, where $P$ is the common feature dimension. Only $(W, b)$ is broadcast and averaged, while the backbone parameters $\theta_m$ stay local. At each communication round $t$, the server broadcasts the current head $(W_t, b_t)$ to all clients. Each client trains this head together with its private backbone on local data and uploads only the resulting head $(\widetilde{W}_{m,t}, \widetilde{b}_{m,t})$. The server then sets $(W_{t+1}, b_{t+1}) = \sum_m p_m(\widetilde{W}_{m,t}, \widetilde{b}_{m,t})$.

**The unconstrained feature model.** The UFM replaces the backbone features $H_m(X_m; \theta_m) \in \mathbb{R}^{P \times N_m}$ with free features $H_m \in \mathbb{R}^{P \times N_m}$, giving the local objective

$$\mathcal{L}_m(W, b, H_m) = \frac{1}{2N_m} \left\| W H_m + b\mathbf{1}^\top - Y_m \right\|_F^2 + \frac{\lambda_H}{2N_m} \|H_m\|_F^2 + \frac{\lambda_W}{2} \|W\|_F^2 , \tag{1}$$

with $\lambda_H, \lambda_W > 0$; the bias is unregularized and the squared error sums over the $C$ outputs. In the rounds above, each client now optimizes $H_m$ in place of $\theta_m$, and the server still averages only $(W, b)$. The feature penalty is a common abstraction for standard penalties and normalizations on the backbone parameters, whose effects on the features can be complicated (Mixon et al., 2022; Fang et al., 2021). The centralized reference is another UFM with one head for the pooled data and minimizes

$$\mathcal{L}_{\text{cen}}(W, b, \{H_m\}, \{Y_m\}) = \sum_m p_m \mathcal{L}_m(W, b, H_m). \tag{2}$$

We compare the limit of the server's head Gram matrix $G_t = W_t W_t^\top$ with the Gram matrix of an optimal head of this centralized objective.

**Target moments and the fully active regime.** Here we introduce and summarize quantities important for later analysis:

$$\mu_m = \frac{Y_m \mathbf{1}}{N_m}, \quad \mu_g = \sum_m p_m \mu_m, \tag{3}$$

$$\Sigma_m = \frac{Y_m Y_m^\top}{N_m} - \mu_m \mu_m^\top, \quad \Sigma_{\text{within}} = \sum_m p_m \Sigma_m, \quad \Sigma_\mu = \sum_m p_m (\mu_m - \mu_g)(\mu_m - \mu_g)^\top, \tag{4}$$

$$\Sigma_{\text{cen}} = \Sigma_{\text{within}} + \Sigma_\mu. \tag{5}$$

For our theoretical analysis, we assume $P \geq C$ and $\lambda_{\min}(\Sigma_m) > \lambda_H \lambda_W$ for every client (hence $N_m \geq C + 1$). The same eigenvalue bound holds for $\Sigma_{\text{within}}$ and $\Sigma_{\text{cen}}$. We refer to these conditions as the *fully active regime*, under which every client's optimal head Gram matrix is positive definite.

**Selecting locally optimal heads.** In the fully active regime, the global minimizers of (1) share the same bias and head Gram matrix, while the head itself remains nonunique up to orthogonal transformations (Section 4.1) (Andriopoulos et al., 2025). Optimality alone therefore does not determine the server update. To resolve this ambiguity, we add the proximal regularizer of FedProx (Li et al., 2020), $\frac{\rho}{2}\|W - W_t\|_F^2$, to each client's objective. For a fixed broadcast $W_t$ of full row rank, the heads of the global minimizers of the penalized objective converge, as $\rho \downarrow 0$, to the unique optimal head of the original objective that is closest to $W_t$ in Frobenius norm (Lemma 1). In the next section, we analyze the dynamics of the shared head Gram matrix $G_t$ induced by imposing this limiting procedure at every round, and also characterize the gap between its limit and the optimal head Gram matrix of centralized training.

# 4 THEORETICAL ANALYSIS

## 4.1 THE OPTIMAL PARAMETERS AT EACH CLIENT

We first express the local optimization problem in terms of the head alone. Given a dataset $Y$ and fixed $W$, the minimization of the loss over $(b, H)$ gives

$$b^\star(W) = \mu, \qquad H^\star(W) = W^\top A^{-1}(Y - \mu \mathbf{1}^\top),$$

where $\mu$ is the sample mean of $Y$ and $A = WW^\top + \lambda_H I_C$. Substituting these expressions into the loss yields the profiled objective

$$F(W; T) = \frac{\lambda_H}{2} \operatorname{tr}(TA^{-1}) + \frac{\lambda_W}{2} \|W\|_F^2 , \tag{6}$$

where $T$ is the sample covariance of $Y$. This profiled objective is strictly convex with respect to the Gram matrix $G = WW^\top$, so the solution is unique and the corresponding stationary condition is $(G + \lambda_H I_C)^2 = (\lambda_H/\lambda_W)T$. Hence, the solution is given by the following function:

$$\phi(T) := \alpha T^{1/2} - \lambda_H I_C, \quad \alpha = \sqrt{\frac{\lambda_H}{\lambda_W}}. \tag{7}$$

This accords with the regression-UFM optimum of Andriopoulos et al. (2024) (Appendix A.1). Using this function $\phi$, the optimal head Gram matrices for the local and centralized objectives are written as $G_m = \phi(\Sigma_m)$ and $G_{\text{cen}} = \phi(\Sigma_{\text{cen}})$, respectively. Similarly, $G_{\text{within}} = \phi(\Sigma_{\text{within}})$ is the optimal head Gram matrix of the averaged profiled objective $\sum_m p_m F(W; \Sigma_m) = F(W; \Sigma_{\text{within}})$.

## 4.2 SELECTING THE OPTIMAL HEAD

As shown above, the local objective determines a unique optimal head Gram matrix $G_m$. However, since $F(W; T)$ depends on $W$ only through $WW^\top$, the optimal head remains nonunique up to right orthogonal transformations. We resolve this ambiguity by selecting the optimal head closest to the broadcast $W$:

$$\Pi_m(W) = \operatorname*{arg\,min}_{U : UU^\top = G_m} \|U - W\|_F^2 , \qquad G = WW^\top \succ 0. \tag{8}$$

For any feasible $U$, the squared distance expands as

$$\|U - W\|_F^2 = \operatorname{tr} G_m + \operatorname{tr} G - 2 \operatorname{tr}(UW^\top).$$

Finding the closest optimal head therefore reduces to maximizing $\operatorname{tr}(UW^\top)$, a rectangular Procrustes problem. Its minimum squared distance equals the squared BW distance between the two Gram matrices (Bhatia et al., 2019):

$$d_{\text{BW}}^2(G_{\text{ref}}, G_{\text{tar}}) := \operatorname{tr} G_{\text{ref}} + \operatorname{tr} G_{\text{tar}} - 2 \operatorname{tr}\left[ (G_{\text{ref}}^{1/2} G_{\text{tar}} G_{\text{ref}}^{1/2})^{1/2} \right].$$

The following lemma gives the unique minimizing head and shows how it is selected by a vanishing head-only proximal penalty.

**Lemma 1** (Closest optimal head and proximal selection)**.** *Let client $m$ be fully active and let the broadcast head $W$ satisfy $G = WW^\top \succ 0$. (i) The minimizer in* (8) *is unique. It equals $\Pi_m(W) = T_m(G)W$ with $T_m(G) = G^{-1/2}(G^{1/2} G_m G^{1/2})^{1/2} G^{-1/2} \succ 0$, and $\|\Pi_m(W) - W\|_F^2 = d_{\text{BW}}^2(G, G_m)$. (ii) For $\rho > 0$, every global minimizer $(U_\rho, b_\rho, H_\rho)$ of $\mathcal{L}_m(U, b, H) + \frac{\rho}{2}\|U - W\|_F^2$ has $b_\rho = \mu_m$ and $\mathcal{L}_m(U_\rho, b_\rho, H_\rho) - \min \mathcal{L}_m \le \frac{\rho}{2} d_{\text{BW}}^2(G, G_m)$, and $U_\rho \to \Pi_m(W)$ as $\rho \downarrow 0$.*

Part (i) follows from the classical Procrustes characterization of the BW distance (Bhatia et al., 2019), and guarantees that each client has a unique optimal head closest to the common broadcast. By choosing these heads, the selection rule resolves the orthogonal ambiguity and makes the server update well defined. Part (ii) shows that the head-only proximal penalty, of the form used in FedProx (Li et al., 2020), selects $\Pi_m(W)$ as $\rho \downarrow 0$, provided that the penalized objective is globally minimized for each $\rho > 0$. We impose this limiting procedure at every round, before the server averages the returned heads. The proof is given in Appendix A.2.

### 4.3 CONVERGENCE TO THE BW BARYCENTER

Under the selection rule, averaging the returned heads induces a closed recursion for the shared Gram matrix $G_t = W_t W_t^\top$. The following theorem shows that this recursion converges to the BW barycenter of the clients' optimal Gram matrices.

**Theorem 1** (Convergence to the Bures–Wasserstein barycenter)**.** *In the fully active regime of Section 3, let $G_0 = W_0 W_0^\top \succ 0$ and suppose each client returns $(\Pi_m(W_t), \mu_m)$. Then every $G_t$ is positive definite, $b_t = \mu_g$ for $t \geq 1$, and the autonomous Gram dynamics are*

$$G_{t+1} = \mathcal{G}(G_t) := G_t^{-1/2} \left[ \sum_m p_m (G_t^{1/2} G_m G_t^{1/2})^{1/2} \right]^2 G_t^{-1/2}. \tag{9}$$

*For every such initialization, $G_t$ converges to the unique positive-definite BW barycenter $G_\star$, characterized by*

$$G_\star = \operatorname*{arg\,min}_{X \succ 0} \sum_m p_m d_{\mathrm{BW}}^2(X, G_m), \qquad G_\star = \sum_m p_m (G_\star^{1/2} G_m G_\star^{1/2})^{1/2}. \tag{10}$$

*Proof sketch.* By Lemma 1, $\Pi_m(W_t) = T_m(G_t) W_t$, where each $T_m(G)$ is positive definite and satisfies $T_m(G)\, G\, T_m(G) = G_m$. All clients act on the same broadcast, so their average is $W_{t+1} = S(G_t) W_t$, where $S(G) = \sum_m p_m T_m(G) \succ 0$. Consequently $G_{t+1} = S(G_t) G_t S(G_t) \succ 0$. Substituting the expression for $T_m$ gives (9), while averaging the local biases gives $\mu_g$.

The resulting Gram map is the barycenter iteration of Álvarez Esteban et al. (2016). Their convergence theorem applies because the clients' optimal Gram matrices are positive definite, the weights are positive and sum to one, and $G_0 \succ 0$. It yields the limit and characterization in (10). Appendices A.2–A.3 give the projection proof and the full correspondence with that theorem. □

The limiting Gram matrix is therefore the BW barycenter of the local optimal Gram matrices $G_m = \phi(\Sigma_m)$. It depends only on the client covariances, aggregation weights, and regularization parameters, and is independent of the choice of $G_0 \succ 0$ and the feature dimension $P \geq C$. Centralized training instead yields $G_{\mathrm{cen}} = \phi(\Sigma_{\mathrm{cen}})$, which generally differs from this barycenter.

### 4.4 THE GAP TO CENTRALIZED TRAINING AND ITS CORRECTION

The barycenter fixed-point equation yields the following decomposition of the gap $G_{\mathrm{cen}} - G_\star$ into three positive-semidefinite terms.

**Proposition 1** (Three positive-semidefinite terms of the gap)**.** *Under the assumptions of Theorem 1, choose any $W_\star$ satisfying $W_\star W_\star^\top = G_\star$ and set $U_m^\star = \Pi_m(W_\star)$. Then $\sum_m p_m U_m^\star = W_\star$ and*

$$G_{\mathrm{cen}} - G_\star = \mathcal{M}_\mu + \mathcal{M}_\Sigma + \mathcal{M}_A. \tag{11}$$

*Here $\mathcal{M}_\mu = G_{\mathrm{cen}} - G_{\mathrm{within}}$, $\mathcal{M}_\Sigma = G_{\mathrm{within}} - \sum_m p_m G_m$, and $\mathcal{M}_A = \sum_m p_m G_m - G_\star$. All three matrices are positive semidefinite (PSD). Hence $G_\star \preceq \sum_m p_m G_m \preceq G_{\mathrm{within}} \preceq G_{\mathrm{cen}}$. The gap vanishes if and only if all client means agree and all client covariances agree. Moreover,*

$$\operatorname{tr} \mathcal{M}_A = \sum_m p_m d_{\mathrm{BW}}^2(G_\star, G_m). \tag{12}$$

*Proof sketch.* The fixed-point equation implies $\sum_m p_m T_m(G_\star) = I_C$, and hence $\sum_m p_m U_m^\star = W_\star$. Therefore, $\sum_m p_m G_m - G_\star = \sum_m p_m (U_m^\star - W_\star)(U_m^\star - W_\star)^\top$. Add and subtract $G_{\mathrm{within}}$ and $\sum_m p_m G_m$ in $G_{\mathrm{cen}} - G_\star$ to obtain (11). Square-root monotonicity and concavity make $\mathcal{M}_\mu$ and $\mathcal{M}_\Sigma$ PSD, respectively; $\mathcal{M}_A$ is a covariance matrix.

A zero total gap forces every PSD term to vanish. The mean term then forces the means to agree, and zero averaging covariance forces the returned heads, hence the client covariances, to agree. The converse follows by substitution. Taking traces and using the distance identity of Lemma 1 gives the BW variance in (12). Appendix B provides these steps in detail and bounds this variance by the pairwise BW distances between the $G_m$. □

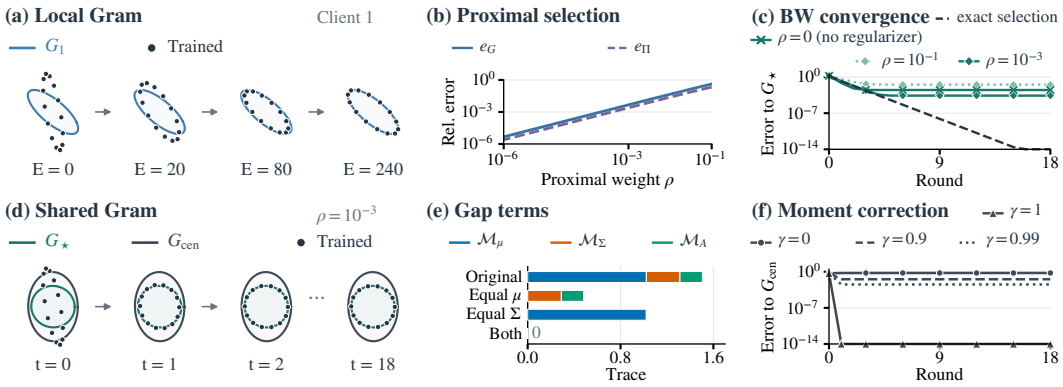


Figure 2: **UFM numerical checks on one fixed instance** ($C = P = 2$, $M = 3$, $\lambda_H = \lambda_W = 0.1$; Appendix D.1). Ellipses $\{G^{1/2}u : \|u\|_2 = 1\}$ represent Gram matrices $G$, and dots trace trained ones. $E$ counts full-batch local steps, and $\gamma$ scales the moment correction. Errors below $10^{-14}$ are drawn at the axis floor.

The term $\mathcal{M}_\mu$ captures the scatter of client means removed by clientwise centering, while $\mathcal{M}_\Sigma$ is the Jensen gap associated with the operator concavity of $\phi$. The term $\mathcal{M}_A$ is the weighted scatter matrix of the selected heads, and its trace equals the BW variance of the clients' optimal Gram matrices. Thus, resolving the orthogonal ambiguity does not by itself eliminate the gap to centralized training.

To recover the centralized Gram matrix, we make every client's profiled objective equal to $F(W; \Sigma_{\text{cen}})$. Since $F(W; T)$ is affine in $T$, this can be achieved by adding the head-only moment correction:

$$C_m(W) = \frac{\lambda_H}{2} \operatorname{tr}\{(\Sigma_{\text{cen}} - \Sigma_m)(WW^\top + \lambda_H I_C)^{-1}\}. \tag{13}$$

It satisfies $F(W; \Sigma_m) + C_m(W) = F(W; \Sigma_{\text{cen}})$ for every $W$. Thus all corrected clients have the same optimal Gram matrix $G_{\text{cen}}$.

**Proposition 2** (One-round recovery of the centralized Gram matrix)**.** *In the fully active regime of Section 3, suppose $W_t W_t^\top \succ 0$. If every client globally minimizes $\mathcal{L}_m + C_m$ and returns the optimal head closest to $W_t$, the returned head weight matrices $\widetilde{W}_{m,t}$ coincide, so $G_{t+1} = G_{\text{cen}}$, and the returned biases $\mu_m$ average to $b_{t+1} = \mu_g$.*

*Proof sketch.* The corrected clients have the same optimal Gram matrix $G_{\text{cen}}$ and receive the same broadcast. Uniqueness of the closest optimal head makes their returned weight matrices $\widetilde{W}_{m,t}$ identical, while their biases remain $\mu_m$. Averaging proves the claim. Appendix C also verifies existence of the corrected global minimizers. □

The correction requires a one-time exchange of client means and covariances, with $O(C^2)$ entries per client. The local feature solutions still use clientwise-centered targets, so matching the head Gram matrix and aggregate bias need not reproduce centralized predictions (Appendix C).

## 5 Experiments

We first check the UFM predictions numerically, then test whether they describe head geometry and dynamics in trained networks.

### 5.1 UFM numerical experiments

Figure 2 illustrates the theoretical results on one instance with two targets and three clients. Panel (a) shows joint optimization of $(W, b, H)$ bringing the local Gram matrix toward $G_m$. Panel (b) shows the relative errors to the optimal Gram ($e_G$) and the selected head ($e_\Pi$) decreasing as $\rho$ decreases, as predicted by Lemma 1.

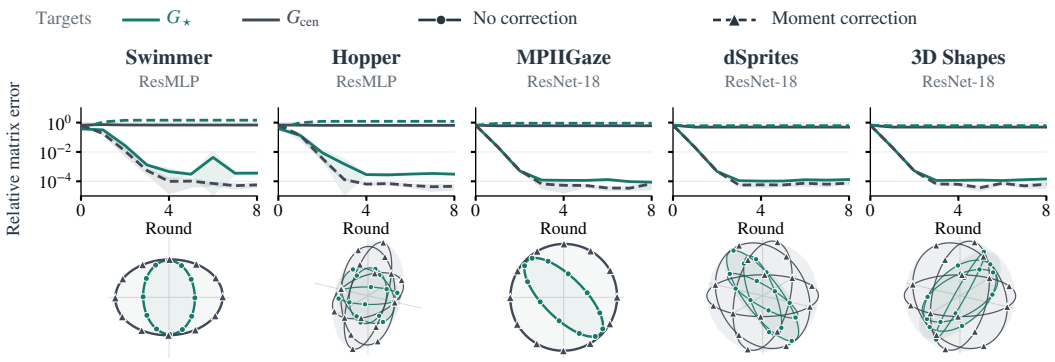


Figure 3: **Proximal training with and without moment correction** ($\rho = 10^{-3}$). Colors indicate the reference Gram; solid and dashed curves denote training without and with correction. Curves show seed means $\pm 1$ standard deviation. Below, outlines show $G_\star$ and $G_{\text{cen}}$, and markers show final Grams for seed 0. Figure 9 gives the other five settings.

Panel (c) compares the shared Gram matrices obtained by exact selection and by numerical optimization with and without the proximal penalty. Exact selection converges to $G_\star$, as predicted by Theorem 1. With numerically optimized heads, the final shared Gram is closer to $G_\star$ at $\rho = 10^{-3}$ than without the penalty. Panel (d) shows this $\rho = 10^{-3}$ trajectory as ellipses: the shared Gram approaches $G_\star$, while a gap to $G_{\text{cen}}$ remains. Appendix D.1 specifies the numerical objectives, solvers and initializations.

Panel (e) checks the gap decomposition of Proposition 1. Equalizing client means removes $\mathcal{M}_\mu$, equalizing covariances removes $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$, and equalizing both closes the gap. Panel (f) applies $\gamma C_m$ under exact selection: full correction ($\gamma = 1$) recovers $G_{\text{cen}}$ in one round, as Proposition 2 predicts, whereas partial correction leaves a gap. Appendix D repeats these checks on random instances of several sizes.

### 5.2 DNN EXPERIMENTS

We test whether trained networks approach $G_\star$ without correction and $G_{\text{cen}}$ with correction, and whether the proximal penalty improves agreement with these predictions.

#### 5.2.1 EXPERIMENTAL SETTING

**Datasets and models.** We use five tabular and five image regression datasets. The tabular datasets are Beijing Multi-Site Air Quality (Chen, 2017; Zhang et al., 2017) and four MuJoCo tasks (Swimmer, Hopper, HalfCheetah and Walker; Todorov et al., 2012) from JAT (Gallouédec et al., 2024). The image datasets are MPIIGaze (Zhang et al., 2015), 300W-LP (Zhu et al., 2016), dSprites (Matthey et al., 2017), 3D Shapes (Burgess & Kim, 2018) and CIFAR-geo, constructed from CIFAR-10 (Krizhevsky, 2009). We partition the data to give clients different target means and covariances (Appendix E.1). The backbones are ResMLP (Ma et al., 2025) for tabular data and ResNet-18 (He et al., 2016) for images.

**Training conditions.** Each client trains its head and private backbone using minibatch AdamW on (1). Here $H_m$ contains the backbone outputs, whose squared norms are penalized as in the UFM. We use $E = 4000$ local epochs per round to approach local optimality and average only the heads over eight rounds. Ordinary training uses $\rho = 0$; proximal training adds $\rho\|W - W_t\|_F^2/2$ with $\rho = 10^{-3}$ to approximate the selection rule. Each is run with and without $C_m$, giving four procedures, with five seeds $\{0, 1, 2, 3, 4\}$ per dataset (Appendix E.5).

**Evaluation.** We compute $G_\star$ and $G_{\text{cen}}$ from training-target moments, client weights and regularization, independently of the trained networks. Relative matrix error $e_F(G, G') = \|G - G'\|_F/\|G'\|_F$ measures agreement in scale and shape. Endpoint errors compare $G_8$ with $G_\star$ without

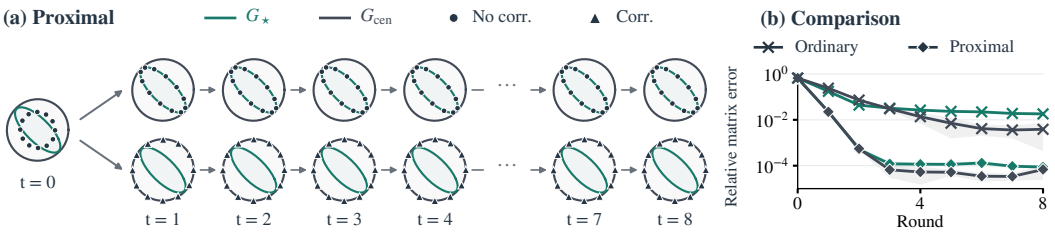


Figure 4: **Gram evolution on MPIIGaze.** (a) Seed-0 Gram matrices under proximal training. (b) Relative errors to $G_\star$ without correction (green) and to $G_{\text{cen}}$ with correction (graphite gray), with five-seed means $\pm 1$ standard deviation.

correction and with $G_{\text{cen}}$ under correction. For uncorrected runs, trajectory errors instead compare with (9) from the same initial Gram (Appendix F.1).

### 5.2.2 Experimental results

**Convergence toward the BW prediction.** Without correction, proximal training approaches $G_\star$ while remaining far from $G_{\text{cen}}$ (Figure 3). Across the 50 dataset–seed combinations, the median endpoint error to $G_\star$ is $2.90 \times 10^{-4}$, compared with $1.63 \times 10^{-2}$ for ordinary training. The proximal penalty improves endpoint agreement in every pair, and the proximal trajectories closely follow the predicted Gram iteration (Appendix F.1).

**Moment correction.** With correction, the shared Gram instead approaches $G_{\text{cen}}$ (Figure 3). Across the same 50 combinations, the median endpoint error is $7.08 \times 10^{-5}$ with the proximal penalty and $5.06 \times 10^{-3}$ without it (Appendix F.1). Figure 4 illustrates these different endpoints and the improvement from the proximal penalty on MPIIGaze.

**Additional results.** Clientwise target centering removes the mean term but leaves a gap, consistent with Proposition 1 (Appendix F.2). Proximal training agrees best with $G_\star$ at an intermediate penalty weight (Appendix F.4). Aligning trained heads to the broadcast also improves agreement with $G_\star$ and lowers the objective after averaging compared with the same unaligned heads (Appendix F.5). The correspondence persists across other architectures, larger training sets and, with sufficient local training, full-model averaging (Appendix F.6).

## 6 Discussion

**Limitations.** Our main analysis assumes that clients reach their local optima in every round, unlike typical federated training. It also assumes the fully active regime and does not cover rank-deficient optimal head Gram matrices. The UFM does not model features of unseen inputs, so our analysis provides no test-error guarantees.

**Perspectives.** To partially resolve the above limitations, we performed additional experiments on test error and an analysis of finite local training. Compared with ordinary training, proximal training yields comparable test error on the image datasets, while proximal training with moment correction yields a modest median reduction (Appendix F.7). For finite local training, we studied profiled updates, which minimize features and bias exactly before each head step. For small total step length $\eta E$, we established a locally attracting Gram fixed point near $G_{\text{within}}$, the fixed point for one local step per round, and quantify its first-order displacement (Appendix G). Numerical checks support the small-step predictions; a separate sweep over local step counts shows a gradual shift toward the BW barycenter (Appendix G.4).

**Conclusion.** Under the proposed selection rule, we characterized the UFM Gram dynamics through the BW barycenter, decomposed the gap to centralized training, and recovered the centralized optimal Gram in one round by moment correction. UFM and neural-network experiments support the predicted Gram behavior.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to suggest ideas for the proofs of our mathematical claims. Additionally, we used them to suggest improvements to the English writing, to check and adjust the LaTeX formatting, to search for related work beyond the studies already known to us so that the paper is positioned accurately, to search for related theoretical results, and to review and organize code. The authors reviewed and approved all revised text and verified all core theoretical results presented in this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

**Theory.** Section 3 defines the federated UFM, including the local and centralized objectives and the fully active conditions assumed by every result, and Section 4 states each result with its assumptions. The appendices give complete proofs: Appendix A derives the optimal head Gram matrix and proves Lemma 1 and Theorem 1, Appendix B proves Proposition 1, and Appendix C proves Proposition 2. Appendix G states and proves Theorem 2 on finite local steps.

**Experiments.** Appendix D describes how the UFM instances are generated and solved, gives the fixed instance of Figure 2 with its random seed, and repeats each numerical check on many random instances of several sizes. For the DNN experiments, Appendix E describes the datasets with their preprocessing, client partitions and licenses (Appendix E.1), the architectures, the federated training procedure with all hyperparameters (Appendix E.3), the evaluation (Appendix E.4), and all runs together with the compute used (Appendix E.5). The fully active condition holds in every reported DNN experiment (Appendix E.3), and the predictions $G_\star$ and $G_{\mathrm{cen}}$ are computed from the training-target moments, client weights and regularization parameters alone, independently of the trained networks. Each DNN experiment is repeated over random seeds: five seeds $\{0, 1, 2, 3, 4\}$ for the main experiments and at least the three seeds $\{0, 1, 2\}$ for each supplementary experiment (Appendix E.5). We report seed means with standard-deviation bands, medians over all dataset–seed combinations, and paired comparisons within each combination; Appendix F reports the results for every setting. The supplementary material contains code that reproduces Figure 2, the training, evaluation and data-preparation code with the configurations of all 200 runs behind Figures 3, 4, 9 and 10, and a Swimmer demonstration that retrains seed 0 of the four training procedures and compares the trained head Grams with $G_\star$ and $G_{\mathrm{cen}}$.

## REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. doi: 10.1137/100805741. URL https://doi.org/10.1137/100805741.

George Andriopoulos, Zixuan Dong, Li Guo, Zifan Zhao, and Keith Ross. The prevalence of neural collapse in neural multivariate regression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126417–126451. Curran Associates, Inc., 2024. doi: 10.52202/079017-4015. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4748b6b6ca49f04b6a8cfce1d5f9a70-Paper-Conference.pdf.

George Andriopoulos, Soyuj Jung Basnet, Juan Guevara, Li Guo, and Keith Ross. Neural multivariate regression: Qualitative insights from the unconstrained feature model, 2025. URL https://arxiv.org/abs/2505.09308.

Tina Behnia and Christos Thrampoulidis. Supervised contrastive representation learning: Landscape analysis with unconstrained features. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 575–580, 2024. doi: 10.1109/ISIT57864.2024.10619192.

Daniele Berardini, Vito Paolo Pastore, and Vittorio Murino. Distribution alignment for one-shot federated learning via optimal transport, 2026. URL https://arxiv.org/abs/2606.16655.

Rajendra Bhatia. *Operator Monotone and Operator Convex Functions*, pp. 112–151. Springer New York, New York, NY, 1997. ISBN 978-1-4612-0653-8. doi: 10.1007/978-1-4612-0653-8_5. URL https://doi.org/10.1007/978-1-4612-0653-8_5.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. ISSN 0723-0869. doi: https://doi.org/10.1016/j.exmath.2018.01.002. URL https://www.sciencedirect.com/science/article/pii/S0723086918300021.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Zachary Charles and Jakub Konečný. Convergence and accuracy trade-offs in federated learning and meta-learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2575–2583. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/charles21a.html.

Song Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5RK5G.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10572–10586. Curran Associates, Inc., 2022. doi: 10.52202/068431-0768. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/449590dfd5789cc7043f85f8bb7afa47-Paper-Conference.pdf.

Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 270–290. PMLR, 16–19 Aug 2022. URL https://proceedings.mlr.press/v145/e22b.html.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021. doi: 10.1073/pnas.2103091118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2103091118.

Quentin Gallouédec, Edward Beeching, Clément Romac, and Emmanuel Dellandréa. Jack of all trades, master of some, a multi-purpose transformer agent, 2024. URL https://arxiv.org/abs/2402.09844.

X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *Journal of Machine Learning Research*, 25(192):1–48, 2024. URL http://jmlr.org/papers/v25/23-1215.html.

Wanli Hong and Shuyang Ling. Beyond unconstrained features: Neural collapse for shallow neural networks with general data. *Journal of Machine Learning Research*, 27(82):1–46, 2026. URL http://jmlr.org/papers/v27/24-1429.html.

Jaehee Jang, Heoneok Ha, Dahuin Jung, and Sungroh Yoon. Fedclassavg: Local representation learning for personalized federated learning on heterogeneous neural networks. In *Proceedings of the 51st International Conference on Parallel Processing*, ICPP '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450397339. doi: 10.1145/3545008.3545073. URL https://doi.org/10.1145/3545008.3545073.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IPrzNbddXV.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28060–28094. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24ai.html.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5319–5329, October 2023.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations, 2020. URL https://arxiv.org/abs/2001.01523.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.011. URL https://www.sciencedirect.com/science/article/pii/S1063520321001123. Special Issue on Harmonic Analysis and Machine Learning.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5972–5984. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2f2b265625d76a6704b08093c652fd79-Paper.pdf.

Chuang Ma, Tomoyuki Obuchi, and Toshiyuki Tanaka. Neural collapse in cumulative link models for ordinal regression: An analysis with unconstrained feature model. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 102065–102112. Curran Associates, Inc., 2025. doi: 10.52202/085713-3415. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/93da65de953f738e7160ef1125e288d1-Paper-Conference.pdf.

Paul Mangold, Alain Oliviero Durmus, Aymeric Dieuleveut, Sergey Samsonov, and Eric Moulines. Refined analysis of constant step size federated averaging and federated richardson-romberg extrapolation. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 5023–5031. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/mangold25a.html.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, Jul 2022. ISSN 2730-5724. doi: 10.1007/s43670-022-00027-5. URL https://doi.org/10.1007/s43670-022-00027-5.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Reese Pathak and Martin J Wainwright. Fedsplit: an algorithmic framework for fast federated optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7057–7066. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4ebd440d99504722d80de606ea8507da-Paper.pdf.

Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9RUHPlladgh.

Mingjia Shi, Yuhao Zhou, Qing Ye, and Jiancheng Lv. Unconstrained feature model and its general geometric patterns in federated learning: Local subspace minority collapse. In Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, and Chaojie Li (eds.), *Neural Information Processing*, pp. 449–464, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-8132-8.

Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EXnIyMVTL8s.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22045–22055. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fb2697869f56484404c8ceee2985b01d-Paper.pdf.

Peter Súkeník, Marco Mondelli, and Christoph Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52991–53024. Curran Associates, Inc., 2023. doi: 10.52202/075280-2306. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a60c43ba078b723d3d517d28c50ded4c-Paper-Conference.pdf.

Peter Súkeník, Christoph Lampert, and Marco Mondelli. Neural collapse vs. low-rank bias: Is deep neural collapse really optimal? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 138250–138288. Curran Associates, Inc., 2024. doi: 10.52202/079017-4388. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f9c2ab8d429044e0c35bcece2ff6d123-Paper-Conference.pdf.

Yunfei Teng, Wenbo Gao, François Chalus, Anna E Choromanska, Donald Goldfarb, and Adrian Weller. Leader stochastic gradient descent for distributed training of deep learning models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fea16e782bc1b1240e4b3c797012e289-Paper.pdf.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27225–27238. Curran Associates, Inc., 2022. doi: 10.52202/068431-1974. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ae54ce310476218f26dd48c1626d5187-Paper-Conference.pdf.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21478–21505. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tirer22a.html.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34301–34329. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tirer23a.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BkluqlSFDS.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/564127c03caab942e503ee6f810f54fd-Paper.pdf.

Diyuan Wu and Marco Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67499–67536. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wu25u.html.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Zikai Xiao, Zihan Chen, Liyinglan Liu, YANG FENG, Joey Tianyi Zhou, Jian Wu, Wanlu Liu, Howard Yang, and Zuozhu Liu. Fedloge: Joint local and generic federated learning under long-tailed data. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 50357–50376, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/db174d373133dcc6bf83bc98e4b681f8-Paper-Conference.pdf.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 09 2017. ISSN 1364-5021. doi: 10.1098/rspa.2017.0457. URL https://doi.org/10.1098/rspa.2017.0457.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27179–27202. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22c.html.

Guogang Zhu, Xuefeng Liu, Jianwei Niu, Shaojie Tang, and Xinghao Wu. The other side of the coin: Unveiling the downsides of model aggregation in federated learning from a layer-peeled perspective, 2025. URL https://arxiv.org/abs/2502.03231.

Heng Zhu, Harsh Vardhan, and Arya Mazumdar. Effectiveness of distributed gradient descent with local steps for overparameterized models, 2026. URL https://arxiv.org/abs/2412.07971.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

Pedro C. Álvarez Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2016.04.045. URL https://www.sciencedirect.com/science/article/pii/S0022247X16300907.

## APPENDIX CONTENTS

Appendices A–C give the proofs of the main results. Appendices D–F report the UFM and DNN experiments. Appendix G collects supplementary analyses and their numerical checks.


**A From local optima to the BW limit** **17**
A.1 Profiling and the optimal head Gram matrix . . . 17
A.2 Unique selected head weights and the proximal limit . . . 17
A.3 The closed Gram recursion and its BW limit . . . 18

**B The gap to centralized training** **19**
B.1 The three positive-semidefinite terms . . . 19
B.2 The exact zero set . . . 19
B.3 BW variance and pairwise bounds . . . 19

**C Moment correction and the scope of recovery** **20**
C.1 A common corrected profile and recovery . . . 20
C.2 The two correction terms and required moments . . . 21
C.3 Predictions with local and shared biases . . . 21

**D UFM numerical experiments** **22**
D.1 The fixed instance of Figure 2 . . . 22
D.2 Selection by a vanishing proximal regularizer . . . 23
D.3 Convergence to the BW barycenter . . . 24
D.4 The three terms of the gap to centralized training . . . 24
D.5 One-round recovery by the moment correction . . . 26

**E Details of the DNN experimental setup** **27**
E.1 Datasets . . . 27
E.2 Models . . . 28
E.3 Federated training . . . 28
E.4 Evaluation . . . 29
E.5 Experiment overview . . . 29

**F Supplementary DNN experimental results** **31**
F.1 Selection and moment correction . . . 31
F.2 Removing the mean contribution . . . 31
F.3 Local epochs and the endpoint . . . 33
F.4 Dependence on the proximal weight . . . 34
F.5 Selection by alignment . . . 35
F.6 Architecture, data size and full-model averaging . . . 36
F.7 Test error . . . 37

**G Supplementary analysis** **38**
G.1 Finite-step dynamics of the profiled UFM . . . 38
G.2 Proof of Theorem 2 . . . 40
G.3 Dependence of the displacement on client covariances . . . 44
G.4 Numerical checks of finite-step dynamics . . . 46

# A FROM LOCAL OPTIMA TO THE BW LIMIT

This appendix gives the proofs for Sections 4.1–4.3. We use the fully active assumptions of Section 3, together with each result's additional hypotheses.

## A.1 PROFILING AND THE OPTIMAL HEAD GRAM MATRIX

We derive the profile and the local and centralized optimal Gram matrices used in Section 4.1. For a target matrix $Y \in \mathbb{R}^{C\times N}$, write

$$\mu = Y\mathbf{1}/N, \qquad \widetilde{Y} = Y - \mu\mathbf{1}^\top, \qquad T = \widetilde{Y}\widetilde{Y}^\top/N, \qquad A = WW^\top + \lambda_H I_C,$$

where $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector. At fixed $(W, b)$, the minimizer of (1) with respect to $H$ is unique thanks to the strict convexity and takes

$$H(W,b) = (W^\top W + \lambda_H I_P)^{-1} W^\top (Y - b\mathbf{1}^\top) = W^\top A^{-1}(Y - b\mathbf{1}^\top).$$

Substitution into the loss gives

$$\frac{\lambda_H}{2}\operatorname{tr}(TA^{-1}) + \frac{\lambda_H}{2}(b-\mu)^\top A^{-1}(b-\mu) + \frac{\lambda_W}{2}\|W\|_F^2 .$$

Since $A \succ 0$, the unique minimizing bias is $b = \mu$, giving $H^\star(W) = W^\top A^{-1}\widetilde{Y}$ and the profile (6). Writing $G = WW^\top$, we minimize

$$f_T(G) = \frac{\lambda_H}{2}\operatorname{tr}[T(G + \lambda_H I_C)^{-1}] + \frac{\lambda_W}{2}\operatorname{tr} G$$

over $G \succeq 0$, all representable since $P \geq C$. For $X, T \succ 0$ and nonzero symmetric $D$,

$$\frac{d^2}{dt^2}\operatorname{tr}[T(X+tD)^{-1}]\Big|_{t=0} = 2\operatorname{tr}(KMK) > 0,$$
$$K = X^{-1/2}DX^{-1/2}, \qquad M = X^{-1/2}TX^{-1/2} \succ 0.$$

Hence $f_T$ is strictly convex. Its interior stationary equation is

$$-\frac{\lambda_H}{2}(G+\lambda_H I_C)^{-1}T(G+\lambda_H I_C)^{-1} + \frac{\lambda_W}{2} I_C = 0, \qquad (G+\lambda_H I_C)^2 = \frac{\lambda_H}{\lambda_W}T.$$

Under the fully active condition $\lambda_{\min}(T) > \lambda_H\lambda_W$, its solution is $G = \phi(T) = \sqrt{\lambda_H/\lambda_W}\,T^{1/2} - \lambda_H I_C \succ 0$. For any $W$ satisfying $WW^\top = \phi(T)$, the triple $(W, \mu, H^\star(W))$ is a global minimizer of the full loss (Andriopoulos et al., 2024).

With $p_m = N_m/N$, (2) is the same loss on concatenated client data, whose moments are $(\mu_g, \Sigma_{\text{cen}})$. Thus the local and centralized optimal head Gram matrices are $G_m = \phi(\Sigma_m)$ and $G_{\text{cen}} = \phi(\Sigma_{\text{cen}})$.

## A.2 UNIQUE SELECTED HEAD WEIGHTS AND THE PROXIMAL LIMIT

The following lemma establishes the selection rule of Section 4.2. Taking $G_{\text{tar}} = G_m$ and $T = \Sigma_m$ proves Lemma 1; its loss bound follows from (15).

**Lemma 2** (Positive-definite rectangular projection and selection). *Let $G_{\text{tar}} \succ 0$ and let $W \in \mathbb{R}^{C\times P}$ have full row rank, with $G = WW^\top \succ 0$ and $P \geq C$. The unique minimizer of $\|U - W\|_F^2$ subject to $UU^\top = G_{\text{tar}}$ is*

$$\Pi_{G_{\text{tar}}}(W) = T_{G_{\text{tar}}}(G)W, \qquad T_{G_{\text{tar}}}(G) = G^{-1/2}(G^{1/2}G_{\text{tar}}G^{1/2})^{1/2}G^{-1/2}, \qquad (14)$$

*and its attained squared distance is $d^2_{\text{BW}}(G, G_{\text{tar}})$.*

*For the proximal-selection statement, let $Y \in \mathbb{R}^{C\times N}$ have mean $\mu$ and covariance $T$ as defined in Appendix A.1, and let $F(\cdot; T)$ be given by* (6)*, with $\lambda_H, \lambda_W > 0$. Assume that $\lambda_{\min}(T) > \lambda_H\lambda_W$ and $G_{\text{tar}} = \phi(T)$. For the fixed reference $W$, global minimizers of*

$$F(U;T) + \frac{\rho}{2}\|U - W\|_F^2$$

*exist for every $\rho > 0$. Any choice of these minimizers $U_\rho$ satisfies*

$$U_\rho \to \Pi_{G_{\mathrm{tar}}}(W) \qquad \text{as } \rho \downarrow 0.$$

*For the full UFM loss (1) formed from $Y$ with the same head-only penalty, every global minimizer $(U_\rho, b_\rho, H_\rho)$ satisfies $b_\rho = \mu$ and $H_\rho = H^\star(U_\rho)$, where $H^\star$ is defined in Appendix A.1.*

*Proof.* Write $U = G_{\mathrm{tar}}^{1/2} R$, where $RR^\top = I_C$. The squared distance differs by a constant from $-2\,\mathrm{tr}(RZ)$, where $Z = W^\top G_{\mathrm{tar}}^{1/2}$ has full column rank. Let $Z = \Phi D \Psi^\top$ be its thin singular value decomposition; all diagonal entries of $D$ are positive. Then $\mathrm{tr}(RZ) \le \mathrm{tr}\, D$. Equality forces every diagonal entry of the contraction $\Psi^\top R \Phi$ to equal one. Writing $\phi_j$ and $\psi_j$ for the columns of $\Phi$ and $\Psi$, this means $\psi_j^\top R \phi_j = 1$. Since $\|R\phi_j\| \le 1$ and $\|\psi_j\| = 1$, equality in Cauchy–Schwarz gives $R\phi_j = \psi_j$ for every $j$, hence $R\Phi = \Psi$. Complete $\Phi$ to an orthogonal matrix $[\Phi, \Phi_\perp]$ and write $R = \Psi\Phi^\top + Q\Phi_\perp^\top$. The row-orthonormality constraint gives $I_C + QQ^\top = I_C$, so $Q = 0$. Thus the optimizer is unique.

Since $(T_{G_{\mathrm{tar}}}(G)W)(T_{G_{\mathrm{tar}}}(G)W)^\top = T_{G_{\mathrm{tar}}}(G)\, G\, T_{G_{\mathrm{tar}}}(G) = G_{\mathrm{tar}}$ holds, $U = T_{G_{\mathrm{tar}}}(G)W$ is feasible, and attains the maximum:

$$\mathrm{tr}[UW^\top] = \mathrm{tr}[T_{G_{\mathrm{tar}}}(G)G] = \mathrm{tr}[(G^{1/2} G_{\mathrm{tar}} G^{1/2})^{1/2}] = \mathrm{tr}\, D.$$

To justify the last equality, set $X = G^{1/2} G_{\mathrm{tar}}^{1/2}$. Then

$$X^\top X = G_{\mathrm{tar}}^{1/2} G G_{\mathrm{tar}}^{1/2} = Z^\top Z = \Psi D^2 \Psi^\top, \qquad XX^\top = G^{1/2} G_{\mathrm{tar}} G^{1/2}.$$

Since $XX^\top$ and $X^\top X$ have the same eigenvalues, the eigenvalues of $G^{1/2} G_{\mathrm{tar}} G^{1/2}$ are the squared singular values of $Z$. Taking the positive square root therefore gives

$$\mathrm{tr}\Big[(G^{1/2} G_{\mathrm{tar}} G^{1/2})^{1/2}\Big] = \sum_{i=1}^{C} D_{ii} = \mathrm{tr}\, D.$$

Hence $T_{G_{\mathrm{tar}}}(G)W$ attains the trace upper bound and thus is the unique solution discussed above. Substituting this maximum into $\|U - W\|_F^2 = \mathrm{tr}\, G_{\mathrm{tar}} + \mathrm{tr}\, G - 2\,\mathrm{tr}(UW^\top)$ gives the BW distance formula.

For the proximal limit, put $F_\star = \min_U F(U;T)$ and $U^\dagger = \Pi_{G_{\mathrm{tar}}}(W)$. Coercivity from $\lambda_W \|U\|_F^2 / 2$ ensures a global minimizer $U_\rho$ for each $\rho > 0$. Comparison with $U^\dagger$ gives

$$0 \le F(U_\rho;T) - F_\star, \qquad F(U_\rho;T) - F_\star + \frac{\rho}{2}\|U_\rho - W\|_F^2 \le \frac{\rho}{2}\left\|U^\dagger - W\right\|_F^2. \tag{15}$$

Thus $\|U_\rho - W\|_F \le \left\|U^\dagger - W\right\|_F$ and $F(U_\rho;T) \to F_\star$. The heads $U_\rho$ thus lie in a fixed compact ball centered at $W$, so every sequence of $U_\rho$ with $\rho \downarrow 0$ has a convergent subsequence by the Bolzano–Weierstrass theorem. By continuity, every such subsequential limit minimizes $F(\cdot;T)$ and is no farther from $W$ than $U^\dagger$. Since $U^\dagger$ is the unique head closest to $W$ among the global minimizers of $F(\cdot;T)$, every such limit equals $U^\dagger$.

Since the proximal penalty depends only on the head, the conditional bias and feature solutions remain unchanged. Thus every global minimizer of the penalized full objective satisfies $b_\rho = \mu$ and $H_\rho = H^\star(U_\rho)$, as derived in Appendix A.1, and its original loss equals $F(U_\rho;T)$. Consequently, (15) gives

$$0 \le \mathcal{L}(U_\rho, b_\rho, H_\rho) - \min \mathcal{L} = F(U_\rho;T) - F_\star \le \frac{\rho}{2}\|U^\dagger - W\|_F^2.$$

The excess loss therefore tends to zero as $\rho \downarrow 0$. □

### A.3 The closed Gram recursion and its BW limit

We prove Theorem 1 by deriving the Gram recursion and checking the hypotheses of the BW barycenter convergence theorem. Set $T_m(G) := T_{G_m}(G)$ and $S(G) := \sum_m p_m T_m(G)$. By Lemma 2, we have the head update $W_{t+1} = \sum_m p_m \Pi_m(W_t) = S(G_t)W_t$, implying the following closed Gram recursion:

$$G_{t+1} = S(G_t) G_t S(G_t).$$

Substituting $T_m$ gives (9). Starting from $G_0 \succ 0$, induction gives $T_m(G_t) \succ 0$, $S(G_t) \succ 0$ and $G_{t+1} \succ 0$. The returned biases $\mu_m$ average to $b_{t+1} = \mu_g$.

**Convergence of the Gram recursion.** For positive weights summing to one and positive-definite $G_m$, the iteration (9) converges from every $G_0 \succ 0$ to the unique positive-definite BW barycenter satisfying (10) (Álvarez Esteban et al., 2016, Theorem 4.2 and Remark 4.3). Its variational characterization is the Gaussian barycenter characterization (Agueh & Carlier, 2011; Bhatia et al., 2019). Full activity supplies $G_m \succ 0$, so this proves Theorem 1.

If $G_1, \ldots, G_M$ commute with each other, $(\sum_m p_m G_m^{1/2})^2$ satisfies the fixed-point equation in their common eigenbasis, so uniqueness identifies it with $G_\star$.

# B THE GAP TO CENTRALIZED TRAINING

This appendix proves the gap characterization in Proposition 1 of Section 4.4. We establish the PSD decomposition, characterize equality, and relate the averaging term to BW distances.

## B.1 THE THREE POSITIVE-SEMIDEFINITE TERMS

The weighted scatter of the selected head weights supplies the averaging term in (11). The fixed-point equation in (10), conjugated by $G_\star^{-1/2}$, gives $S(G_\star) = I_C$. Hence any factor $W_\star$ of $G_\star$ satisfies $\sum_m p_m \Pi_m(W_\star) = W_\star$. More generally, the weighted scatter identity at any round gives

$$\sum_m p_m G_m - G_{t+1} = \sum_m p_m (\Pi_m(W_t) - W_{t+1})(\Pi_m(W_t) - W_{t+1})^\top \succeq 0. \tag{16}$$

Expand the weighted scatter using $\sum_m p_m \Pi_m(W_t) = W_{t+1}$ and $\Pi_m(W_t)\Pi_m(W_t)^\top = G_m$ to obtain this identity. At any factor $W_\star$ of $G_\star$, this scatter is $\mathcal{M}_A = \sum_m p_m G_m - G_\star$, independently of that factor; convergence of $W_t$ is not required.

Adding and subtracting $G_{\text{within}}$ and $\sum_m p_m G_m$ gives (11). For $\alpha = \sqrt{\lambda_H/\lambda_W}$,

$$\mathcal{M}_\mu = \alpha(\Sigma_{\text{cen}}^{1/2} - \Sigma_{\text{within}}^{1/2}) \succeq 0, \qquad \mathcal{M}_\Sigma = \alpha \left( \Sigma_{\text{within}}^{1/2} - \sum_m p_m \Sigma_m^{1/2} \right) \succeq 0.$$

Since $\Sigma_{\text{cen}} = \Sigma_{\text{within}} + \Sigma_\mu \succeq \Sigma_{\text{within}}$, square-root monotonicity gives $\mathcal{M}_\mu \succeq 0$. Operator concavity of $\phi$ gives its Jensen gap $\mathcal{M}_\Sigma \succeq 0$ (Bhatia, 1997, Theorems V.1.9 and V.2.5). Equation (16) makes $\mathcal{M}_A$ the PSD weighted scatter of $U_m^\star = \Pi_m(W_\star)$ around $W_\star$. This proves the Loewner chain of Proposition 1.

## B.2 THE EXACT ZERO SET

We next prove the equality condition in Proposition 1: the gap vanishes exactly when client means and covariances agree. A zero sum forces each PSD term to vanish. Square-root injectivity gives $\mathcal{M}_\mu = 0 \iff \Sigma_\mu = 0$. Since $\operatorname{tr} \Sigma_\mu = \sum_m p_m \|\mu_m - \mu_g\|^2$ with $p_m > 0$, all means then agree. Likewise $\mathcal{M}_A = 0$ gives $\sum_m p_m \|U_m^\star - W_\star\|_F^2 = 0$, so all $U_m^\star$, hence all $G_m$, agree. Injectivity of $\phi$ implies equal covariances. Conversely, equal means and covariances give $\Sigma_\mu = 0$ and $G_m = G_{\text{cen}} = G_\star$, making all three terms zero.

## B.3 BW VARIANCE AND PAIRWISE BOUNDS

Choose any factor $W_\star W_\star^\top = G_\star$ and set $U_m^\star = \Pi_m(W_\star)$. Taking the trace of

$$\mathcal{M}_A = \sum_m p_m (U_m^\star - W_\star)(U_m^\star - W_\star)^\top$$

and using $\operatorname{tr}(AA^\top) = \|A\|_F^2$, we obtain

$$\operatorname{tr} \mathcal{M}_A = \sum_m p_m \|U_m^\star - W_\star\|_F^2 .$$

Each $U_m^\star$ is the factor of $G_m$ closest to $W_\star$. Hence Lemma 2 gives

$$\|U_m^\star - W_\star\|_F^2 = d_{\text{BW}}^2(G_\star, G_m).$$

Substitution proves (12).

In addition to this, we prove the following pairwise bounds mentioned in Section 4.4:

$$D \le \operatorname{tr} \mathcal{M}_A \le 2D, \qquad D := \sum_{m<n} p_m p_n d^2_{\mathrm{BW}}(G_m, G_n). \tag{17}$$

For the lower bound, we first express the scatter around $W_\star$ in terms of pairwise distances. Expanding the squared distances and using $\sum_m p_m U^\star_m = W_\star$ gives

$$\begin{aligned}\sum_{m<n} p_m p_n \left\|U^\star_m - U^\star_n\right\|_F^2 &= \frac{1}{2}\sum_{m,n} p_m p_n \left\|U^\star_m - U^\star_n\right\|_F^2 \\ &= \sum_m p_m \left\|U^\star_m\right\|_F^2 - \left\|W_\star\right\|_F^2 \\ &= \operatorname{tr} \mathcal{M}_A.\end{aligned}$$

For each pair $(m, n)$, Lemma 2 gives

$$d^2_{\mathrm{BW}}(G_m, G_n) = \min_{UU^\top = G_m} \left\|U - U^\star_n\right\|_F^2 .$$

Since $U^\star_m U^{\star\top}_m = G_m$, the head $U^\star_m$ is a feasible choice in this minimization. Therefore,

$$d^2_{\mathrm{BW}}(G_m, G_n) \le \left\|U^\star_m - U^\star_n\right\|_F^2 .$$

Multiplying by $p_m p_n$ and summing over $m < n$ yields $D \le \operatorname{tr} \mathcal{M}_A$.

For the upper bound, define

$$V(X) := \sum_m p_m d^2_{\mathrm{BW}}(X, G_m).$$

By the barycenter characterization in (10), $G_\star$ is a global minimizer of $V$. Since each $G_n$ is positive definite and therefore feasible, we have $V(G_\star) \le V(G_n)$ for every $n$. Averaging these inequalities with weights $p_n$ gives

$$\operatorname{tr} \mathcal{M}_A = V(G_\star) \le \sum_n p_n V(G_n) = \sum_{n,m} p_n p_m d^2_{\mathrm{BW}}(G_n, G_m) = 2D.$$

The last equality follows because the terms with $m = n$ are zero, while each pair with $m < n$ is counted twice.

# C Moment correction and the scope of recovery

This appendix proves the recovery result in Proposition 2 of Section 4.4 and explains its scope. We identify the moments needed for the correction and compare the resulting predictions with centralized training.

## C.1 A common corrected profile and recovery

Since $C_m(W)$ depends only on $W$, minimizing over the bias and features for a fixed $W$ gives the same conditional solutions as in Appendix A.1:

$$b^\star_m(W) = \mu_m, \qquad H^\star_m(W) = W^\top A^{-1}(Y_m - \mu_m 1^\top), \qquad A = WW^\top + \lambda_H I_C.$$

Substituting these solutions into the corrected objective gives

$$\begin{aligned}\min_{b,H}\{\mathcal{L}_m(W, b, H) + C_m(W)\} &= F(W; \Sigma_m) + C_m(W) \\ &= \frac{\lambda_H}{2} \operatorname{tr}\left[(\Sigma_m + \Sigma_{\mathrm{cen}} - \Sigma_m)A^{-1}\right] + \frac{\lambda_W}{2}\left\|W\right\|_F^2 \\ &= F(W; \Sigma_{\mathrm{cen}}).\end{aligned}$$

Thus every client has the same profiled objective for the head.

By Appendix A.1, this common objective is minimized exactly by the heads satisfying

$$WW^\top = G_{\text{cen}} = \phi(\Sigma_{\text{cen}}) \succ 0.$$

Such heads exist because $P \geq C$. Pairing any such head with the conditional bias and feature solutions above therefore gives a global minimizer of each client's corrected full objective.

Now fix a common broadcast $W_t$ of full row rank. Under the selection rule, each client chooses the head closest to $W_t$ among the factors of $G_{\text{cen}}$. By Lemma 2, this head is unique, so

$$\widetilde{W}_{m,t} = \Pi_{G_{\text{cen}}}(W_t), \qquad \widetilde{b}_{m,t} = \mu_m \quad \text{for every } m.$$

Since $\sum_m p_m = 1$, server averaging gives

$$\begin{aligned} W_{t+1} &= \sum_m p_m \widetilde{W}_{m,t} = \Pi_{G_{\text{cen}}}(W_t), \\ G_{t+1} &= W_{t+1} W_{t+1}^\top = G_{\text{cen}}, \\ b_{t+1} &= \sum_m p_m \mu_m = \mu_g. \end{aligned}$$

This proves Proposition 2.

Finally, Appendix A.2 applies to the common profiled objective: any choice of global minimizers

$$U_\rho \in \arg\min_U \left\{ F(U; \Sigma_{\text{cen}}) + \frac{\rho}{2} \|U - W_t\|_F^2 \right\}$$

satisfies $U_\rho \to \Pi_{G_{\text{cen}}}(W_t)$ as $\rho \downarrow 0$. Thus the same selection can be imposed by globally minimizing the penalized objective for each $\rho > 0$ and taking this limit before server averaging.

### C.2 The two correction terms and required moments

Using $\Sigma_{\text{cen}} = \Sigma_{\text{within}} + \Sigma_\mu$, we write

$$C_m(W) = \frac{\lambda_H}{2} \operatorname{tr}(\Sigma_\mu A^{-1}) + \frac{\lambda_H}{2} \operatorname{tr}\left[(\Sigma_{\text{within}} - \Sigma_m) A^{-1}\right], \qquad A = WW^\top + \lambda_H I_C.$$

The first term restores the mean-scatter contribution removed by local centering. The second replaces $\Sigma_m$ by $\Sigma_{\text{within}}$ in each profiled objective; its weighted average is zero when evaluated at a common $W$.

Each client uploads $N_m$, $\mu_m$, and $\Sigma_m$ once. With $N = \sum_m N_m$, the server computes

$$\mu_g = \frac{1}{N} \sum_m N_m \mu_m, \qquad \Sigma_{\text{cen}} = \frac{1}{N} \sum_m N_m (\Sigma_m + \mu_m \mu_m^\top) - \mu_g \mu_g^\top,$$

and broadcasts $\Sigma_{\text{cen}}$ so that each client can evaluate $C_m(W)$. This requires $O(C^2)$ additional entries per client, or $O(MC^2)$ in total.

### C.3 Predictions with local and shared biases

Fix a common head $W$ with $G = WW^\top = G_{\text{cen}}$, and let $A = G + \lambda_H I_C$ and $d_m = \mu_m - \mu_g$. The corrected client centers its targets using $\mu_m$, whereas centralized training uses $\mu_g$. Their optimal features on client $m$'s samples are therefore

$$\begin{aligned} H_m^{\text{corr}} &= W^\top A^{-1} (Y_m - \mu_m \mathbf{1}^\top), \\ H_m^{\text{cen}} &= W^\top A^{-1} (Y_m - \mu_g \mathbf{1}^\top). \end{aligned}$$

Hence

$$H_m^{\text{corr}} - H_m^{\text{cen}} = -W^\top A^{-1} d_m \mathbf{1}^\top. \tag{18}$$

After aggregation, each client uses the shared bias $\mu_g$ while retaining its locally optimized features $H_m^{\text{corr}}$. By (18), its prediction difference from centralized training is

$$(W H_m^{\text{corr}} + \mu_g \mathbf{1}^\top) - (W H_m^{\text{cen}} + \mu_g \mathbf{1}^\top) = -G A^{-1} d_m \mathbf{1}^\top.$$

If the client instead retains its local bias $\mu_m$, the prediction difference becomes

$$(WH_m^{\mathrm{corr}} + \mu_m \mathbf{1}^\top) - (WH_m^{\mathrm{cen}} + \mu_g \mathbf{1}^\top) = \lambda_H A^{-1} d_m \mathbf{1}^\top.$$

The latter equality uses $I_C - GA^{-1} = \lambda_H A^{-1}$. Both prediction differences are nonzero whenever $\mu_m \neq \mu_g$. Thus the correction recovers the centralized head Gram matrix and aggregate bias, but need not reproduce centralized training predictions.

# D UFM NUMERICAL EXPERIMENTS

This appendix checks the results of Section 4 numerically under the fully active conditions of Section 3. Appendix D.1 specifies the fixed instance and the solvers behind Figure 2. Appendices D.2–D.5 repeat each check on random instances of several sizes, in the order of Section 4: selection by the proximal regularizer (Lemma 1), convergence to the BW barycenter (Theorem 1), the three terms of the gap (Proposition 1) and the moment correction (Proposition 2). All results of Section 4 hold for arbitrary positive weights $p_m$ summing to one, with $\mu_g$, $\Sigma_{\mathrm{within}}$ and $\Sigma_{\mathrm{cen}}$ formed with these weights. The instances in Appendices D.3–D.5 therefore draw the weights at random instead of setting $p_m = N_m/N$. Curves show medians over instances; in Figures 5 and 7, shaded bands show interquartile ranges. Appendix G reports the numerical checks of the finite-step analysis.

## D.1 THE FIXED INSTANCE OF FIGURE 2

Figure 2 illustrates the results of Section 4 on one fixed instance with $C = P = 2$, $M = 3$, $N_m = 256$, $p_m = 1/3$ and $\lambda_H = \lambda_W = 0.1$. Writing $R_\theta$ for planar rotation through $\theta$ degrees, the client moments and the initial Gram matrix are

$$\begin{aligned}
\Sigma_1 &= R_{-40}\,\mathrm{diag}(3, 0.11)R_{-40}^\top, & \mu_1 &= (0, -2.2)^\top,\\
\Sigma_2 &= R_{10}\,\mathrm{diag}(1.1, 0.45)R_{10}^\top, & \mu_2 &= (0, 0)^\top,\\
\Sigma_3 &= R_{55}\,\mathrm{diag}(2.4, 0.13)R_{55}^\top, & \mu_3 &= (0, 2.2)^\top,\\
G_0 &= R_{100}\,\mathrm{diag}(2.5, 0.09)R_{100}^\top, & W_0 &= G_0^{1/2}.
\end{aligned}$$

Each target matrix is generated by centering and whitening a Gaussian draw with random seed 0, then applying the prescribed mean and covariance. The empirical moments equal these values up to rounding, and concatenating the three equal-size matrices gives the same $\Sigma_{\mathrm{cen}}$ as (5). A Gram matrix $G$ is drawn as the ellipse $\{G^{1/2}u : \|u\|_2 = 1\}$, whose semiaxes are the square roots of its eigenvalues.

**Local optimization and selection.** Panel (a) follows client 1 during joint training of $(W, b, H)$ with Adam, a learning rate of $0.02$ and no proximal regularizer, starting from $W_0$, $b = \mu_1$ and the ridge-optimal $H$ at $W_0$. Its snapshots show the head Gram matrix after $E = 0, 20, 80$ and $240$ full-batch steps, together with client 1's optimal Gram matrix $G_1$, on the coordinate scale of panel (d). The other panels use the profiled objective (6), in which $(b, H)$ are minimized out in closed form. For $\rho > 0$, the clients minimize the penalized profiled objective with SciPy's `trust-exact` method, using the analytic gradient and Hessian. For a returned head $W_\rho$, we measure the relative Gram error and the relative selection error

$$e_G(\rho) = \frac{\left\|W_\rho W_\rho^\top - G_m\right\|_F}{\|G_m\|_F}, \qquad e_\Pi(\rho) = \frac{\|W_\rho - \Pi_m(W)\|_F}{\|\Pi_m(W)\|_F}.$$

For panel (b), the broadcast is fixed at $W_0$, and $\rho$ decreases over 11 logarithmically spaced values from $10^{-1}$ to $10^{-6}$; the first solve starts from $W_0$, and each later one from the previous solution. Panel (b) shows the largest of each error over the three clients; at $\rho = 10^{-6}$, they are $4.66 \times 10^{-6}$ and $2.48 \times 10^{-6}$.

**Convergence to the barycenter.** For panels (c) and (d), each client starts its local solve at the selected head $\Pi_m(W_t)$ and then minimizes the penalized objective numerically, for 18 rounds at each $\rho \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. Panel (c) shows $\rho = 10^{-1}$ and $10^{-3}$, together with exact selection and with unregularized L-BFGS started from the broadcast head in each round. After 18

rounds, the relative error of $G_t$ to $G_\star$ is $3.49 \times 10^{-2}$ at $\rho = 10^{-1}$, $2.55 \times 10^{-4}$ at $\rho = 10^{-3}$ and $3.11 \times 10^{-3}$ without the regularizer; with exact selection, it falls below $10^{-14}$ by round 16. Over all 303 profiled solves, the largest gradient norm is $4.14 \times 10^{-9}$. Panel (d) shows the shared Gram matrix at $\rho = 10^{-3}$ at rounds 0, 1, 2 and 18, on one coordinate scale. The reference $G_\star$ agrees to a relative error of $1.45 \times 10^{-15}$ with an independent minimization of the variational objective in (10) over Cholesky factors.

**Gap and correction.** Panel (e) evaluates the three terms of Proposition 1 for the original moments and for three modifications: equal means $\mu_m = \mu_g$, equal covariances $\Sigma_m = \Sigma_{\text{within}}$, and both. All other quantities remain fixed. With the original moments, the traces of $\mathcal{M}_\mu$, $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$ are 1.02, $2.90 \times 10^{-1}$ and $1.93 \times 10^{-1}$. Equalizing the means removes only $\mathcal{M}_\mu$, equalizing the covariances removes $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$, and equalizing both removes the gap, up to rounding. The computed terms also satisfy (11) and (12) and are positive semidefinite, up to rounding. Panel (f) applies the scaled correction $\gamma C_m$ under exact selection, for $\gamma \in \{0, 0.9, 0.99, 1\}$; client $m$ then has the optimal Gram matrix $\phi((1-\gamma)\Sigma_m + \gamma\Sigma_{\text{cen}})$. After one round, the relative error of $G_t$ to $G_{\text{cen}}$ is $8.08 \times 10^{-16}$ with the full correction, and $3.53 \times 10^{-2}$ and $3.46 \times 10^{-3}$ for $\gamma = 0.9$ and 0.99. Without correction, it settles at the relative gap of this instance, $5.78 \times 10^{-1}$.

### D.2 Selection by a vanishing proximal regularizer

We test Lemma 1(ii) on 128 random fully active instances for each $C \in \{2, 4, 8\}$ and $P \in \{C, 2C\}$. We set $\lambda_H = \lambda_W = 1$ and $N_m = 64$. The client means are standard normal, and the covariances have random orthogonal eigenbases with eigenvalues drawn uniformly from $[2, 5]$; the samples match these moments exactly. At a fixed broadcast $W$ with singular values drawn uniformly from $[0.7, 1.5]$, each client minimizes $\mathcal{L}_m + \frac{\rho}{2}\|W_m - W\|_F^2$ jointly over $(W_m, b, H)$, using SciPy's `trust-exact` solver with analytic derivatives. Each weight is warm-started from the solution at the next larger weight. Figure 5 shows the medians and interquartile ranges of the errors $e_G(\rho)$ and $e_\Pi(\rho)$ of Appendix D.1. For $\rho \le 10^{-1}$, both errors decrease in proportion to $\rho$. At the smallest weight shown, $\rho = 1.78 \times 10^{-6}$, the medians of $e_G$ lie between $5.15 \times 10^{-7}$ and $6.47 \times 10^{-7}$, and those of $e_\Pi$ between $2.70 \times 10^{-7}$ and $3.67 \times 10^{-7}$. At $\rho = 0$, the objective determines the Gram matrix but not the head. The left slot of each panel shows this difference: exact optimal heads, sampled on the optimal orbit, have Gram errors of at most $3.30 \times 10^{-15}$, whereas unregularized L-BFGS returns from six independent standard-normal head initializations per instance all reach the optimal Gram matrix, with Gram errors below $10^{-3}$, but their selection errors have a median of 1.42.

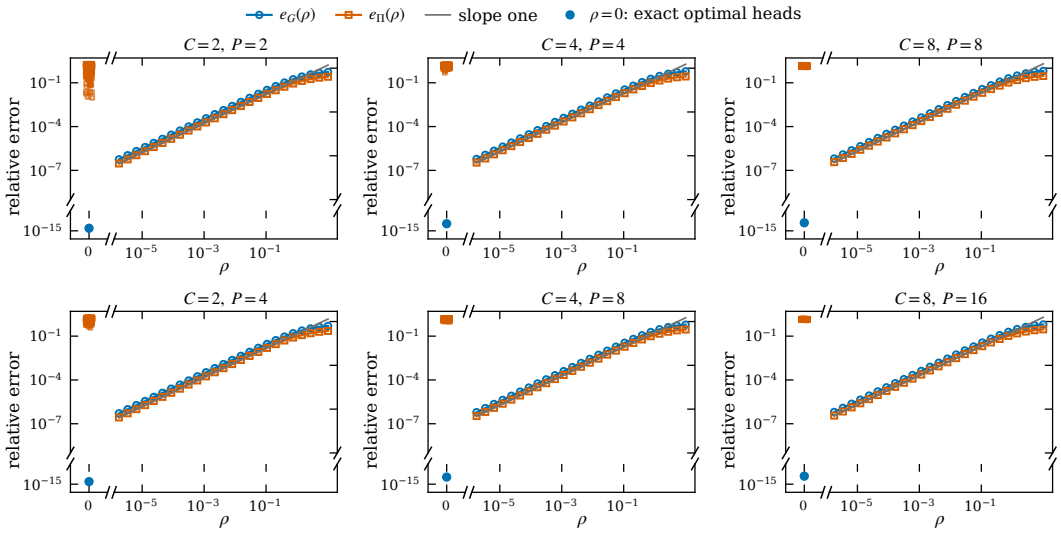


Figure 5: **Selection by a vanishing proximal regularizer (Lemma 1).** Relative Gram error $e_G(\rho)$ (circles) and selection error $e_\Pi(\rho)$ (squares) in six $(C, P)$ settings, with medians and interquartile ranges over 128 instances. The gray line has slope one. At $\rho = 0$, the circle marks exact optimal heads and the squares mark unregularized L-BFGS returns from random starts.

In networks trained with a fixed local budget, agreement with $G_\star$ is best at an intermediate weight (Appendix F.4). At $\rho = 10^{-5}$, finite local training leaves the heads farther from the selected heads, and at large $\rho$ the clients move away from their optima.

### D.3 Convergence to the BW barycenter

We test Theorem 1 on 16 random fully active instances for each $C \in \{2, 4, 8\}$ and $M \in \{3, 8\}$, whose client covariances do not commute. We set $P = 2C$ and $\lambda_H = \lambda_W = 1$. Each $G_m$ has a random orthogonal eigenbasis and eigenvalues drawn uniformly from $[0.4, 4]$, and we set $\Sigma_m = (G_m + I_C)^2$. The entries of the client means $\mu_m$ and the initial head $W_0$ are independent standard normal. Each client has $N_m = 2C$ samples whose empirical mean and covariance equal the prescribed $\mu_m$ and $\Sigma_m$, and the aggregation weights are drawn from a symmetric Dirichlet distribution with concentration 2 per client. Figure 6 shows the median relative error $\|G_t - G_\star\|_F / \|G_\star\|_F$ of the shared Gram matrix over 18 rounds, with $G_\star$ obtained from a separate run of (9) to convergence.

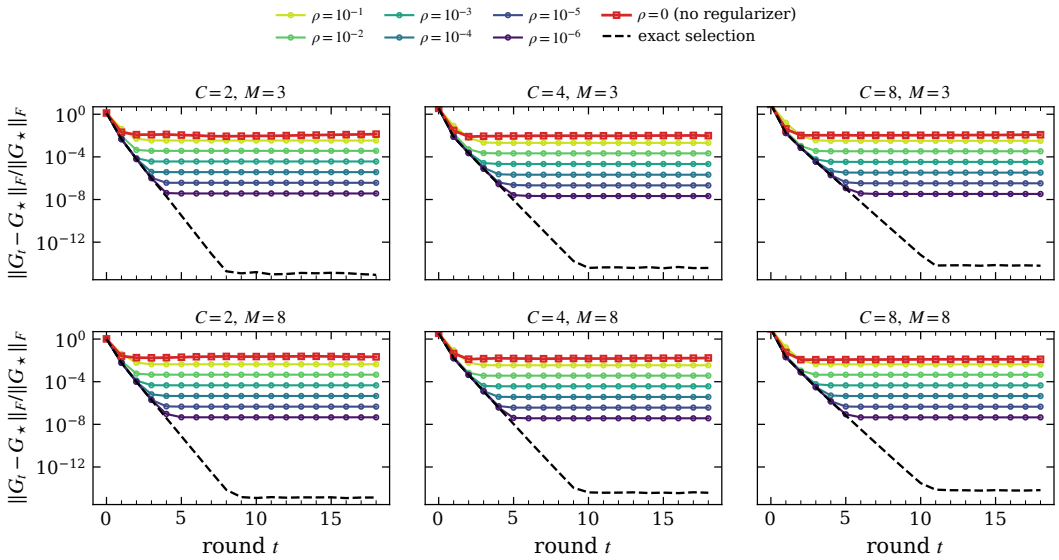


Figure 6: **Convergence to the BW barycenter (Theorem 1).** Median relative error of the shared Gram matrix to $G_\star$ over 16 instances in each of six problem sizes. Dashed: exact selection. Colored circles: proximal returns at $\rho = 10^{-1}, \ldots, 10^{-6}$. Red squares: unregularized joint training from the broadcast head.

With exact selection, the error decreases geometrically to below $10^{-14}$. For six weights $\rho \in \{10^{-1}, \ldots, 10^{-6}\}$, each client minimizes the penalized objective jointly over $(W_m, b, H)$ with SciPy's `trust-exact` solver and analytic derivatives, starting from $\Pi_m(W_t)$ and its ridge-optimal features. The error then levels off at a value roughly proportional to $\rho$. Unregularized joint training with Adam from the broadcast head levels off between $9.84 \times 10^{-3}$ and $2.13 \times 10^{-2}$. Trained networks show the same ordering (Table 4): proximal training closely follows the exact Gram iteration, while ordinary training ends near $G_\star$ with a larger error.

### D.4 The three terms of the gap to centralized training

We test the three predictions of Proposition 1: each term is positive semidefinite, the gap vanishes exactly when the client means and covariances agree, and $\operatorname{tr} \mathcal{M}_A$ equals the BW variance of the clients' optimal Gram matrices. By (17), this variance lies between $D$ and $2D$, where $D = \sum_{m<n} p_m p_n d^2_{\mathrm{BW}}(G_m, G_n)$ is the pairwise dispersion. Figure 7 reports these checks. All panels use $\lambda_H = \lambda_W = 0.1$ and standard-normal base client means.

**Which heterogeneity produces which term.** Panels (a) and (b) use two controlled sweeps with $C = 3$ and $M = 4$, each with the eight seeds $\{55, 56, 57, 58, 59, 60, 61, 62\}$ at every strength $s$.

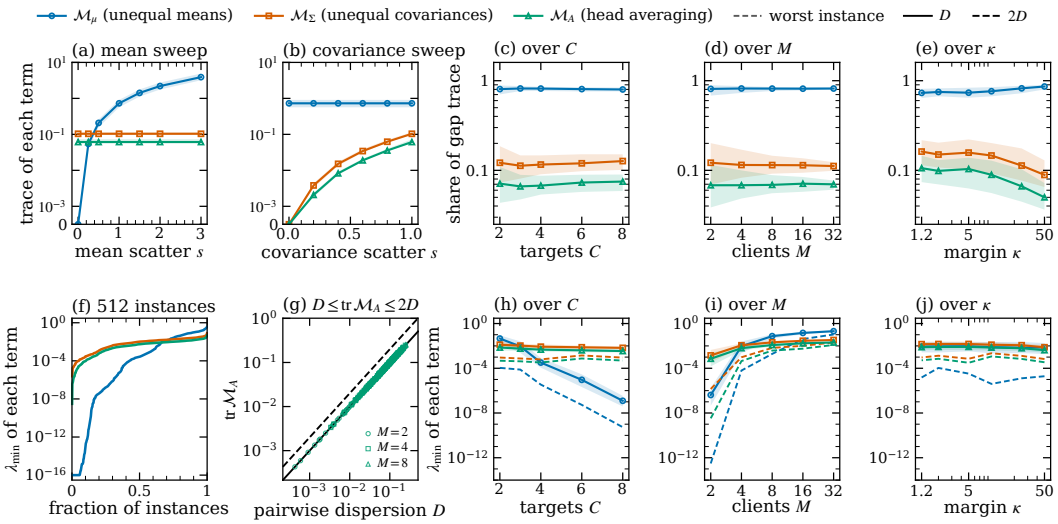


Figure 7: **The three terms of the gap $G_{\mathrm{cen}} - G_\star$ (Proposition 1).** Blue, orange and green identify $\mathcal{M}_\mu$, $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$ throughout; circles, squares and triangles mark the terms in (a–e) and (h–j), while in (g) they distinguish $M = 2, 4, 8$. Curves are medians and bands are interquartile ranges. (a, b) Trace of each term when the client means or the client covariances are varied with strength $s$ while the other is held fixed. The vertical axis is linear below $10^{-3}$ so that exact zeros are visible. (c–e) Share of the gap trace carried by each term against the number of targets $C$, the number of clients $M$ and a lower bound $\kappa$ on the fully active margin. (f) Smallest eigenvalue of each term over 512 instances, sorted; numerical zeros are drawn at $10^{-16}$. (g) $\mathrm{tr}\,\mathcal{M}_A$ against the pairwise dispersion $D$, with the bounds $D$ and $2D$ of (17). (h–j) Smallest eigenvalue of each term over the sweeps of (c–e); dashed curves show the worst of the 256 instances in each setting.

For each seed, the base covariances have random orthogonal eigenbases and eigenvalues uniform in $[0.3, 3]$. The weights are normalized independent uniform draws from $[0.5, 1.5]$. The mean sweep scales the separation of the client means by $s \in \{0, 0.25, 0.5, 1, 1.5, 2, 3\}$ and keeps the client covariances fixed. The term $\mathcal{M}_\mu$ then grows from exactly zero at $s = 0$ to a median trace of 3.90 at $s = 3$, while $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$ keep their median traces of $1.04 \times 10^{-1}$ and $6.12 \times 10^{-2}$. The covariance sweep moves each client covariance along $\Sigma_m(s) = \Sigma_{\mathrm{within}} + s(\Sigma_m - \Sigma_{\mathrm{within}})$ for $s \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, which keeps the client means and the weighted within-client covariance fixed. Now $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$ grow from zero, with traces below $2 \times 10^{-14}$ at $s = 0$, to the same two values, and $\mathcal{M}_\mu$ keeps its median trace of $7.30 \times 10^{-1}$. Equalizing the moments of the 512 instances described below gives the same pattern: equal means remove $\mathcal{M}_\mu$, equal covariances remove $\mathcal{M}_\Sigma$ and $\mathcal{M}_A$, and equalizing both removes the gap. Every removed term has trace below $2 \times 10^{-14}$.

**Positive semidefiniteness.** Panel (f) sorts the smallest eigenvalue of each term over 512 random fully active instances, with $M \in \{2, 4, 8\}$ clients, $C = P \in \{2, 3, 5, 8\}$, randomly rotated covariances with eigenvalues drawn uniformly from $[0.25, 3]$, and symmetric Dirichlet weights with concentration 2 per client. No eigenvalue is negative beyond rounding: the minimum over all 1536 values is $-2.42 \times 10^{-15}$. Sixty values of $\mathcal{M}_\mu$ lie below $10^{-12}$, and 29 of them are drawn at the floor of the panel. They belong to two-client instances with $C \ge 5$, where the mean scatter $\Sigma_\mu$ has rank one, so the trailing eigenvalues of $\phi(\Sigma_{\mathrm{within}} + \Sigma_\mu) - \phi(\Sigma_{\mathrm{within}})$ are small but, for generic client means, not exactly zero. Panels (h)–(j) repeat the test on 256 new instances per setting while one parameter varies: the number of targets $C$ with $M = 4$, the number of clients $M$ with $C = 3$, and the lower bound $\kappa$ on the fully active margin with $C = 3$ and $M = 4$. These sweeps use the weight distribution of panel (f); the $C$ and $M$ sweeps also use its covariance distribution. In the last sweep, the covariance eigenvalues are drawn uniformly from $[\kappa\lambda_H\lambda_W, 3]$, and $\kappa$ decreases from 50 to 1.2, toward the boundary of the fully active regime. The smallest eigenvalue of the worst instance stays positive in every setting; its minimum, $3.07 \times 10^{-13}$, occurs for $\mathcal{M}_\mu$ with two clients.

**The BW variance.** Panel (g) plots $\operatorname{tr} \mathcal{M}_A$ against $D$ for the 512 instances. Every point lies between the lines $D$ and $2D$. The ratio $\operatorname{tr} \mathcal{M}_A / D$ ranges over $[1.000, 1.006]$ and equals one for two clients. In that case the barycenter lies on the BW geodesic between $G_1$ and $G_2$, at distance $p_2 d_{\mathrm{BW}}(G_1, G_2)$ from $G_1$ (Agueh & Carlier, 2011), so $\operatorname{tr} \mathcal{M}_A = p_1 p_2 d^2_{\mathrm{BW}}(G_1, G_2) = D$ exactly. The identity (12) holds to a relative residual of at most $2.38 \times 10^{-12}$, and the weighted scatter of the selected heads, $\sum_m p_m (U_m^\star - W_\star)(U_m^\star - W_\star)^\top$, matches $\mathcal{M}_A = \sum_m p_m G_m - G_\star$ to a relative residual of at most $3.24 \times 10^{-12}$.

**Composition of the gap.** Panels (c)–(e) show the share of the gap trace carried by each term on the instances of panels (h)–(j), whose client means are standard normal. The term $\mathcal{M}_\mu$ carries 73–86% of the gap trace, $\mathcal{M}_\Sigma$ carries 9–16% and $\mathcal{M}_A$ carries 5–11% (medians). The relative gap $\|G_{\mathrm{cen}} - G_\star\|_F \,/\, \|G_{\mathrm{cen}}\|_F$ has medians between $1.68 \times 10^{-1}$ and $3.04 \times 10^{-1}$. In networks, removing the mean term by clientwise centering also leaves a positive gap (Appendix F.2), consistent with the remaining terms $\mathcal{M}_\Sigma + \mathcal{M}_A$.

### D.5 One-round recovery by the moment correction

We test Proposition 2 under exact selection on the six problem sizes of Figure 6, with 32 random fully active instances per size whose client covariances do not commute. Proposition 2 predicts that the corrected clients return identical heads, so that $G_t$ reaches $G_{\mathrm{cen}}$ after one round, while the local biases $\mu_m$ average to $\mu_g$. The eigenvalues of each $G_m$ are uniform in $[0.4, 4]$, the client means are standard normal, the weights are Dirichlet, $P = 2C$ and $\lambda_H = \lambda_W = 1$. Every client returns the selected head of its corrected objective, so each round is a closed-form map and involves no solver.

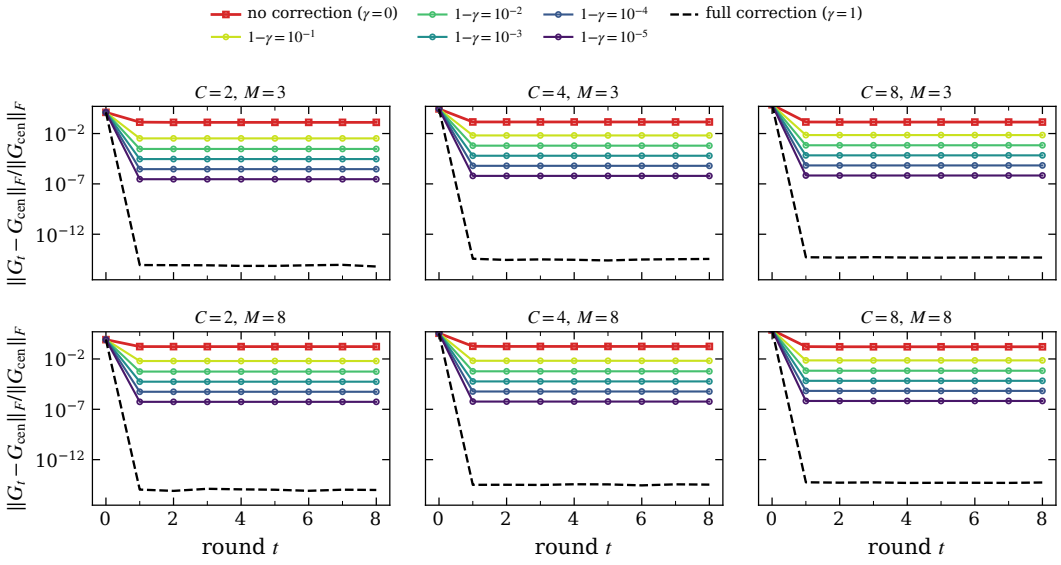


Figure 8: **One-round recovery of the centralized Gram matrix (Proposition 2).** Median relative error of the shared Gram matrix to $G_{\mathrm{cen}}$ over 32 instances in each of six problem sizes, under exact selection. Red squares: no correction; the plateau is the gap of Proposition 1. Dashed: the correction $C_m$ of (13). Colored: the scaled correction $\gamma C_m$, with deficit $1 - \gamma$ from $10^{-1}$ to $10^{-5}$.

Without the correction, $G_t$ converges to $G_\star$, and its relative distance to $G_{\mathrm{cen}}$ settles at the relative gap of the instance, with medians between $1.29 \times 10^{-1}$ and $1.75 \times 10^{-1}$. With the correction $C_m$, the relative distance after one round is at most $7.92 \times 10^{-15}$ over all 192 instances. The colored curves scale the correction to $\gamma C_m$. Because $F(W; T)$ is affine in $T$, client $m$ then minimizes $F(W; (1-\gamma)\Sigma_m + \gamma \Sigma_{\mathrm{cen}})$, and the remaining distance is proportional to the deficit $1 - \gamma$ from $10^{-1}$ to $10^{-5}$.

A second check solves the corrected profiled objective numerically, with L-BFGS and a positive proximal weight, in a fixed instance with $M = 4$ and $C = P = 3$. At the final weight $\rho = 10^{-5}$, one corrected round brings $G_t$ within a relative error of $9.13 \times 10^{-6}$ of $G_{\mathrm{cen}}$, and the four returned

heads lie within a Frobenius distance of $3.37 \times 10^{-7}$ of the first. With half of the correction, the returned heads differ by $1.66 \times 10^{-1}$, and $G_t$ misses $G_{\mathrm{cen}}$ by a relative error of $1.03 \times 10^{-1}$. Across 512 random instances, the identity $F(W; \Sigma_m) + C_m(W) = F(W; \Sigma_{\mathrm{cen}})$ holds with residuals of at most $3.30 \times 10^{-16}$ in value and $9.51 \times 10^{-15}$ in gradient. In trained networks, corrected proximal training approaches $G_{\mathrm{cen}}$ within the first few rounds (Appendix F.1).

# E Details of the DNN experimental setup

This appendix describes the datasets, models, federated training and evaluation used in Section 5 and Appendix F. Appendix E.5 defines the experimental settings and summarizes the main and supplementary experiments.

## E.1 Datasets

The DNN experiments use five tabular and five image datasets; Table 1 lists their input and target dimensions, numbers of clients and sample counts. For Beijing Multi-Site Air Quality (Chen, 2017; Zhang et al., 2017), we predict $PM_{2.5}$ and $NO_2$ concentrations from $PM_{10}$ concentration, temperature, pressure, dew-point temperature, precipitation, wind speed, month and hour. The other tabular datasets are four MuJoCo environments (Swimmer, Hopper, HalfCheetah and Walker; Todorov et al., 2012), in which, following Andriopoulos et al. (2024), actions are predicted from observations. For each environment we use the first 4,000 training transitions from JAT (Gallouédec et al., 2024).

The image datasets differ in how directly they pose multivariate regression. MPIIGaze (Zhang et al., 2015) and 300W-LP (Zhu et al., 2016) are multivariate regression datasets by design, for gaze and head-pose estimation: we predict gaze pitch and yaw from eye images and head pitch, yaw and roll from face images. dSprites (Matthey et al., 2017) and 3D Shapes (Burgess & Kim, 2018) are public datasets labeled with their generative factors. Regressing these factors from images has been studied before (Schott et al., 2022), but without a standard protocol, so we choose the regressed factors and the data split: horizontal position, vertical position and scale for dSprites, and object hue, scale and orientation for 3D Shapes. We construct CIFAR-geo as a regression task: we randomly shift and zoom each CIFAR-10 (Krizhevsky, 2009) image and predict the horizontal translation, vertical translation and log scale. MuJoCo observations and action targets retain their stored scale in both training and evaluation. The targets of the other datasets are centered and scaled with pooled training statistics, in training and at evaluation.

For the main experiments, we divide each dataset among clients using two partition schemes. For CIFAR-geo, we design the client groups to have different distributions of shift and zoom. For the other nine datasets, we project each training target onto the leading eigenvector of the pooled training-target second-moment matrix, sort the training samples by this projection and divide them into equally sized groups, one per client. Both schemes give the clients different target means and covariances.

The sample counts in Table 1 are totals over all clients. In CIFAR-geo, each client holds about 3,000 training samples, and the test samples are split evenly at random, 750 per client. In the other nine datasets, every client holds the same number of training samples. In the other four image datasets, test samples are assigned by their target projections to the nearest client training interval, giving test sets of different sizes. This target-based assignment provides an oracle evaluation protocol.

All datasets are used under their published terms. Beijing Multi-Site Air Quality is released under CC BY 4.0. Apache 2.0 covers dSprites, 3D Shapes and the JAT dataset, from which we take the MuJoCo transitions. MPIIGaze is released under CC BY-NC-SA 4.0 for non-commercial scientific use. 300W-LP is synthesized from 300-W, whose data are provided for research purposes only. The official CIFAR-10 page states no license and asks users to cite Krizhevsky (2009).

Table 1: **Datasets and sample counts** for the main DNN experiments. Test error is reported only for the image datasets (Appendix F.7).

| Dataset | Input dimension | Target dimension | Clients | Training samples | Test samples |
|---|---|---|---|---|---|
| *Tabular datasets* | | | | | |
| Beijing | 8 | 2 | 4 | 9,000 | — |
| Swimmer | 8 | 2 | 4 | 4,000 | — |
| Hopper | 11 | 3 | 4 | 4,000 | — |
| HalfCheetah | 17 | 6 | 4 | 4,000 | — |
| Walker | 17 | 6 | 4 | 4,000 | — |
| *Image datasets* | | | | | |
| MPIIGaze | $36 \times 60 \times 1$ | 2 | 8 | 12,000 | 3,000 |
| 300W-LP | $64 \times 64 \times 3$ | 3 | 8 | 12,000 | 3,000 |
| dSprites | $64 \times 64 \times 1$ | 3 | 8 | 12,000 | 3,000 |
| 3D Shapes | $64 \times 64 \times 3$ | 3 | 8 | 12,000 | 3,000 |
| CIFAR-geo | $40 \times 40 \times 3$ | 3 | 4 | 12,000 | 3,000 |

### E.2 Models

The main experiments use a residual MLP (ResMLP) for the tabular datasets and ResNet-18 (He et al., 2016) for the image datasets. The ResMLP follows the architecture of Ma et al. (2025), with three residual blocks of width 1024 and PReLU activations (He et al., 2015). All networks are trained from scratch and output a 512-dimensional feature through a linear layer with no activation.

Architecture comparisons also use ResNet-34 (He et al., 2016) and a residual CNN implemented as a ResNet-18 variant. The residual CNN has four stages of two basic blocks with 64, 128, 256 and 512 channels. It replaces ResNet-18's BatchNorm with eight-group GroupNorm (Wu & He, 2018) and its $7 \times 7$ stride-2 stem and max pooling with a $3 \times 3$ stride-1 stem. The first block of each of the last three stages downsamples with stride 2. For ResNet-18 and ResNet-34, the classification layer is replaced by a 512-to-512 linear feature projection after global average pooling. The residual CNN uses the same pooling and projection.

### E.3 Federated training

Unless stated otherwise, all DNN experiments share the round protocol, local objective and optimizer below; the four training procedures differ only in the terms added to the objective. In each round, all clients receive the shared head $(W_t, b_t)$, train it together with their private backbone, and upload the resulting head. The server averages these heads with weights $p_m = N_m/N$. Each backbone and any BatchNorm running statistics remain private throughout training. The main experiments start clients from the same randomly initialized backbone and use eight communication rounds and $E = 4000$ local epochs per round. Appendix F.3 varies $E$, and Appendix F.6 also studies full-model averaging.

Training uses a minibatch version of (1), with $H_m$ given by the network features, $\lambda_H = \lambda_W = 0.01$ and zero optimizer weight decay. The head bias is unregularized. The fully active condition $\lambda_{\min}(\Sigma_m) > \lambda_H \lambda_W$ holds for every client in the reported experiments: the smallest client covariance eigenvalue is $9.11 \times 10^{-3}$, compared with $\lambda_H \lambda_W = 10^{-4}$.

The main experiments compare four training procedures. Ordinary training optimizes this objective. Proximal training adds $\rho \|W - W_t\|_F^2/2$ with $\rho = 10^{-3}$ to approximate the selection rule of Lemma 1. Corrected ordinary training adds the head-only moment correction $C_m(W)$ in (13), computed from client and pooled training-target moments. Corrected proximal training adds both terms.

The optimizer is AdamW (Loshchilov & Hutter, 2019), reinitialized for each client's local training in every round. Its learning rate follows a cosine schedule from $10^{-3}$ to $10^{-5}$ within each round. We use shuffled minibatches of size 256 and clip the global gradient norm at 5.

### E.4 EVALUATION

We evaluate each run by the geometry of the shared head, the local fit and prediction error of the trained models, and display the results as curves over rounds and as ellipses.

**Head geometry and dynamics.** The theoretical targets $G_\star$ and $G_{\mathrm{cen}}$ are computed from training-target moments, client weights and regularization. The procedures without correction are compared with $G_\star$, and the corrected procedures with $G_{\mathrm{cen}}$. For Gram matrices $G$ and $G'$, we report relative matrix error and direction error:

$$e_F(G, G') = \frac{\|G - G'\|_F}{\|G'\|_F}, \qquad d(G, G') = \left\| \frac{G}{\|G\|_F} - \frac{G'}{\|G'\|_F} \right\|_F.$$

Endpoint errors compare the final shared Gram $G_8$ with these fixed targets. For training without correction, trajectory error compares $G_t$ with the exact recursion (9) started from the same $G_0$, taking the median $e_F$ over rounds 1–8 for each run. Table 4 reports medians of these run-level quantities. For these runs, the weighted distance from the uploaded heads to the selected heads is

$$\delta_t = \sum_m p_m \|\widetilde{W}_{m,t} - \Pi_m(W_t)\|_F, \qquad \delta_{\mathrm{last}} = \delta_7.$$

**Local fitting and prediction.** For training without correction, we evaluate the original objective (1) on each trained local model in evaluation mode before server averaging and report the relative gap $(\mathcal{L}_m - \mathcal{L}_m^\star)/\mathcal{L}_m^\star$. For the eigenvalues $s_i$ of $\Sigma_m$, the regression-UFM optimum (Andriopoulos et al., 2024) gives

$$\mathcal{L}_m^\star = \sum_{i: s_i > \lambda_H \lambda_W} \left( \sqrt{\lambda_H \lambda_W s_i} - \frac{\lambda_H \lambda_W}{2} \right) + \frac{1}{2} \sum_{i: s_i \le \lambda_H \lambda_W} s_i.$$

Prediction uses the final shared head with each client's own backbone. MSE averages squared errors over samples and output coordinates, weighting each client by the number of samples in its evaluated subset; it excludes all penalties and correction terms.

**Curves and geometric displays.** The round curves show means of per-seed errors with one sample standard deviation, using the same seeds across compared procedures. Geometric displays use seed 0 and draw the boundary of $\mathcal{E}(G) = \{G^{1/2}u : \|u\|_2 \le 1\}$, with common coordinates and equal axis units. HalfCheetah and Walker use the first three target coordinates of their six-dimensional Grams; all reported errors use the full matrices.

### E.5 EXPERIMENT OVERVIEW

We call a dataset paired with a network a setting. The main experiments use ten settings, pairing each tabular dataset with ResMLP and each image dataset with ResNet-18. Every setting is trained with the four procedures of Appendix E.3 and five seeds $\{0, 1, 2, 3, 4\}$; Table 2 shows this coverage.

The supplementary comparisons in Table 3 use three seeds $\{0, 1, 2\}$. Mean centering and the local-epoch and proximal-weight sweeps use the same ten settings. The architecture comparisons add the residual CNN and ResNet-34 on the five image datasets; the data-size comparison enlarges the training sets of CIFAR-geo and 3D Shapes; and the full-model comparison averages the backbones as well as the heads on the five tabular datasets. The prediction comparison uses the five image settings. Comparisons between training procedures pair runs by setting, seed, feature and head penalties, and local-epoch budget.

The alignment experiments in Appendix F.5, which Table 3 does not list, impose the selection rule by aligning each trained head to the broadcast head instead of adding the proximal regularizer. They compare aligned, proximal and ordinary training on six settings (Beijing, Swimmer, Hopper, HalfCheetah, 3D Shapes and CIFAR-geo), with five seeds per setting and three seeds for the moment-corrected comparison, and measure the effect of alignment on a single aggregation in 60 runs.

Table 2: **Main DNN experiments.** Coverage by setting and procedure in Appendix F.1.

| ✓: experiment repeated with five random seeds $\{0, 1, 2, 3, 4\}$ | | | | |
|---|---|---|---|---|
| | Without correction | | With moment correction | |
| Setting | Ordinary | Proximal | Ordinary | Proximal |
| *Tabular datasets* | | | | |
| Beijing | ✓ | ✓ | ✓ | ✓ |
| Swimmer | ✓ | ✓ | ✓ | ✓ |
| Hopper | ✓ | ✓ | ✓ | ✓ |
| HalfCheetah | ✓ | ✓ | ✓ | ✓ |
| Walker | ✓ | ✓ | ✓ | ✓ |
| *Image datasets* | | | | |
| MPIIGaze | ✓ | ✓ | ✓ | ✓ |
| 300W-LP | ✓ | ✓ | ✓ | ✓ |
| dSprites | ✓ | ✓ | ✓ | ✓ |
| 3D Shapes | ✓ | ✓ | ✓ | ✓ |
| CIFAR-geo | ✓ | ✓ | ✓ | ✓ |

The first five settings use ResMLP; the five image settings use ResNet-18. Each procedure uses eight rounds and $E = 4000$. Proximal training adds $\rho\|W - W_t\|_F^2/2$ with $\rho = 10^{-3}$. Moment correction adds $C_m$ from Proposition 2.

Table 3: **Supplementary DNN experiments.** Coverage by setting and condition; sweep columns give coverage at every parameter value.

| ✓: experiment repeated with three random seeds $\{0, 1, 2\}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | F.2 | F.3 | F.4 | F.6 | | | | | | | F.7 |
| Setting | Means | $E$ | $\rho$ | CNN | CNN $+C_m$ | RN34 | Data | Full $E = 4000$ | Full $E = 400$ | Full $+C_m$ | Predict |
| *Tabular datasets* | | | | | | | | | | | |
| Beijing | ✓ | ✓ | ✓ | — | — | — | — | ✓ | ✓ | ✓ | — |
| Swimmer | ✓ | ✓ | ✓ | — | — | — | — | ✓ | ✓ | ✓ | — |
| Hopper | ✓ | ✓ | ✓ | — | — | — | — | ✓ | ✓ | ✓ | — |
| HalfCheetah | ✓ | ✓ | ✓ | — | — | — | — | ✓ | ✓ | ✓ | — |
| Walker | ✓ | ✓ | ✓ | — | — | — | — | ✓ | ✓ | ✓ | — |
| *Image datasets* | | | | | | | | | | | |
| MPIIGaze | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | — | — | — | — | ✓ |
| 300W-LP | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | — | — | — | — | ✓ |
| dSprites | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | — | — | — | — | ✓ |
| 3D Shapes | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | — | — | — | ✓ |
| CIFAR-geo | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | — | — | — | ✓ |

An em dash marks a setting outside that comparison. Means: clientwise target centering. $E$: seven local-epoch budgets. $\rho$: seven proximal weights. CNN and RN34: residual CNN and ResNet-34. CNN + $C_m$: residual CNN with moment correction at $E = 4000$. Data: larger training sets with $E = 1000$. Full: backbone and head averaging at the stated budget. Full + $C_m$: full-model averaging with moment correction at $E = 4000$. Predict: training and test MSE for ordinary, proximal and corrected proximal training.

**Compute.** Computation used NVIDIA GH200 GPUs (GH200 120GB model; 97,871 MiB of visible device memory). Each run used one GPU, with up to eight independent runs sharing a GPU. We estimate 1,415 occupied GPU-hours for the 764 successful runs underlying the reported DNN results by merging their execution intervals on each GPU, including intervals shared with other runs. Experiment configurations, manifests, results and figure-source hashes are retained in the experimental records.

# F Supplementary DNN experimental results

This appendix reports the DNN comparisons behind Section 5 and their extensions. Appendix F.1 compares ordinary and proximal training with and without moment correction. Appendices F.2–F.4 test the gap decomposition and the dependence on the local budget and the proximal weight, and Appendix F.5 imposes the selection rule by alignment instead of the proximal regularizer. Appendix F.6 extends the comparison to other architectures, larger training sets and full-model averaging, and Appendix F.7 examines test error, which lies outside the UFM analysis. Appendix E.5 lists the settings and seeds of each comparison.

## F.1 Selection and moment correction

We first compare ordinary and proximal training across all main experiments, measuring agreement with $G_\star$ without correction and with $G_{\mathrm{cen}}$ under moment correction. Table 4 summarizes these comparisons. Without correction, the proximal regularizer improves endpoint agreement with $G_\star$ in every pair. It also reduces the median trajectory error relative to the exact iteration and the median distance to the selected heads.

Table 4: **Selection and moment correction.** Medians over the same 50 dataset–seed combinations. Without correction, endpoint errors compare with $G_\star$ and trajectory errors with the exact iteration from each run's $G_0$. All corrected errors compare with $G_{\mathrm{cen}}$. Bold marks the smaller value in each column.

| | Without correction | | | With moment correction | | |
|---|---|---|---|---|---|---|
| Training | Endpoint $e_F$ | Trajectory $e_F$ | $\delta_{\mathrm{last}}$ | Round 1 $d$ | Round 1 $e_F$ | Round 8 $e_F$ |
| Ordinary | $1.63 \times 10^{-2}$ | $3.44 \times 10^{-2}$ | $1.85 \times 10^{-1}$ | $8.35 \times 10^{-2}$ | $3.08 \times 10^{-1}$ | $5.06 \times 10^{-3}$ |
| Proximal | $\mathbf{2.90 \times 10^{-4}}$ | $\mathbf{3.39 \times 10^{-4}}$ | $\mathbf{4.47 \times 10^{-3}}$ | $\mathbf{4.21 \times 10^{-3}}$ | $\mathbf{2.29 \times 10^{-2}}$ | $\mathbf{7.08 \times 10^{-5}}$ |

With moment correction, proximal training reaches a smaller residual to $G_{\mathrm{cen}}$ in the first round and improves further over subsequent rounds. Figures 3 and 9 show these dynamics for all ten settings under proximal training. Appendix D.5 reports the corresponding UFM check under exact selection.

Figure 10 gives the ordinary-training counterpart to Figures 3 and 9, using the same seeds, axis limits and geometric scales. Figure 11 directly compares ordinary and proximal training against $G_\star$ without correction and against $G_{\mathrm{cen}}$ with correction.

## F.2 Removing the mean contribution

To test the gap decomposition of Proposition 1, we remove the between-client mean term by centering each client's targets and measure the remaining gap under ordinary training. Centering retains each client's covariance. We evaluate this change on all ten settings, using ResMLP for the tabular datasets and ResNet-18 for the image datasets, with three seeds per setting. The pooled reference then becomes $G_{\mathrm{within}}$, while $G_\star$ remains unchanged. In all 30 dataset–seed combinations, $G_{\mathrm{within}} - G_8$ is positive definite, consistent with a remaining covariance and averaging gap. The median direction error to $G_\star$ is $9.48 \times 10^{-3}$, and the median ratio of the observed gap trace to $\mathrm{tr}(\mathcal{M}_\Sigma + \mathcal{M}_A)$ is $1.19$ (Figure 12).

Centering removes only the mean term, so the covariance and averaging terms remain together in this comparison. Two of the 30 runs have a direction error to $G_\star$ above $0.05$, and in four runs this error differs from that of the uncentered control by more than $0.02$.

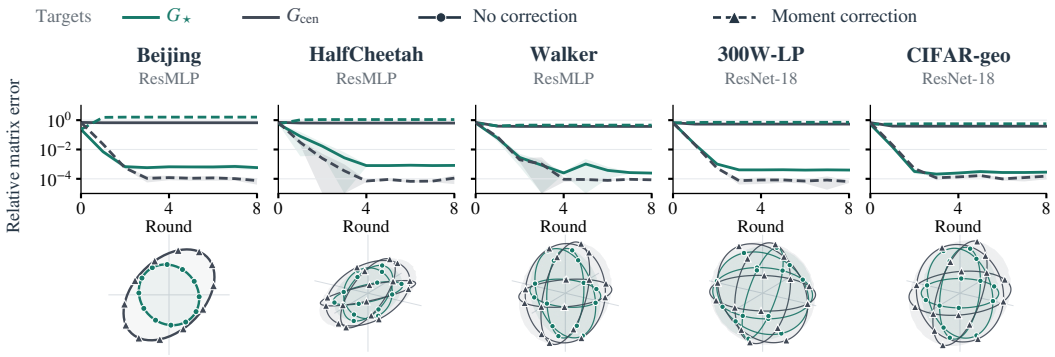


Figure 9: **Proximal training in the other five settings.** The design matches Figure 3; six-dimensional settings use coordinates 1–3. All runs use $\rho = 10^{-3}$.

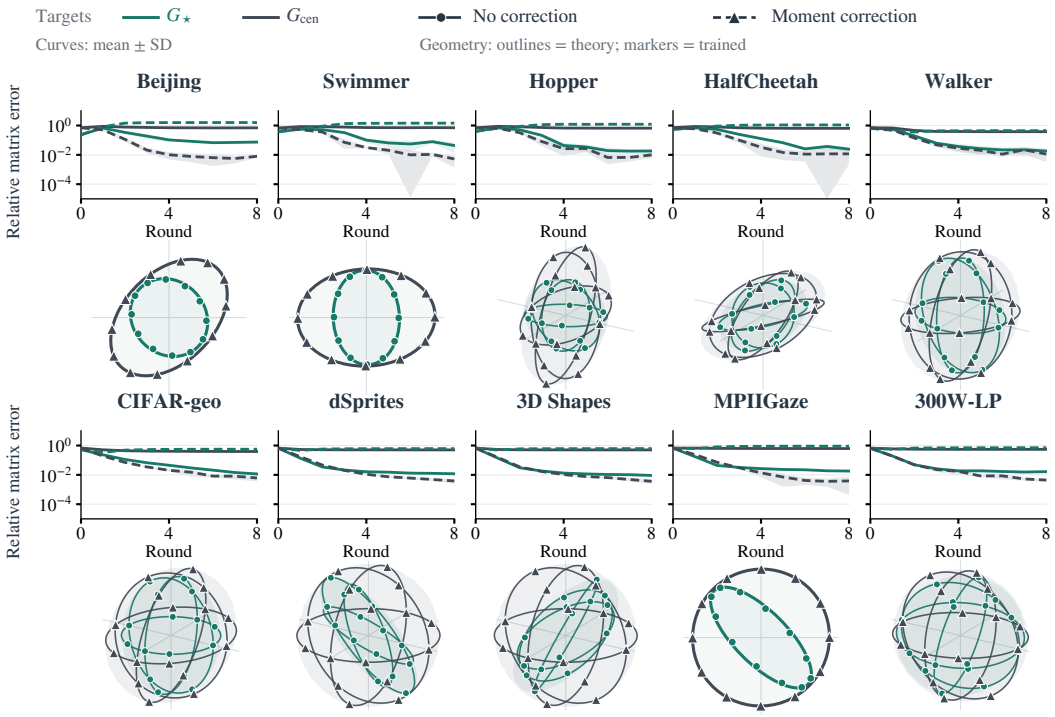


Figure 10: **Ordinary training with and without moment correction.** The design matches Figures 3 and 9, with $\rho = 0$. Curves show seed means $\pm 1$ standard deviation; geometry uses seed 0. Both variants retain the feature and head penalties.

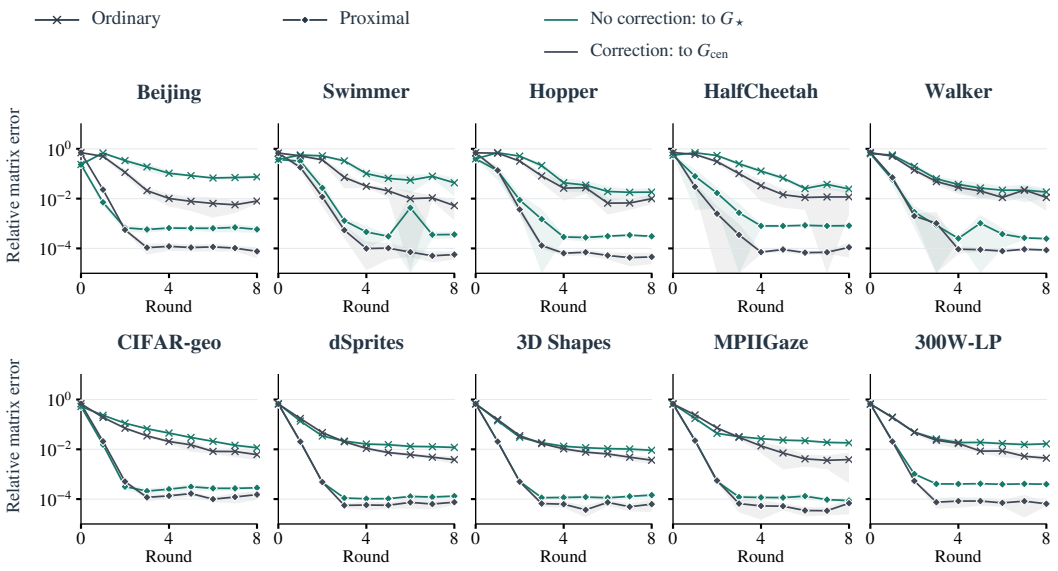


Figure 11: **Direct comparison of ordinary and proximal training.** Green curves measure training without correction against $G_\star$; graphite gray curves measure training with moment correction against $G_{\text{cen}}$. Within each color, crosses and filled diamonds denote ordinary and proximal training; both use solid lines. Curves use the same five seeds per setting as Figures 3 and 10, with means and one-standard-deviation bands. All panels share their axis limits.

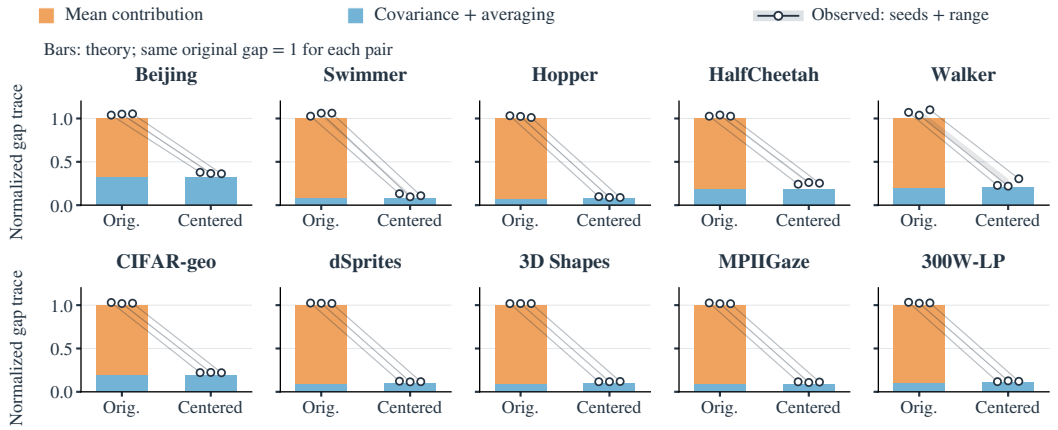


Figure 12: **Removing the mean contribution leaves a residual gap.** Stacked bars separate the predicted mean contribution $\mathrm{tr}(\mathcal{M}_\mu)$ from $\mathrm{tr}(\mathcal{M}_\Sigma + \mathcal{M}_A)$ before and after centering each client's targets. Both conditions use the original theoretical gap trace as denominator, computed per seed before averaging the components. Open markers show all three paired seeds, joined by thin lines; pale shading spans their range. Their observed gaps are $\mathrm{tr}(G_{\text{cen}} - G_8)$ before centering and $\mathrm{tr}(G_{\text{within}} - G_8)$ afterward, on the same scale.

### F.3 Local epochs and the endpoint

To see how close ordinary training comes to $G_\star$ as local training grows, we vary the number of local epochs $E$ per round. The local-budget sweep covers all ten settings with $E \in \{50, 100, 200, 400, 1000, 2000, 4000\}$ and seeds $\{0, 1, 2\}$, keeping the other training parameters fixed. We compare the direction error of the server Gram $G_8$ to $G_\star$ across local training budgets (Figure 13). The median over setting-level medians is $4.57 \times 10^{-1}$ at $E = 50$, $1.76 \times 10^{-1}$ at $E = 400$, $4.33 \times 10^{-2}$ at $E = 1000$ and $1.39 \times 10^{-2}$ at $E = 4000$. The first tested budget with

median direction error below 0.1 is $E = 400$ for Beijing and CIFAR-geo, $E = 1000$ for Walker and the other four image settings, and $E = 2000$ for Swimmer, Hopper and HalfCheetah. All budgets use eight rounds, so total local training grows with $E$, by a factor of 80 from $E = 50$ to $E = 4000$. A similar qualitative trend toward the BW prediction is observed in the profiled UFM experiments in Appendix G.4, although the update rules and evaluation protocols differ.

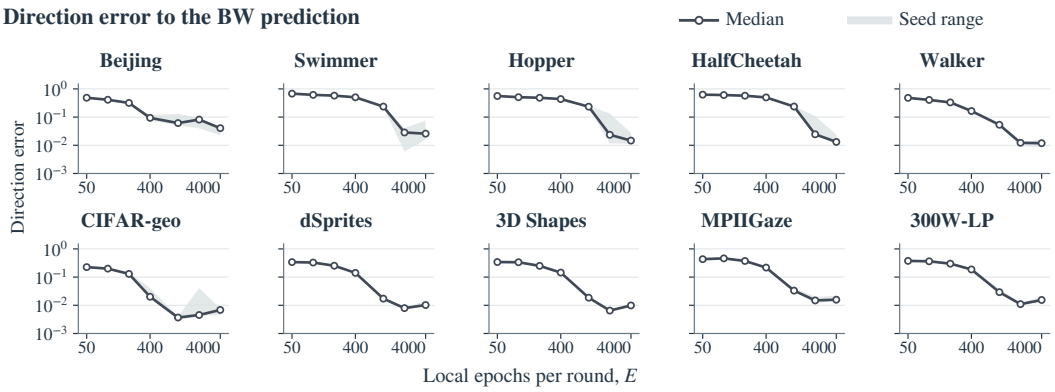


Figure 13: **Local epochs and the endpoint.** Direction error $d(G_8, G_\star)$ versus local epochs $E$, with shared logarithmic axes. Curves are medians over seeds $\{0, 1, 2\}$ and bands span these seeds. Top row: ResMLP; bottom row: ResNet-18.

### F.4 Dependence on the proximal weight

To see how the proximal weight affects agreement with the selection rule and with the local optima, we vary $\rho$ and measure the endpoint error, the distance to the selected heads and the relative original-objective gap. The weight sweep uses all ten settings with seeds $\{0, 1, 2\}$, giving the same 30 dataset–seed combinations at every weight. It compares $\rho = 0$ with $\rho \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and every run uses eight rounds and $E = 4000$ local epochs.

At $\rho = 10^{-3}$, the distance to the selected heads and the endpoint relative matrix error $e_F(G_8, G_\star)$ both decrease relative to $\rho = 0$ in all 30 dataset–seed combinations. The median final distance $\delta_{\text{last}}$ falls from $1.91 \times 10^{-1}$ to $5.33 \times 10^{-3}$, while the median relative original-objective gap rises from $1.43 \times 10^{-5}$ to $4.05 \times 10^{-5}$. Increasing $\rho$ further moves the clients away from their unpenalized optima and worsens endpoint agreement (Table 5, Figure 14). At $\rho = 10^{-5}$, the median $e_F$ is $7.38 \times 10^{-3}$, compared with $1.71 \times 10^{-2}$ at $\rho = 0$ and $3.08 \times 10^{-4}$ at $\rho = 10^{-3}$. In Lemma 1, the selected head arises from minimizing each penalized objective globally and then letting $\rho \downarrow 0$, whereas with finite local training a very small weight has a weaker effect. The medians of $\delta_{\text{last}}$ and $e_F$ are therefore U-shaped in $\rho$, with minima at $\rho = 10^{-3}$ (Table 5).

Table 5: **Proximal weight and local fitting.** Medians at the final round over the same 30 dataset–seed combinations at each weight. Relative gaps use the original loss without the proximal regularizer and first take the median across clients. Bold marks the smallest value in each column.

| Proximal weight $\rho$ | Distance to selected heads $\delta_{\text{last}}$ | Relative original-objective gap | $e_F(G_8, G_\star)$ |
|---|---|---|---|
| 0 | $1.91 \times 10^{-1}$ | $\mathbf{1.43 \times 10^{-5}}$ | $1.71 \times 10^{-2}$ |
| $10^{-5}$ | $1.19 \times 10^{-1}$ | $1.49 \times 10^{-5}$ | $7.38 \times 10^{-3}$ |
| $10^{-4}$ | $3.73 \times 10^{-2}$ | $1.53 \times 10^{-5}$ | $9.14 \times 10^{-4}$ |
| $10^{-3}$ | $\mathbf{5.33 \times 10^{-3}}$ | $4.05 \times 10^{-5}$ | $\mathbf{3.08 \times 10^{-4}}$ |
| $10^{-2}$ | $3.31 \times 10^{-2}$ | $1.40 \times 10^{-3}$ | $4.17 \times 10^{-3}$ |
| $10^{-1}$ | $1.32 \times 10^{-1}$ | $1.81 \times 10^{-2}$ | $4.26 \times 10^{-2}$ |
| 1 | $3.56 \times 10^{-1}$ | $1.57 \times 10^{-1}$ | $3.80 \times 10^{-1}$ |

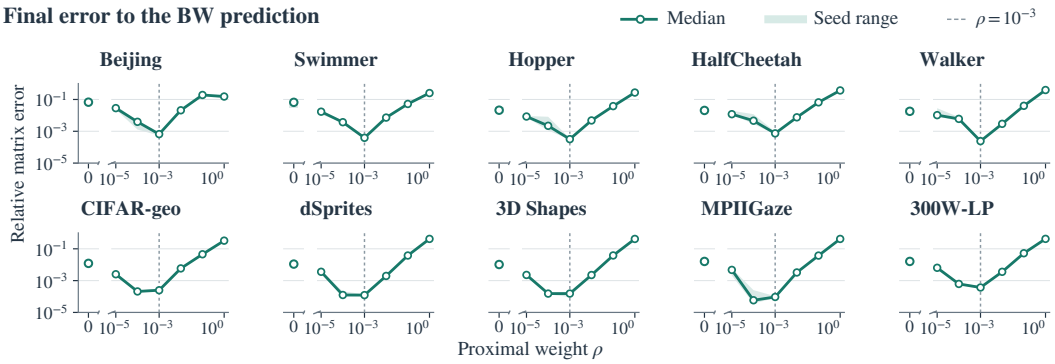


Figure 14: **Weight of the proximal regularizer.** Endpoint relative matrix error $e_F(G_8, G_\star)$ with $E = 4000$. In each setting, curves are medians over seeds $\{0, 1, 2\}$ and pale bands span these seeds. The zero-weight baseline is separate from the positive logarithmic axis; dashed guides mark $\rho = 10^{-3}$.

### F.5 Selection by alignment

Besides the proximal regularizer, the selection rule can be imposed by aligning each trained head to the broadcast head. We use this second implementation to examine (1) whether the endpoint again reaches $G_\star$, (2) whether moment correction still recovers $G_{\mathrm{cen}}$, and (3) how much averaging heads in mismatched coordinates raises the training objective. Each round has two stages: local training followed by an orthogonal alignment to the broadcast head. Local training uses the objective in Appendix E.3 with $\rho = 0$, either with or without moment correction. Let $\widehat{W}_{m,t}$ and $\widehat{h}_{m,t}(x)$ be the resulting head and feature vector. Client $m$ computes

$$Q_{m,t} \in \operatorname*{arg\,min}_{Q \in \mathbb{R}^{P \times P}:\, Q^\top Q = I_P} \|\widehat{W}_{m,t} Q - W_t\|_F^2,$$
$$\widetilde{W}_{m,t} = \widehat{W}_{m,t} Q_{m,t}, \qquad \widetilde{h}_{m,t}(x) = Q_{m,t}^\top \widehat{h}_{m,t}(x).$$

The client applies the feature transformation to its final linear feature layer, including its bias, and retains that layer for the next round. The head bias is unchanged by alignment. This joint transformation preserves the local predictions, head Gram matrix and feature and head penalties. The server then averages the aligned heads and their biases with weights $p_m = N_m/N$.

The implementation computes the aligned head from $C \times C$ Gram matrices and obtains $Q_{m,t}$ by SVD-based orthogonal completion. When both the returned and broadcast heads have full row rank, the aligned head is the unique head with the observed Gram matrix that is closest to $W_t$ in Frobenius norm, by the projection formula in Appendix A.2. When local training reaches the UFM-optimal Gram $G_m$, this gives the selected head $\Pi_m(W_t)$ in Lemma 1(i).

**Geometry over communication rounds.** We compare ordinary, proximal ($\rho = 10^{-3}$) and aligned training on six settings: Beijing, Swimmer, Hopper and HalfCheetah with ResMLP, and 3D Shapes and CIFAR-geo with ResNet-18. Each setting uses all five seeds $\{0, 1, 2, 3, 4\}$, giving 30 matched comparisons. Runs use eight rounds and $E = 4000$ local epochs, with the other training parameters from Appendix E.3. Alignment reduces the distance to the selected heads, the trajectory error and the endpoint error to $G_\star$ in all 30 comparisons with ordinary training, and brings the median endpoint error to $1.10 \times 10^{-4}$, compared with $3.24 \times 10^{-4}$ under proximal training and $1.86 \times 10^{-2}$ under ordinary training (Table 6). Relative to proximal training, alignment reduces trajectory error in all 30 comparisons and endpoint error to $G_\star$ in 29.

The moment-corrected comparison uses the same six settings with seeds $\{0, 1, 2\}$, giving 18 matched comparisons. Under exact UFM optimization, the corrected local optima share $G_{\mathrm{cen}}$, and alignment to the common broadcast selects the same head at every client, which yields the one-round recovery in Proposition 2. **In the trained networks, corrected alignment reaches a median relative error of $8.38 \times 10^{-5}$ to $G_{\mathrm{cen}}$ after the first round, whereas corrected proximal training**

**needs three rounds to reach a comparable error** ($2.38 \times 10^{-2}$, $6.54 \times 10^{-4}$ and $9.84 \times 10^{-5}$ after rounds 1, 2 and 3).

Table 6: **Alignment, proximal training and ordinary training.** Medians over 30 matched dataset–seed combinations without correction and 18 with moment correction. The distance $\delta_{\mathrm{last}}$ and trajectory error are defined in Appendix E.4. Without correction, endpoint error compares $G_8$ with $G_\star$; corrected errors compare $G_t$ with $G_{\mathrm{cen}}$. Bold marks the smallest value in each column.

| | Without correction ($n = 30$) | | | Corrected ($n = 18$) | |
|---|---|---|---|---|---|
| Training | $\delta_{\mathrm{last}}$ | Trajectory $e_F$ | Endpoint $e_F$ | Round 1 $e_F$ | Round 8 $e_F$ |
| Ordinary | $1.87 \times 10^{-1}$ | $6.75 \times 10^{-2}$ | $1.86 \times 10^{-2}$ | $5.20 \times 10^{-1}$ | $7.07 \times 10^{-3}$ |
| Proximal | $6.35 \times 10^{-3}$ | $3.63 \times 10^{-4}$ | $3.24 \times 10^{-4}$ | $2.38 \times 10^{-2}$ | $7.06 \times 10^{-5}$ |
| Aligned | $\mathbf{1.54 \times 10^{-4}}$ | $\mathbf{1.12 \times 10^{-4}}$ | $\mathbf{1.10 \times 10^{-4}}$ | $\mathbf{8.38 \times 10^{-5}}$ | $\mathbf{6.27 \times 10^{-5}}$ |

**Cost of mismatched coordinates.** To isolate the effect of solution multiplicity on one server update, we aggregate the same first-round local models twice, once as trained and once after alignment, and evaluate the weighted training objective (2) with the corresponding private features and the same head bias. The same six settings, with five seeds each and common or independently initialized backbones, give 60 runs; all clients start from the same head and train for $E = 4000$ epochs without correction or a proximal regularizer. Alignment lowers the objective in all 60 runs and removes a median of $4.6\%$ to $23.7\%$ of it across settings, similarly for the two initializations (Table 7). Averaging heads in mismatched coordinates thus raises the training objective within a single round.

Table 7: **Cost of mismatched coordinates in one aggregation.** The same first-round local models are averaged twice, with the heads as trained and after alignment to the broadcast head, giving training objectives (2) $L_{\mathrm{raw}}$ and $L_{\mathrm{al}}$. Each entry is the median over five seeds of $(L_{\mathrm{raw}} - L_{\mathrm{al}})/L_{\mathrm{raw}}$, the fraction of the objective that alignment removes. Columns give the initialization of the private backbones; all clients start from the same head. Tabular settings use ResMLP and image settings use ResNet-18.

| | Fraction of the objective removed by alignment | |
|---|---|---|
| Setting | Common backbone init. | Independent backbone init. |
| *Tabular datasets* | | |
| Beijing | 15.7% | 15.4% |
| Swimmer | 10.1% | 10.5% |
| Hopper | 17.7% | 18.1% |
| HalfCheetah | 23.6% | 23.7% |
| *Image datasets* | | |
| 3D Shapes | 4.6% | 6.0% |
| CIFAR-geo | 7.0% | 8.3% |

### F.6 Architecture, data size and full-model averaging

To test whether the agreement with $G_\star$ depends on the backbone, the amount of training data or the averaging scheme, we repeat the experiments with other backbones, larger training sets and full-model averaging. Table 8 first compares network architectures and data sizes, then full-model averaging.

**Architecture and data size.** The residual CNN and ResNet-34 comparisons use ordinary training on all five image datasets with seeds $\{0, 1, 2\}$. The larger-data setting, also with ordinary training and ResNet-18, increases 3D Shapes from 12,000 to 48,000 training samples, keeping its eight clients, and CIFAR-geo from 12,000 to 40,000 training samples partitioned into eight clients by target projection. The corresponding main CIFAR-geo experiment uses four designed client groups. Both larger-data experiments use $E = 1000$ local epochs per round, compared with $E = 4000$ in the main experiments. In all three comparisons, the median direction error to $G_\star$ lies between $9.55 \times 10^{-3}$ and $1.45 \times 10^{-2}$, and every run is closer in direction to $G_\star$ than to $G_{\mathrm{cen}}$ (Table 8).

The comparison between ordinary and corrected training in Appendix F.1 also covers the five residual-CNN settings, with three seeds each. All 65 ordinary-training runs, 50 from the main experiments and 15 with the residual CNN, are closer in direction to $G_\star$ than to $G_{\mathrm{cen}}$. For the residual CNN, moment correction reduces the median direction error to $G_{\mathrm{cen}}$ from $5.24 \times 10^{-1}$ to $1.93 \times 10^{-3}$.

Table 8: **Architecture, data size and full-model averaging.** Median direction errors over all runs in each comparison. Closer to $G_\star$ is the percentage of runs closer in direction to $G_\star$ than to $G_{\mathrm{cen}}$; corrected runs target $G_{\mathrm{cen}}$. Architecture comparisons use $E = 4000$ unless specified; other training conditions are described in the text.

| Training / architecture | To $G_\star$ | To $G_{\mathrm{cen}}$ | Closer to $G_\star$ |
|---|---|---|---|
| *Architecture and data* | | | |
| Residual CNN | $1.45 \times 10^{-2}$ | $5.24 \times 10^{-1}$ | 100% |
| Residual CNN + correction | $5.05 \times 10^{-1}$ | $1.93 \times 10^{-3}$ | 0% |
| ResNet-34 | $1.29 \times 10^{-2}$ | $5.19 \times 10^{-1}$ | 100% |
| Larger data, $E = 1000$ | $9.55 \times 10^{-3}$ | $5.63 \times 10^{-1}$ | 100% |
| *Full-model averaging* | | | |
| $E = 400$ | $4.41 \times 10^{-1}$ | $1.82 \times 10^{-1}$ | 27% |
| $E = 4000$ | $2.14 \times 10^{-2}$ | $6.20 \times 10^{-1}$ | 100% |
| $E = 4000$ + correction | $6.22 \times 10^{-1}$ | $8.08 \times 10^{-3}$ | 0% |

**Full-model averaging.** Under full-model averaging, the server averages the backbone as well as the head, so the clients also share their feature extractors. These experiments use the five tabular datasets with ResMLP and three seeds each. With $E = 400$, the endpoint is far from both $G_\star$ and $G_{\mathrm{cen}}$, with median direction errors of $4.41 \times 10^{-1}$ and $1.82 \times 10^{-1}$. With $E = 4000$, the median direction error to $G_\star$ falls to $2.14 \times 10^{-2}$, and every run is closer in direction to $G_\star$ than to $G_{\mathrm{cen}}$. Adding moment correction at $E = 4000$ reduces the median direction error to $G_{\mathrm{cen}}$ from $6.20 \times 10^{-1}$ to $8.08 \times 10^{-3}$ (Figure 15). With sufficient local training, the endpoint under full-model averaging therefore stays close to $G_\star$, and moment correction still moves it to $G_{\mathrm{cen}}$.

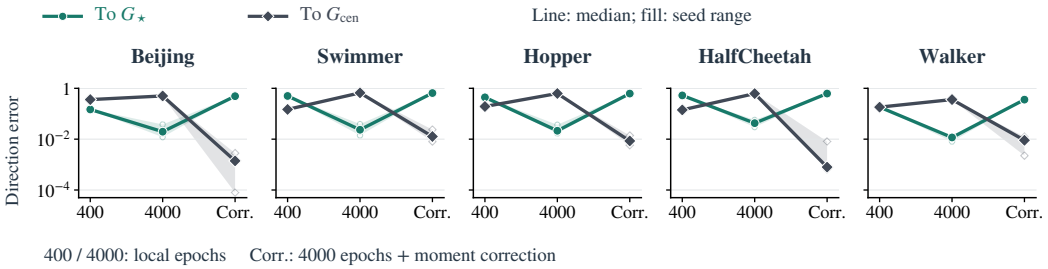


Figure 15: **Local training and moment correction under full-model averaging.** Each panel shows one tabular dataset with ResMLP and compares the same three seeds across ordinary training with $E = 400$, ordinary training with $E = 4000$, and corrected training with $E = 4000$. Colors distinguish direction error to $G_\star$ and $G_{\mathrm{cen}}$. Curves show medians; pale bands span the three seeds. The first transition changes the local budget; the second adds correction at a fixed budget.

### F.7 Test error

We next examine whether the proximal regularizer and the moment correction change the prediction error on test samples. This question lies outside the UFM analysis. The UFM treats the features of the training samples as free variables and describes how the training objective is optimized; it does not model the features that a backbone produces for unseen inputs, and hence does not predict test error. The preceding experiments show that this analysis predicts the endpoint of training across datasets and architectures. We add a comparison of test error for two reasons: (1) test error ultimately matters for any training procedure, and (2) the two interventions may affect test error differently:

- The moment correction moves the head Gram to the optimum of centralized training, using target moments that the clients share with the server, and may therefore lower test error.
- The proximal regularizer adds an extra constraint to each local update, and it is unclear whether this constraint impairs generalization.

The comparison uses the main-experiment runs on the five image datasets with ResNet-18 and seeds $\{0, 1, 2\}$, 15 dataset–seed combinations in total. Ordinary training, proximal training and proximal training with moment correction share the recipe $\lambda_H = \lambda_W = 0.01$, zero weight decay, eight rounds and $E = 4000$. All three keep the feature and head penalties; proximal training adds $\rho = 10^{-3}$, and the corrected procedure also adds $C_m$. Prediction error is evaluated as in Appendix E.4, with test samples assigned to clients as in Appendix E.1, and MSE changes are paired within each dataset–seed combination.

The proximal regularizer tightens agreement with $G_\star$, and adding correction moves the endpoint to $G_{\mathrm{cen}}$ (Table 9). Both interventions lower training MSE in all 15 comparisons, with median changes of $-0.2\%$ and $-2.0\%$. Proximal training changes test MSE by less than $1.5\%$ in every comparison and lowers it in 9 of 15. Adding correction lowers test MSE in 13 of 15 comparisons, with a median change of $-1.5\%$; the mean change ranges from $-2.3\%$ on dSprites to $+0.3\%$ on MPIIGaze (Figure 16). On these datasets, the proximal regularizer thus constrains training while keeping test error within $1.5\%$ of ordinary training, and the moment correction lowers test error slightly on average.

Table 9: **Geometry and prediction under one training recipe.** Medians over the 15 image dataset–seed combinations. Geometry columns are relative matrix errors. The loss is the MSE, and loss changes are paired against ordinary training before taking medians. **Negative changes indicate lower loss, which is better.**

| Training | $e_F(G_8, G_\star)$ | $e_F(G_8, G_{\mathrm{cen}})$ | Train loss change | Test loss change |
|---|---|---|---|---|
| Ordinary | $1.37 \times 10^{-2}$ | $5.04 \times 10^{-1}$ | – | – |
| Proximal | $1.54 \times 10^{-4}$ | $4.96 \times 10^{-1}$ | $-0.2\%$ | $-0.1\%$ |
| Proximal + moment correction | $6.23 \times 10^{-1}$ | $7.03 \times 10^{-5}$ | $-2.0\%$ | $-1.5\%$ |

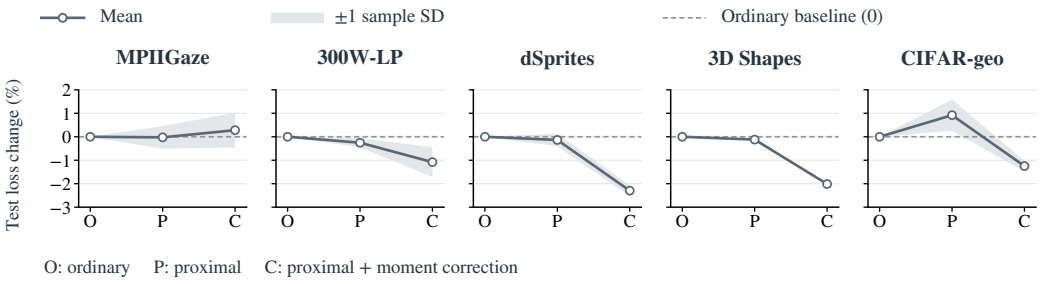


Figure 16: **Paired test-MSE changes on the five image datasets.** Values are $100(\mathrm{MSE}/\mathrm{MSE}_{\mathrm{O}} - 1)$, with the same seed's ordinary-training MSE as denominator for all procedures; O, P and C denote ordinary training, proximal training, and proximal training with moment correction. Curves show means over seeds $\{0, 1, 2\}$, and bands span one sample standard deviation; the dashed line is zero. **Negative values indicate lower loss, which is better.**

# G Supplementary analysis

## G.1 Finite-step dynamics of the profiled UFM

We study a federated update in which each client performs gradient descent on its profiled objective $F_m(W) = F(W; \Sigma_m)$. Each gradient step corresponds to first minimizing the local loss over the features and bias at the current head, then taking a full-batch gradient step on the head. We call this a *profiled step*.

At each round, all clients start from the broadcast $W_t$, take $E$ profiled steps with learning rate $\eta$, and return their heads for weighted averaging. Since each profiled step uses the full local batch, $E$ also counts the local epochs per round, as in the DNN experiments. We write $h = \eta E$ and denote the server Gram matrix by $G_t = W_t W_t^\top$.

**One local step per round.** For $E = 1$, the server update is

$$W_{t+1} = W_t - \eta \sum_m p_m \nabla F(W_t; \Sigma_m) = W_t - \eta \nabla F(W_t; \Sigma_{\text{within}}),$$

where the second equality follows from the affine dependence of $F(W; T)$ on $T$. Thus the server performs gradient descent on the averaged profiled objective

$$\bar{F}(W) := \sum_m p_m F_m(W) = F(W; \Sigma_{\text{within}}).$$

In the fully active regime, its optimal Gram matrix is

$$G_{\text{within}} = \phi(\Sigma_{\text{within}}).$$

Every head with this Gram matrix is a fixed point of the one-step update. We use $G_{\text{within}}$ as the reference for studying multiple local steps.

**Comparison with joint updates.** For comparison, suppose each client takes one simultaneous full-batch gradient step on $\mathcal{L}_m$ in $(W, b, H_m)$, starting from the common head $(W, b)$. The server averages the updated heads, while the features remain local.

The averaged updates of $W$ and $b$ equal those of gradient descent on the pooled objective (2). Since $\nabla_{H_m} \mathcal{L}_{\text{cen}} = p_m \nabla_{H_m} \mathcal{L}_m$, the feature update is

$$H_m^+ = H_m - \eta \nabla_{H_m} \mathcal{L}_m = H_m - \frac{\eta}{p_m} \nabla_{H_m} \mathcal{L}_{\text{cen}}.$$

Thus the entire round corresponds to gradient descent on $\mathcal{L}_{\text{cen}}$, with learning rate $\eta$ for $(W, b)$ and $\eta/p_m$ for $H_m$. If these joint updates converge to a pooled global minimizer, their head Gram converges to $G_{\text{cen}}$. The profiled update instead follows $\bar{F}$, whose optimal Gram is $G_{\text{within}}$. The objectives underlying the two procedures therefore have generally different optimal Gram matrices, even with one local step per round.

**Displacement under multiple local steps.** We now study how the Gram fixed point changes when $E > 1$. To state the result, write the gradient of the profiled objective as

$$\nabla F(W; T) = 2\Lambda_T(G) W, \qquad G = WW^\top,$$

where

$$\Lambda_T(G) = \frac{\lambda_W}{2} I_C - \frac{\lambda_H}{2} (G + \lambda_H I_C)^{-1} T (G + \lambda_H I_C)^{-1}.$$

We use the abbreviations

$$\Lambda_m = \Lambda_{\Sigma_m}, \qquad \bar{\Lambda} = \Lambda_{\Sigma_{\text{within}}}, \qquad R = \Sigma_{\text{within}}^{1/2}, \qquad \alpha = \sqrt{\lambda_H / \lambda_W},$$

so that $G_{\text{within}} = \alpha R - \lambda_H I_C$.

Let $f(G) = \bar{F}(W)$ whenever $WW^\top = G$, and define

$$\begin{aligned} \psi(G) &= 2 \sum_m p_m \operatorname{tr}\big(\Lambda_m(G) G \Lambda_m(G)\big), \\ \mathcal{H}[X] &= \frac{\lambda_W}{2\alpha} (R^{-1} X + X R^{-1}). \end{aligned} \tag{19}$$

These quantities satisfy

$$\psi(WW^\top) = \frac{1}{2} \sum_m p_m \left\| \nabla F_m(W) \right\|_F^2, \qquad \mathcal{H} = \operatorname{Hess} f(G_{\text{within}}).$$

Thus $\psi$ measures the weighted squared norms of the local head gradients, while $\mathcal{H}$ is the Hessian of the averaged objective with respect to the Gram matrix. The operator $\mathcal{H}$ is positive definite on $\operatorname{Sym}(C)$, the space of symmetric $C \times C$ matrices.

For sufficiently small $h$, the following theorem establishes a locally attracting Gram fixed point near $G_{\text{within}}$ and gives its first-order displacement from $G_{\text{within}}$.

**Theorem 2** (Fixed point and local convergence of profiled updates)**.** *Assume the fully active conditions of Section 3. For an integer $E \geq 1$ and learning rate $\eta > 0$, define the server update*

$$\Phi_{\eta,E}(W) = \sum_m p_m \mathrm{GD}_\eta^E[F_m](W),$$

*where $\mathrm{GD}_\eta^E[F_m](W)$ denotes $E$ gradient steps on $F_m$, starting from $W$.*

*There exist constants $r, \varepsilon_0, K > 0$ such that, for every $P \geq C$ and every $(\eta, E)$ satisfying $h = \eta E \leq \varepsilon_0$, the following hold.*

1. Gram dynamics and a unique nearby fixed point. *The next Gram matrix depends only on the current Gram matrix, so the update defines a map*

   $$G_{t+1} = \Phi_{\eta,E}^G(G_t).$$

   *This map has exactly one fixed point in $\|G - G_{\mathrm{within}}\|_F \leq r$. We denote this positive-definite fixed point by $G_{\eta,E}$.*

2. Displacement from the one-step fixed point. *Define*

   $$\Delta = \frac{1}{2}\mathcal{H}^{-1}[\nabla\psi(G_{\mathrm{within}})]\,,$$

   *where $\psi$ and $\mathcal{H}$ are given in* (19)*. Then*

   $$\|G_{\eta,E} - G_{\mathrm{within}} - \eta(E-1)\Delta\|_F \leq K(\eta E)^2.$$

   *For $E = 1$, the fixed point is exactly $G_{\eta,1} = G_{\mathrm{within}}$.*

3. Convergence from nearby initializations. *If $G_0$ is sufficiently close to $G_{\eta,E}$, then for all $t \geq 0$,*

   $$\|G_t - G_{\eta,E}\|_F \leq \left(1 - \frac{1}{2}h\theta_{\min}\right)^t \|G_0 - G_{\eta,E}\|_F\,,$$

   *where*

   $$\theta_{\min} = 2\lambda_W\left(1 - \sqrt{\frac{\lambda_H\lambda_W}{\lambda_{\min}(\Sigma_{\mathrm{within}})}}\right) > 0.$$

*The constants $r, \varepsilon_0, K$ depend only on the client covariances, weights, regularization parameters and $C$; they are independent of $\eta$, $E$ and $P$. The convergence statement concerns the Gram matrices $G_t$. No convergence of the heads $W_t$ is asserted.*

### G.2 Proof of Theorem 2

*Proof.* We write $\Phi = \Phi_{\eta,E}$ for the head update and $\Phi^G = \Phi_{\eta,E}^G$ for its induced Gram map. All remainder estimates below hold uniformly in a fixed neighborhood of $G_{\mathrm{within}}$, for sufficiently small $h = \eta E$. Their constants may depend on the client covariances, weights, regularization parameters and $C$, but not on $\eta$, $E$ or $P$. For linear operators on $\mathrm{Sym}(C)$, we use the norm induced by the Frobenius norm and denote the identity operator by $\mathrm{Id}$.

*1. Exact Gram updates and the local expansion.* Fix a broadcast head $W$ with Gram $G = WW^\top$. For client $m$, let $W_{m,k}$ be the head after $k$ local steps and let $G_{m,k} = W_{m,k}W_{m,k}^\top$, with $W_{m,0} = W$ and $G_{m,0} = G$. Define

$$A_{m,k} = I_C - 2\eta\Lambda_m(G_{m,k}).$$

The head update is $W_{m,k+1} = A_{m,k}W_{m,k}$. Since $A_{m,k}$ is symmetric, its Gram satisfies

$$G_{m,k+1} = A_{m,k}G_{m,k}A_{m,k}. \tag{20}$$

Thus each local Gram trajectory depends only on the initial Gram $G$. The returned head is $W_{m,E} = M_m(G)W$, where

$$M_m(G) = A_{m,E-1}\cdots A_{m,0}.$$

Writing $M(G) = \sum_m p_m M_m(G)$, the server update and its Gram are

$$\Phi(W) = M(G)W, \qquad \Phi^G(G) = M(G)GM(G)^\top.$$

This establishes that the server Gram evolves independently of the choice of head factor $W$.

Choose two closed Frobenius balls centered at $G_{\text{within}}$, with radii $0 < r_1 < r_2$, both contained in the positive-definite cone. Let the initial Gram satisfy $\|G - G_{\text{within}}\|_F \le r_1$. On the larger ball, smoothness and compactness give uniform bounds on the required derivatives and on each local increment: $\|G_{m,k+1} - G_{m,k}\|_F \le c\eta$ for sufficiently small $\eta$. If $h = \eta E$ is small enough that $ch < r_2 - r_1$, induction gives

$$\|G_{m,k} - G\|_F \le c\eta k, \qquad \|G_{m,k} - G_{\text{within}}\|_F \le r_1 + c\eta k < r_2, \qquad 0 \le k \le E.$$

Thus all local iterates remain in the larger ball, where these uniform bounds apply.

The estimates can be made independent of the feature width. Indeed, write $W = G^{1/2}V$, where $VV^\top = I_C$. Local gradient descent from $W$ is the corresponding iteration from $G^{1/2}$ followed by right multiplication by $V$. Since $\|ZV\|_F = \|Z\|_F$ for every $C \times C$ matrix $Z$, the expansion and its remainder can be established using square heads and then transferred to every $P \ge C$.

For a fixed client, abbreviate $g_m = \nabla F_m$ and $W_k = W_{m,k}$. Taylor expansion along the local iterates gives

$$\begin{aligned} W_k - W &= -\eta k g_m(W) + O((\eta k)^2), \\ g_m(W_k) &= g_m(W) - \eta k Dg_m(W)[g_m(W)] + O((\eta k)^2). \end{aligned}$$

Substituting the second expansion into $W_E = W - \eta \sum_{k=0}^{E-1} g_m(W_k)$ yields

$$\mathrm{GD}_\eta^E[F_m](W) = W - h\nabla F_m(W) + \frac{\eta^2 E(E-1)}{2}\nabla^2 F_m(W)[\nabla F_m(W)] + O(h^3). \tag{21}$$

Here $\sum_{k=0}^{E-1} k = E(E-1)/2$, while the summed remainder is bounded by a constant times $\eta^3 \sum_{k=0}^{E-1} k^2 \le h^3/3$. This gives a remainder bound uniform in $E$.

*2. The averaged update and its derivative.* Define

$$\Psi(W) = \frac{1}{2}\sum_m p_m \|\nabla F_m(W)\|_F^2 = \psi(WW^\top).$$

Its gradient is

$$\nabla\Psi(W) = \sum_m p_m \nabla^2 F_m(W)[\nabla F_m(W)].$$

Averaging (21) therefore gives

$$\Phi(W) = W - h\nabla\bar{F}(W) + \frac{\eta^2 E(E-1)}{2}\nabla\Psi(W) + O(h^3).$$

Thus, up to an $O(h^3)$ remainder, one communication round agrees with a gradient step of size $h$ on $\bar{F} - \eta(E-1)\Psi/2$. The identity $\nabla\bar{F} = \sum_m p_m \nabla F_m$ also gives

$$\Psi(W) = \frac{1}{2}\left\|\nabla\bar{F}(W)\right\|_F^2 + \frac{1}{2}\sum_m p_m \left\|\nabla F_m(W) - \nabla\bar{F}(W)\right\|_F^2.$$

The second term measures the dispersion of the client gradients at the common head $W$.

For a symmetric matrix $A$, define the linear operator $\mathcal{S}_A[X] = AX + XA$ on symmetric matrices. Using $\nabla\bar{F}(W) = 2\bar{\Lambda}(G)W$ and $\nabla\Psi(W) = 2\nabla\psi(G)W$, we obtain

$$\Phi^G(G) = G - 2h\mathcal{S}_G[\bar{\Lambda}(G)] + \eta^2 E(E-1)\mathcal{S}_G[\nabla\psi(G)] + 4h^2\bar{\Lambda}(G)G\bar{\Lambda}(G) + O(h^3). \tag{22}$$

We next estimate the derivative of the exact Gram map. Differentiating the recurrence (20) with respect to the initial Gram gives $DG_{m,k}(G) = \mathrm{Id} + O(\eta k)$. The product rule for $M_m(G)$ then gives

$$DM_m(G)[X] = -2\eta \sum_{k=0}^{E-1} P_{m,>k}\, D\Lambda_m(G_{m,k})[DG_{m,k}(G)[X]]\, P_{m,<k},$$

where

$$P_{m,>k} = A_{m,E-1} \cdots A_{m,k+1}, \qquad P_{m,<k} = A_{m,k-1} \cdots A_{m,0}.$$

Empty products equal $I_C$. Both products are $I_C + O(h)$. Using $G_{m,k} = G + O(\eta k)$ in these expressions and averaging over clients gives

$$M(G) = I_C - 2h\bar{\Lambda}(G) + O(h^2),$$
$$DM(G)[X] = -2hD\bar{\Lambda}(G)[X] + O(h^2)\,\|X\|_F\,.$$

Applying the product rule to $M(G)GM(G)^\top$ now yields

$$D\Phi^G(G) = \mathrm{Id} - 2h\big(\mathcal{S}_G \circ D\bar{\Lambda}(G) + \mathcal{S}_{\bar{\Lambda}(G)}\big) + O(h^2). \tag{23}$$

The remainder is uniform in the induced operator norm on a fixed smaller neighborhood of $G_{\mathrm{within}}$.

*3. Existence and uniqueness of a nearby fixed point.* For fixed $(\eta, E)$, define

$$Q(G) = \frac{G - \Phi^G(G)}{2h}, \qquad \mathcal{T} = \mathcal{S}_{G_{\mathrm{within}}} \circ \mathcal{H}.$$

A fixed point of $\Phi^G$ is exactly a zero of $Q$. At the reference Gram, $\bar{\Lambda}(G_{\mathrm{within}}) = 0$ and $D\bar{\Lambda}(G_{\mathrm{within}}) = \mathcal{H}$. Hence (23) implies

$$DQ(G) = \mathcal{T} + O(\|G - G_{\mathrm{within}}\|_F + h).$$

The operators $\mathcal{H}$ and $\mathcal{S}_{G_{\mathrm{within}}}$ are diagonal in the symmetric matrix basis associated with an orthonormal eigenbasis of $R$. Their eigenvalues are positive, so their product $\mathcal{T}$ is self-adjoint, positive definite and invertible.

Consider the auxiliary map

$$\mathcal{K}(G) = G - \mathcal{T}^{-1}Q(G).$$

The estimate for $DQ$ gives

$$D\mathcal{K}(G) = \mathrm{Id} - \mathcal{T}^{-1}DQ(G) = O(\|G - G_{\mathrm{within}}\|_F + h)\,.$$

Also, (22) gives $Q(G_{\mathrm{within}}) = O(h)$, and hence

$$\mathcal{K}(G_{\mathrm{within}}) - G_{\mathrm{within}} = -\mathcal{T}^{-1}Q(G_{\mathrm{within}}) = O(h).$$

Thus there are constants $c_1, c_2 > 0$ such that, uniformly in the neighborhood above for sufficiently small $h$,

$$\|D\mathcal{K}(G)\|_{\mathrm{op}} \le c_1\big(\|G - G_{\mathrm{within}}\|_F + h\big), \quad \|\mathcal{K}(G_{\mathrm{within}}) - G_{\mathrm{within}}\|_F \le c_2 h.$$

Choose $r > 0$ small enough that the closed ball $\|G - G_{\mathrm{within}}\|_F \le r$ lies in this neighborhood and $c_1 r \le 1/4$. Then choose $\varepsilon_0 > 0$ small enough that the estimates above hold for $h \le \varepsilon_0$ and

$$c_1\varepsilon_0 \le \frac{1}{4}, \qquad c_2\varepsilon_0 \le \frac{r}{2}.$$

For every such $h$, these choices ensure

$$\sup_{\|G - G_{\mathrm{within}}\|_F \le r} \|D\mathcal{K}(G)\|_{\mathrm{op}} \le \frac{1}{2}, \qquad \|\mathcal{K}(G_{\mathrm{within}}) - G_{\mathrm{within}}\|_F \le \frac{r}{2}.$$

Since the ball is convex, the derivative bound implies

$$\|\mathcal{K}(G) - \mathcal{K}(G')\|_F \le \frac{1}{2}\,\|G - G'\|_F$$

for any two matrices $G, G'$ in the ball. Moreover, for every $G$ in the ball,

$$\|\mathcal{K}(G) - G_{\mathrm{within}}\|_F \le \|\mathcal{K}(G) - \mathcal{K}(G_{\mathrm{within}})\|_F + \|\mathcal{K}(G_{\mathrm{within}}) - G_{\mathrm{within}}\|_F$$
$$\le \frac{1}{2}\,\|G - G_{\mathrm{within}}\|_F + \frac{r}{2} \le r.$$

Thus $\mathcal{K}$ is a contraction from the closed ball into itself. Banach's fixed-point theorem gives a unique fixed point in this ball, denoted $G_{\eta,E}$. Since

$$\mathcal{K}(G) = G \quad \Longleftrightarrow \quad Q(G) = 0 \quad \Longleftrightarrow \quad \Phi^G(G) = G,$$

this is also the unique fixed point of the Gram update in the ball. The ball is contained in the positive-definite cone, so $G_{\eta,E} \succ 0$.

Finally, the fixed-point identity and the contraction bound give

$$\|G_{\eta,E} - G_{\text{within}}\|_F \le \frac{1}{2}\|G_{\eta,E} - G_{\text{within}}\|_F + \|\mathcal{K}(G_{\text{within}}) - G_{\text{within}}\|_F .$$

Rearranging yields

$$\|G_{\eta,E} - G_{\text{within}}\|_F \le 2c_2 h = O(h).$$

*4. Displacement of the fixed point.* The preceding estimate implies $\bar{\Lambda}(G_{\eta,E}) = O(h)$. Evaluate (22) at $G_{\eta,E}$ and divide the fixed-point equation by $2h$. This gives

$$\mathcal{S}_{G_{\eta,E}}[\bar{\Lambda}(G_{\eta,E})] = \frac{\eta(E-1)}{2}\mathcal{S}_{G_{\eta,E}}[\nabla\psi(G_{\eta,E})] + 2h\bar{\Lambda}(G_{\eta,E})G_{\eta,E}\bar{\Lambda}(G_{\eta,E}) + O(h^2).$$

For $G \succ 0$, the operator $\mathcal{S}_G$ multiplies the $(i,j)$ entry by the sum of two eigenvalues of $G$, when expressed in an eigenbasis of $G$. These sums are bounded away from zero in our neighborhood, so $\mathcal{S}_G^{-1}$ is uniformly bounded there. The quadratic term $2h\bar{\Lambda}(G_{\eta,E})G_{\eta,E}\bar{\Lambda}(G_{\eta,E})$ is $O(h^3)$ because $\bar{\Lambda}(G_{\eta,E}) = O(h)$. Applying $\mathcal{S}_{G_{\eta,E}}^{-1}$ therefore gives

$$\bar{\Lambda}(G_{\eta,E}) = \frac{\eta(E-1)}{2}\nabla\psi(G_{\eta,E}) + O(h^2).$$

We may replace $\nabla\psi(G_{\eta,E})$ by $\nabla\psi(G_{\text{within}})$ with an additional $O(h^2)$ error: the two arguments differ by $O(h)$ and $\eta(E-1) \le h$. Finally, Taylor expansion at $G_{\text{within}}$ yields

$$\mathcal{H}[G_{\eta,E} - G_{\text{within}}] = \frac{\eta(E-1)}{2}\nabla\psi(G_{\text{within}}) + O(h^2).$$

Applying $\mathcal{H}^{-1}$ proves

$$G_{\eta,E} - G_{\text{within}} = \eta(E-1)\Delta + O(h^2), \qquad \Delta = \frac{1}{2}\mathcal{H}^{-1}[\nabla\psi(G_{\text{within}})].$$

The uniform remainder gives the asserted bound with a constant $K$. For $E = 1$, the exact one-step update fixes $G_{\text{within}}$. Uniqueness in the ball therefore gives $G_{\eta,1} = G_{\text{within}}$ exactly.

*5. Local convergence of the Gram iteration.* We now establish contraction of the actual update $\Phi^G$, rather than the auxiliary map $\mathcal{K}$. Let $r_i$ be the eigenvalues of $R$ and $\beta_i = \alpha r_i - \lambda_H > 0$ those of $G_{\text{within}}$. In the corresponding symmetric matrix basis, the eigenvalues of $\mathcal{T}$ are

$$\theta_{ij} = \frac{\lambda_W}{2\alpha}(\beta_i + \beta_j)(r_i^{-1} + r_j^{-1}).$$

Put $s = \sqrt{\lambda_H \lambda_W}$ and $r_{\min} = \min_i r_i$. Then

$$\begin{aligned}\theta_{ij} &= \frac{\lambda_W}{2}\left[\frac{(r_i + r_j)^2}{r_i r_j} - 2s(r_i^{-1} + r_j^{-1})\right] \\ &\ge 2\lambda_W(1 - s/r_{\min}) = \theta_{\min} > 0.\end{aligned}$$

The inequality uses $(r_i + r_j)^2/(r_i r_j) \ge 4$ and $r_i^{-1} + r_j^{-1} \le 2/r_{\min}$. Equality holds when $r_i = r_j = r_{\min}$, so this is the smallest eigenvalue of $\mathcal{T}$.

Because $\mathcal{T}$ is self-adjoint, for $h \le 1/(2\max_{i,j}\theta_{ij})$,

$$\|\mathrm{Id} - 2h\mathcal{T}\|_{\text{op}} = 1 - 2h\theta_{\min}.$$

By (23) and smoothness,

$$D\Phi^G(G) = \mathrm{Id} - 2h\mathcal{T} + O\big(h\,\|G - G_{\text{within}}\|_F + h^2\big).$$

Since $G_{\eta,E} - G_{\text{within}} = O(h)$, there is a constant $C_0 > 0$ such that

$$\|G_{\eta,E} - G_{\text{within}}\|_F \le C_0 h.$$

For any $G$ satisfying $\|G - G_{\eta,E}\|_F \le r_{\text{att}}$, we therefore have

$$\|G - G_{\text{within}}\|_F \le r_{\text{att}} + C_0 h.$$

Thus the ball centered at $G_{\eta,E}$ lies inside the uniqueness ball whenever $r_{\text{att}} + C_0 h \le r$. On this ball, the derivative estimate above gives

$$\left\|D\Phi^G(G)\right\|_{\text{op}} \le 1 - 2h\theta_{\min} + C_1 h r_{\text{att}} + C_2 h^2$$

for constants $C_1, C_2 > 0$ independent of $\eta$, $E$ and $P$. Choose a fixed radius $r_{\text{att}} > 0$ small enough that

$$r_{\text{att}} \le \frac{r}{2}, \qquad C_1 r_{\text{att}} \le \frac{\theta_{\min}}{2}.$$

Then decrease $\varepsilon_0$, if necessary, so that all previous small-$h$ conditions remain valid and

$$C_0 \varepsilon_0 \le \frac{r}{2}, \qquad C_2 \varepsilon_0 \le \theta_{\min}.$$

For every $0 < h \le \varepsilon_0$, these choices ensure that the ball lies inside the uniqueness ball and that

$$\sup_{\|G - G_{\eta,E}\|_F \le r_{\text{att}}} \left\|D\Phi^G(G)\right\|_{\text{op}} \le 1 - \frac{1}{2} h\theta_{\min} =: q, \qquad 0 < q < 1.$$

Since the ball is convex and $\Phi^G(G_{\eta,E}) = G_{\eta,E}$, the mean-value bound gives

$$\left\|\Phi^G(G) - G_{\eta,E}\right\|_F = \left\|\Phi^G(G) - \Phi^G(G_{\eta,E})\right\|_F \le q \left\|G - G_{\eta,E}\right\|_F \le q r_{\text{att}} < r_{\text{att}}.$$

Thus each update decreases the distance to the fixed point and keeps the Gram matrix inside the same ball. The same estimate therefore applies at every subsequent round. For any initial Gram satisfying $\|G_0 - G_{\eta,E}\|_F \le r_{\text{att}}$, iteration yields

$$\|G_t - G_{\eta,E}\|_F \le q^t \|G_0 - G_{\eta,E}\|_F = \left(1 - \frac{1}{2} h\theta_{\min}\right)^t \|G_0 - G_{\eta,E}\|_F,$$

which proves the claimed local convergence. Although $G_{\eta,E}$ depends on $\eta$ and $E$, the radius $r_{\text{att}}$ and the upper bound $\varepsilon_0$ can be chosen independently of $\eta$, $E$ and $P$. □

### G.3 Dependence of the displacement on client covariances

For small $h$, Theorem 2 gives $G_{\eta,E} - G_{\text{within}} = \eta(E-1)\Delta + O(h^2)$. To understand how differences among client covariances affect this displacement, we here evaluate the coefficient $\Delta$ in terms of these covariances. To obtain an explicit prediction, we treat the case of commuting client covariances.

Assume that the client covariances commute pairwise. In a common orthonormal eigenbasis, write

$$\Sigma_m = \text{diag}(\lambda_{m,1}, \dots, \lambda_{m,C}), \qquad \bar{\lambda}_i = \sum_m p_m \lambda_{m,i}.$$

Define

$$r_i = \sqrt{\bar{\lambda}_i}, \qquad \alpha = \sqrt{\lambda_H / \lambda_W}, \qquad s = \sqrt{\lambda_H \lambda_W}.$$

The reference Gram is diagonal in this basis, with entries

$$\bar{g}_i := (G_{\text{within}})_{ii} = \alpha r_i - \lambda_H.$$

To measure the covariance differences in each coordinate, put

$$e_{m,i} = \frac{\lambda_{m,i}}{\bar{\lambda}_i} - 1, \qquad \bar{e}_i = \sum_m p_m e_{m,i}^2.$$

The definition of $\bar{\lambda}_i$ implies $\sum_m p_m e_{m,i} = 0$.

For a diagonal Gram $G = \mathrm{diag}(g_1, \dots, g_C)$, the $i$th diagonal entry of $\Lambda_m(G)$ is

$$\ell_{m,i}(g_i) = \frac{\lambda_W}{2} - \frac{\lambda_H \lambda_{m,i}}{2(g_i + \lambda_H)^2}.$$

Its derivative is

$$\ell'_{m,i}(g_i) = \frac{\lambda_H \lambda_{m,i}}{(g_i + \lambda_H)^3}.$$

The definition of $\psi$ therefore reduces to

$$\psi(G) = 2 \sum_i g_i \sum_m p_m \ell_{m,i}(g_i)^2.$$

At $g_i = \bar{g}_i$, using $\bar{g}_i + \lambda_H = \alpha r_i$ and $\lambda_{m,i} = r_i^2(1 + e_{m,i})$, we obtain

$$\ell_{m,i}(\bar{g}_i) = -\frac{\lambda_W}{2} e_{m,i}, \qquad \ell'_{m,i}(\bar{g}_i) = \frac{\lambda_W}{\alpha r_i}(1 + e_{m,i}).$$

Differentiating $\psi$ with respect to $g_i$ now gives

$$[\nabla \psi(G_{\mathrm{within}})]_{ii} = 2 \sum_m p_m \ell_{m,i}(\bar{g}_i)^2 + 4 \bar{g}_i \sum_m p_m \ell_{m,i}(\bar{g}_i) \ell'_{m,i}(\bar{g}_i) = \lambda_W^2 \left( -\frac{3}{2} + \frac{2s}{r_i} \right) \bar{e}_i.$$

The relations $\sum_m p_m e_{m,i} = 0$ and $\bar{g}_i/(\alpha r_i) = 1 - s/r_i$ are used.

The full gradient at $G_{\mathrm{within}}$ is diagonal as well. Indeed, for every diagonal matrix $D = \mathrm{diag}(\pm 1, \dots, \pm 1)$, we have $\psi(DGD) = \psi(G)$ and $DG_{\mathrm{within}}D = G_{\mathrm{within}}$. Differentiating this invariance gives

$$D \nabla \psi(G_{\mathrm{within}}) D = \nabla \psi(G_{\mathrm{within}}).$$

Since this holds for every choice of signs, all off-diagonal entries of the gradient vanish.

Recall that the displacement coefficient satisfies

$$2\mathcal{H}[\Delta] = \nabla \psi(G_{\mathrm{within}}).$$

In the common eigenbasis, the Hessian acts entrywise as

$$(\mathcal{H}[X])_{ij} = \frac{\lambda_W}{2\alpha}(r_i^{-1} + r_j^{-1}) X_{ij}.$$

All these coefficients are positive, so a diagonal $\nabla \psi(G_{\mathrm{within}})$ gives a diagonal $\Delta$. Writing $\Delta_i := \Delta_{ii}$, we obtain

$$\Delta_i = \frac{\alpha r_i}{2\lambda_W} [\nabla \psi(G_{\mathrm{within}})]_{ii} = \frac{\alpha r_i \lambda_W}{2} \left( -\frac{3}{2} + \frac{2s}{r_i} \right) \bar{e}_i = s \left( s - \frac{3}{4} r_i \right) \bar{e}_i,$$

where the last equality uses $\alpha \lambda_W = s$. Substituting the definitions of $s$, $r_i$ and $\bar{e}_i$ gives

$$\Delta_i = \sqrt{\lambda_H \lambda_W} \left( \sqrt{\lambda_H \lambda_W} - \frac{3}{4} \sqrt{\bar{\lambda}_i} \right) \sum_m p_m \left( \frac{\lambda_{m,i}}{\bar{\lambda}_i} - 1 \right)^2. \tag{24}$$

At fixed $\bar{\lambda}_i$, this coefficient is proportional to the weighted variance $\bar{e}_i$ of the relative client covariance deviations. If $\bar{e}_i = 0$, then $\Delta_i = 0$. For $\bar{e}_i > 0$, $\Delta_i$ is positive when $\bar{\lambda}_i < 16 \lambda_H \lambda_W / 9$, zero at equality, and negative when $\bar{\lambda}_i > 16 \lambda_H \lambda_W / 9$.

For sufficiently small $h$, Theorem 2 gives

$$(G_{\eta,E} - G_{\mathrm{within}})_{ii} = \eta(E-1)\Delta_i + O(h^2).$$

This predicted displacement $\eta(E-1)\Delta_i$ is compared with the numerically measured displacement as shown in Figure 17(d).

### G.4 NUMERICAL CHECKS OF FINITE-STEP DYNAMICS

We numerically examine the fixed points of the profiled Gram update. Figure 17(a–c) compares them with the one-step reference $G_{\mathrm{within}}$ and the BW barycenter $G_\star$ predicted by Theorem 1 under exact closest-head selection. Panels (d) and (e) test the small-$h$ predictions for the displacement and local convergence rate, where $h = \eta E$. Below, $G_{\eta,E}$ denotes the numerically obtained fixed point, including in experiments beyond the small-$h$ regime.

Panels (a–c) use three random noncommuting instances with $C = 3$ and $M = 4$, five learning rates $\eta \in \{0.02, 0.05, 0.1, 0.2, 0.5\}$, and ten local step counts $E$ ranging from 1 to 2000. For each pair $(\eta, E)$, we iterate the server update until the server Gram residual is below $10^{-13}$.

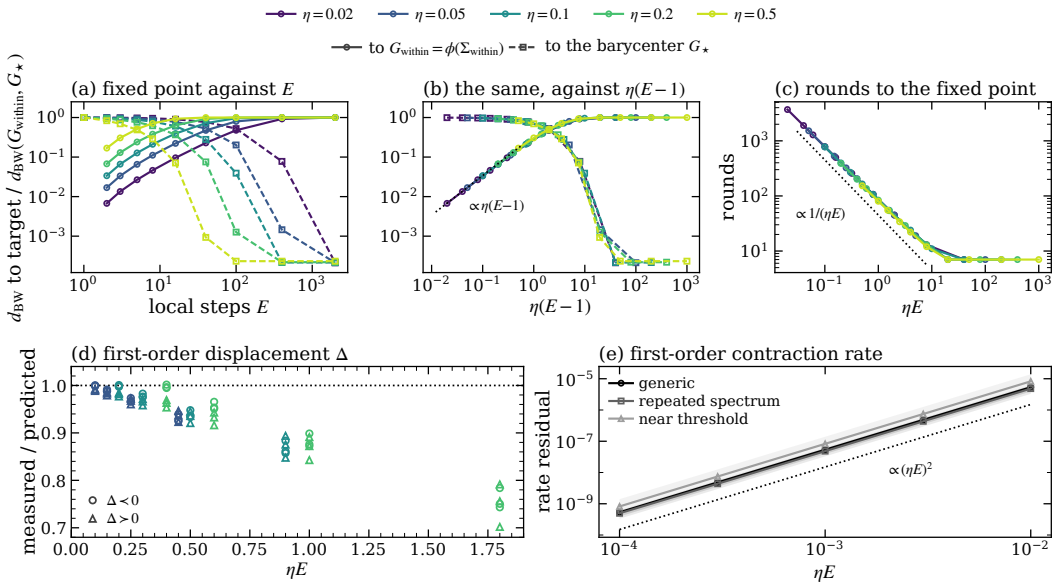


Figure 17: Numerical checks of finite-step profiled dynamics. (a) BW distances from $G_{\eta,E}$ to $G_{\mathrm{within}}$ (circles, solid) and $G_\star$ (squares, dashed), normalized by $d_{\mathrm{BW}}(G_{\mathrm{within}}, G_\star)$; medians over three instances with $C = 3$ and $M = 4$. (b) The same distances plotted against $\eta(E-1)$; the dotted guide has slope one. (c) Communication rounds needed to reach the stopping criterion. (d) Measured displacements divided by the predictions $\eta(E-1)\Delta_i$ in 24 commuting configurations, colored by $\eta$. (e) Absolute errors in the first-order prediction $1 - 2h\theta_{\min}$ for the Jacobian spectral radius; medians and interquartile ranges within each instance family over 780 stability records. The dotted guide has slope two.

**Fixed points and the BW prediction.** Panels (a) and (b) show how the fixed point changes as local training increases. All distances in these panels are normalized by $d_{\mathrm{BW}}(G_{\mathrm{within}}, G_\star)$. For $E = 1$, the numerical fixed point agrees with $G_{\mathrm{within}}$ to within $3.5 \times 10^{-6}$ in these units. For small $h$, its distance from $G_{\mathrm{within}}$ grows approximately linearly with $\eta(E-1)$, consistent with Theorem 2. Panel (b) makes this scaling visible by plotting the same data against $\eta(E-1)$.

With more local training, the fixed points in these examples move away from $G_{\mathrm{within}}$ and closer to $G_\star$. The distances to the two references cross near $\eta(E-1) = 2$, and the normalized distance to $G_\star$ eventually levels off between $1.4 \times 10^{-4}$ and $8.3 \times 10^{-4}$. Thus finite-step profiled training closely approaches the BW prediction in these examples, even though it does not explicitly impose the selection rule assumed in Theorem 1.

**Communication rounds.** Panel (c) shows the communication cost of reaching the fixed point. For small $h$, the median round count decreases approximately as $1/(\eta E)$, from 3698 rounds at $\eta E = 0.02$. It levels off at 7 rounds for $\eta E \geq 20$. The number of local head steps per client is $E$ times the round count, so fewer communication rounds need not mean fewer local updates.

**Local optimality and head selection.** At $E = 2000$ and $\eta \in \{0.1, 0.5\}$, we check whether the returned heads are both locally optimal and close to those prescribed by the selection rule. Writing $W$ for the broadcast and $U_m$ for client $m$'s return, the maximum relative Gram error satisfies

$$\max_m \frac{\left\|U_m U_m^\top - G_m\right\|_F}{\|G_m\|_F} < 4 \times 10^{-15}.$$

In contrast, the selection error

$$\max_m \frac{\|U_m - \Pi_m(W)\|_F}{\|\Pi_m(W)\|_F}$$

ranges from $6 \times 10^{-5}$ to $3 \times 10^{-4}$ across the three noncommuting instances. The heads therefore have nearly optimal local Grams but differ from the closest optimal heads.

This distinction is possible because successive gradient steps multiply the head by symmetric matrices whose ordered product need not be symmetric. This product can differ from the symmetric multiplier $T_m(WW^\top)$ defining $\Pi_m(W)$ in Lemma 1. The selection errors accompany the nonzero distance plateau above; we also observe continued head motion after the server Gram stabilizes in the noncommuting cases with $E > 1$. For comparison, in one commuting instance with $E = 2000$ and $\eta \in \{0.1, 0.5\}$, the selection error is below $7 \times 10^{-15}$. The BW distance from $G_{\eta,E}$ to $G_\star$, normalized by $d_{\mathrm{BW}}(G_{\mathrm{within}}, G_\star)$ as in panels (a) and (b), is below $1.14 \times 10^{-6}$.

**Testing the displacement formula.** Panel (d) tests the coefficient $\Delta_i$ in (24) using 24 commuting configurations. In their common eigenbasis, the first-order prediction is

$$(G_{\eta,E} - G_{\mathrm{within}})_{ii} \approx \eta(E-1)\Delta_i.$$

The panel plots the measured displacement divided by this prediction, so a ratio of one indicates agreement. Across all 72 coordinates, including positive and negative predicted displacements, every measured sign agrees with the prediction. For the eight configurations with $h \le 0.25$, the ratios lie in $[0.96, 1.00]$.

**Testing the local convergence rate.** Panel (e) tests the first-order prediction $1 - 2h\theta_{\min}$ for the Jacobian spectral radius at the fixed point, derived in Appendix G.2. It plots the absolute error

$$\left|\mathrm{spr}\big(D\Phi^G_{\eta,E}(G_{\eta,E})\big) - (1 - 2h\theta_{\min})\right|,$$

where spr denotes spectral radius. This prediction concerns the derivative at the fixed point; the bound $1 - \frac{1}{2}h\theta_{\min}$ in Theorem 2 instead controls contraction throughout a neighborhood.

We evaluate 780 stability records covering generic instances, instances with repeated eigenvalues, and fully active instances near the activation threshold, using $E \in \{2, 4, 8\}$ and

$$h \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}.$$

All measured spectral radii are below one. The prediction errors are consistent with $O(h^2)$ scaling, with a median fitted log–log slope of 1.999. Slopes are fitted separately for each instance and $E$, using $h \le 10^{-3}$ for near-threshold instances and all five values otherwise. Panel (e) displays all five values in every case.