FEDAFFINE: PERSONALIZED FEDERATED LEARNING WITH SHARED AFFINE PREDICTOR FAMILIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated learning must decide which cross-client differences should survive collaboration, yet trained client models conflate persistent population variation with finite-sample estimation error. We call this the population–estimation ambiguity and impose a structural restriction: client population predictors lie near a low-dimensional affine family. A sample-weighted sequence model decomposes risk into unrepresented population variation, client-specific estimation, and pooled estimation, separating personalization capacity from coefficient regularization. This motivates FEDAFFINE, which collaboratively learns a high-dimensional base and shared basis while keeping a short coefficient local. Coefficients can be reconstructed from local data, enabling deployment to held-out clients without persistent training-time state. Under the pFL-Bench FEMNIST protocol, reconstructive FEDAFFINE re-estimates three coefficients from zero and attains weighted accuracy of 90.18% on participating clients and 90.22% on held-out clients, versus 89.48% and 89.41% for Ditto. We additionally report the accuracy–adaptation-compute frontier. On multi-institutional radiotherapy dose prediction, FEDAFFINE achieves site-macro Metric 1 of 3.52 Gy, compared with 4.03 for FedAvg-FT. Controlled experiments validate the fixed-family risk calculation and find predictive but incomplete alignment with a planted span; the analysis does not guarantee neural recovery of that span.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to learn collaboratively without pooling their raw data (McMahan et al., 2017). When differences between client populations call for different predictive rules, a single global model may not serve every client well. Personalized federated learning (PFL) therefore learns a model tailored to each client while benefiting from collaboration across clients (Smith et al., 2017; Collins et al., 2021; Li et al., 2021a). The goal is better prediction on each client's population, not merely a collection of models that differ from one another.

However, the population differences that motivate personalization are not directly observed. Each client observes only a finite dataset, so a trained model reflects both the underlying prediction task and errors in estimating it. Consider training two models on different samples from the same client population: their predictions may disagree even though the population-optimal rule has not changed. Across clients, the same kind of sample-dependent disagreement is superimposed on genuine population differences. Preserving all observed differences can therefore retain estimation error, whereas enforcing agreement indiscriminately can suppress useful personalization. These are not simply opposing preferences for local or global models: the two sources of disagreement call for different statistical treatment. We call this the *population–estimation ambiguity*, which means that observed model disagreement alone does not reveal which differences should be retained through personalization. Figure 1 summarizes this ambiguity and the structural restriction we introduce below.

Personalization is not merely a choice of which parameters remain local; it encodes a claim about which forms of cross-client variation should persist. Existing PFL designs provide several ways to personalize: regularizing local models toward a global reference (Li et al., 2021a), retaining private prediction heads (Collins et al., 2021), or generating models from client embeddings (Shamsian et al., 2021). These choices specify how predictors are coupled, but do not by themselves establish which empirical differences are population properties. In particular, leaving a component local does not

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.

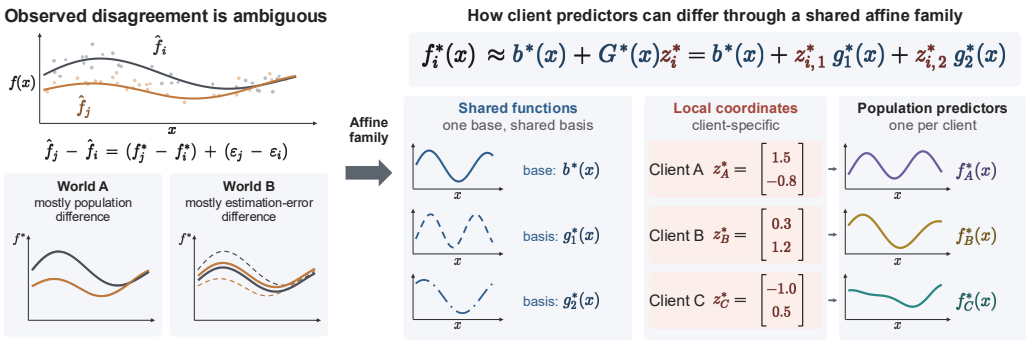

Figure 1: **Population–estimation ambiguity, and how client predictors can differ.** The same trained pair $\widehat{f}_i, \widehat{f}_j$ is compatible with two latent worlds: one in which the gap is mostly a population difference, and one in which it is mostly an estimation-error difference. Restricting client predictors to a shared affine family states which differences are admissible; it does not recover the unrestricted decomposition.

make its learned content population specific. Bayesian and empirical-Bayes approaches address this distinction by modeling how client-specific parameters vary across clients and accounting for uncertainty in estimating these parameters from finite local data (Kotelevskii et al., 2022; Ozkara et al., 2023). We investigate the same statistical problem by asking: *how should client predictors be modeled to accommodate persistent population differences while limiting the influence of finite-sample estimation error?* Our approach is to specify a shared structure of prediction functions, rather than require an explicit probability distribution over client effects.

We hypothesize that persistent differences between client predictors can be approximated by different combinations of a small set of shared functions. Under this hypothesis, each client need not estimate an entire prediction function independently: clients can jointly learn these functions and estimate only their combination coefficients locally. This perspective has a statistical precedent in functional data analysis, where shared patterns of variation support individual estimation from sparse observations (Yao et al., 2005). We express this hypothesis through an affine predictor family, consisting of a shared base function and client-specific linear combinations of shared basis functions:

$$f_i^\star(x) \approx b^\star(x) + \sum_{k=1}^{r_\star} z_{ik}^\star g_k^\star(x). \tag{1}$$

Here, $f_i^\star$ is client i's population-optimal predictor. The shared base $b^\star$ provides a reference predictor, the basis functions $g_k^\star$ describe ways in which predictors can vary, and the coefficients $z_{ik}^\star$ specify their combination for client i. The number $r_\star$ is the dimension of this hypothesized shared family. The base and basis functions may be highly nonlinear in the input; clients differ in the coefficients used to combine them. This is an *affine family* because it adds a shared base to a linear combination of shared functions. Population variation outside the family contributes approximation error.

This assumption changes what is learned collaboratively. The shared functions describe not only a common prediction rule but also how client rules can differ. Distinct coefficient vectors can express distinct predictors while using the same base and basis, so learning these shared functions does not require making all client predictions agree. The hypothesis concerns the organization of heterogeneity: a small set of common functions should account for a substantial part of persistent client variation. It is this relationship among prediction functions, rather than the presence of an additional client variable, that motivates our model.

We implement the family in FEDAFFINE (federated learning with affine predictor families). Shared neural parameters define the base and basis functions. After receiving the current shared model, each client estimates its coefficients from local data with the shared parameters held fixed, followed by joint local updates of the shared parameters and coefficients. Only the shared parameters are communicated. In the reconstructive protocol, local coefficients are estimated anew for the received model and at deployment, so they need not persist between rounds; the Office-Caltech and GDP-HMM specializations instead retain their local identities. Once the family is specified, estimating it

raises further questions about basis dimension and the extent of local adaptation. A sample-weighted linear-estimation benchmark characterizes the corresponding approximation and estimation costs.

Contributions. Our contributions are threefold:

- We formulate an affine model of client population predictors motivated by population–estimation ambiguity, making the structure of client differences an explicit collaborative learning target.
- We develop FEDAFFINE, combining shared base and basis learning with local coefficient estimation. The same shared–local separation supports server-based and decentralized communication.
- We characterize approximation and estimation costs in a fixed-family statistical benchmark, providing a basis for studying model dimension and local adaptation after the predictor family has been defined.

We evaluate FEDAFFINE in four complementary settings: writer-partitioned FEMNIST (Caldas et al., 2019), Office-Caltech domain shift (Li et al., 2021b), label-skewed CIFAR-10/100, and multi-institutional radiotherapy dose prediction. The central application is dense three-dimensional prediction from patient imaging, with hospitals defining the clients (Gao et al., 2025). This setting evaluates the model beyond image classification and artificial client partitions. Label-skewed CIFAR provides a supplementary benchmark; the main application emphasis is on naturally occurring client differences and the medical prediction task.

2 RELATED WORK

We organize personalized federated learning (PFL) by how client predictors are parameterized and estimated. Appendix A gives detailed method-by-method comparisons.

Full-parameter personalization. Full-parameter methods couple complete client models through multitask or cluster structure (Smith et al., 2017; Ghosh et al., 2020; Sattler et al., 2020), client-specific collaboration weights (Zhang et al., 2021; Ye et al., 2023), or initialization, proximal, and interpolation objectives (Fallah et al., 2020; T Dinh et al., 2020; Li et al., 2021a; Deng et al., 2020). They determine how full client estimators interact; our restriction instead concerns the geometry of the resulting prediction functions.

Partial-parameter personalization. FedPer, FedRep, and FedBABU retain client-specific heads over shared representations (Arivazhagan et al., 2019; Collins et al., 2021; Oh et al., 2022); other designs keep local features or normalization, or split shared and local factors (Liang et al., 2020; Li et al., 2021b; Jeong & Hwang, 2022; Guo et al., 2025). Federated Reconstruction estimates a local block on demand (Singhal et al., 2021), while FLUTE studies insufficient shared representations (Liu et al., 2024a). FEDAFFINE instead constrains client predictors to a jointly learned affine span and estimates only their local coordinates in that span.

Auxiliary-variable personalization. pFedHN and PeFLL generate models from client embeddings (Shamsian et al., 2021; Scott et al., 2024); FedEM and FedGMM use mixtures of shared distributions (Marfoq et al., 2021; Wu et al., 2023); and FLoRAL and FLowDUP combine compact client variables with shared low-rank or fixed-random structures (Almansoori et al., 2025; Zakerinia et al., 2026). These are the closest structural precedents. FEDAFFINE places its affine restriction directly on prediction functions—logits for classification—and estimates coefficients from labeled local data. Linear hypernetworks or output adaptors can realize overlapping function families; the distinctions are the coefficient domain, where personalization enters, and how the coefficient is estimated, not merely whether a short client vector exists.

Population variation and estimation uncertainty. Random-effects and functional-data models distinguish persistent individual variation from observation error (Laird & Ware, 1982; Yao et al., 2005), while shared-structure multitask methods use common bases with task-specific coefficients (Ando et al., 2005; Kumar & Daume III, 2012; Ruvolo & Eaton, 2013). Bayesian and empirical-Bayes PFL explicitly model personalized parameters and their uncertainty (Zhang et al., 2022; Kotelevskii et al., 2022; Thapa & Li, 2025; Ozkara et al., 2023); Ditto likewise relates heterogeneity and local

noise to personalization strength (Li et al., 2021a). We study this distinction through an explicit affine function family and its approximation–estimation trade-off. The analysis concerns a fixed family and is not a learning guarantee for the neural implementation.

3 FEDAFFINE: A SHARED AFFINE PREDICTOR FAMILY

3.1 PROBLEM FORMULATION

Consider $N \geq 2$ clients. Client i observes n_i independent samples $S_i = \{(x_{is}, y_{is})\}_{s=1}^{n_i}$ from its distribution $\mathcal{D}_i$. For a prediction function f and loss ℓ, define the population and empirical risks as

$$R_i(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\ell(f(x), y), \qquad \widehat{R}_i(f) = \frac{1}{n_i}\sum_{s=1}^{n_i}\ell(f(x_{is}), y_{is}). \tag{2}$$

Let $\mathcal{F} \subseteq \mathcal{H}$ be a common predictor class, where $\mathcal{H}$ is a Hilbert space of prediction functions (for example, an appropriate L^2 space), and let $f_i^\star \in \arg\min_{f\in\mathcal{F}} R_i(f)$ be a population minimizer. We seek client predictors $\widehat{f}_1, \ldots, \widehat{f}_N$ with small weighted population risk $\sum_{i=1}^N q_i R_i(\widehat{f}_i)$, where $q_i = n_i/n_{\text{tot}}$ and $n_{\text{tot}} = \sum_i n_i$, without exchanging local examples.

For each learned predictor $\widehat{f}_i$, define its estimation error as $\varepsilon_i = \widehat{f}_i - f_i^\star$. Then

$$\widehat{f}_j - \widehat{f}_i = \underbrace{f_j^\star - f_i^\star}_{\text{population difference}} + \underbrace{\varepsilon_j - \varepsilon_i}_{\text{estimation-error difference}}.$$

Observed disagreement alone therefore does not reveal which component should be preserved by personalization (Appendix B.1 gives a concrete example).

Population-level affine approximation. To quantify the approximation in (1), for rank r define the best weighted affine approximation error

$$\mathcal{E}_{\text{aff}}(r) := \inf_{\substack{b,g_1,\ldots,g_r\in\mathcal{H}\\ z_1,\ldots,z_N\in\mathbb{R}^r}} \sum_{i=1}^N q_i \left\| f_i^\star - b - \sum_{k=1}^r z_{ik} g_k \right\|_{\mathcal{H}}^2. \tag{3}$$

Because it is defined in terms of the population predictors $f_i^\star$, this quantity measures structural approximation and contains no finite-sample estimation error. Let $\bar{f}^\star = \sum_j q_j f_j^\star$ denote the weighted mean population predictor, and define the total between-client population variation as $V_{\text{pop}} := \sum_i q_i \left\| f_i^\star - \bar{f}^\star \right\|_{\mathcal{H}}^2$. This is also the approximation error of the best single shared predictor: $\mathcal{E}_{\text{aff}}(0) = V_{\text{pop}}$ (Appendix B.2). Our structural hypothesis is that $\mathcal{E}_{\text{aff}}(r_\star)$ is much smaller than V_{pop} for some modest rank $r_\star$. In other words, a low-dimensional affine family captures most, though not necessarily all, of the between-client population variation.

3.2 NEURAL PARAMETERIZATION AND EMPIRICAL OBJECTIVE

We implement the population-level affine restriction using shared neural parameters ω:

$$f_{\omega,z_i}(x) = b_\omega(x) + G_\omega(x) z_i = b_\omega(x) + \sum_{k=1}^r z_{ik} g_{\omega,k}(x), \qquad z_i \in \mathbb{R}^r, \tag{4}$$

where b_ω is the shared base function and the columns of G_ω are shared basis functions. For classification, f_{ω,z_i} denotes logits; for regression, it denotes the predicted output. The high-dimensional parameters ω are learned collaboratively, whereas client i estimates only the r coefficients in z_i from its local data. The chosen r bounds the dimension of the fitted affine family and need not equal the population approximation rank $r_\star$.

The empirical training problem is

$$\min_{\omega, z_1,\ldots,z_N} \sum_{i=1}^N q_i \widehat{R}_i\big(f_{\omega,z_i}\big). \tag{5}$$

For fixed ω, the map from z_i to $f_{\omega,z_i}(x)$ is affine. Therefore, whenever $\ell(\cdot, y)$ is convex in its prediction argument, the client objective $\widehat{R}_i(f_{\omega,z_i})$ is convex in z_i. This includes cross-entropy applied to affine logits for classification, and squared or absolute loss for regression.

Algorithm 1 FEDAFFINE with server aggregation

Require: Shared initialization ω^0; dimension r; coefficient-estimation settings $(T_{\text{train}}, T_{\text{eval}}, \gamma)$; joint-update steps K; rounds L.

for $t = 0, \dots, L-1$ **do**
  Select clients $\mathcal{S}_t$ and broadcast ω^t.
  for each $i \in \mathcal{S}_t$ in parallel **do**
    $\omega_i \leftarrow \omega^t$ and $z_i \leftarrow 0$.
    With ω_i fixed, take T_{train} local stochastic-gradient steps in z_i as in (6).
    Take K joint local stochastic-gradient steps on $\widehat{R}_i(f_{\omega_i, z_i})$ in (ω_i, z_i).
    Return ω_i and discard z_i.
  end for
  $\omega^{t+1} \leftarrow \sum_{i \in \mathcal{S}_t} \frac{n_i}{\sum_{j \in \mathcal{S}_t} n_j} \omega_i$.
end for
Deployment: freeze ω^L, initialize $z_i \leftarrow 0$, and take T_{eval} coefficient steps on the client's labeled adaptation data as in (6).

3.3 LOCAL COEFFICIENT ESTIMATION AND FEDERATED TRAINING

Given a received shared state ω, client i initializes its coefficient vector at zero and applies T minibatch gradient steps while holding ω fixed:

$$z_i^{(0)} = 0, \qquad z_i^{(s+1)} = z_i^{(s)} - \gamma \nabla_z \widehat{R}_{i, \mathcal{B}_i^{(s)}}\big(f_{\omega, z_i^{(s)}}\big), \quad s = 0, \dots, T-1. \tag{6}$$

Here $\mathcal{B}_i^{(s)} \subseteq S_i$ is the local minibatch at step s. In the reconstructive protocol, the estimate is recomputed for the received shared state, so no client coefficient needs to be retained between rounds.

During training, each selected client first computes its coefficients using T_{train} steps of (6) and then takes K joint local steps in (ω_i, z_i). It returns only ω_i for aggregation and discards z_i. At deployment, the shared state ω^L is frozen. Each participating or held-out client estimates its coefficient using T_{eval} steps on labeled adaptation data kept separate from evaluation data, and then uses f_{ω^L, z_i} for prediction. The full procedure is given in Algorithm 1.

The rank r controls the variation representable by the affine family, whereas T_{train} and T_{eval} control the amount of coefficient fitting within that family. The capacity role of r is analyzed in Section 4. Fixed-map coefficient estimation and finite-step behavior are studied in Appendix B.4, while the dependence on γ, initialization, and basis scale is discussed in Appendix B.4.2.

The method is not restricted to server-based communication. Because only ω is communicated while z_i remains local, server aggregation can be replaced by decentralized communication of the shared state, including gossip or consensus updates as used in decentralized FL (Lian et al., 2017; Koloskova et al., 2020). Appendix C provides the corresponding fixed-state analysis, neural protocol, and a six-site GDP-HMM experiment with random matching.

4 STATISTICAL ANALYSIS OF AFFINE CLIENT HETEROGENEITY

We separate two costs: population variation left outside a rank-r affine family and the finite-sample cost of estimating the retained client-specific coordinates. Proofs and extensions are in Section B.

4.1 POPULATION HETEROGENEITY CAPTURED BY AN AFFINE FAMILY

Write $\delta_i = f_i^\star - \bar{f}^\star$ and define their weighted covariance operator by

$$C = \sum_i q_i (\delta_i \otimes \delta_i), \qquad (u \otimes u)h = \langle u, h \rangle \, u. \tag{7}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ be its eigenvalues, including zeros when necessary.

Proposition 1 (Population approximation by an affine family)**.** *The best rank-r population approximation defined in* (3) *satisfies*

$$\mathcal{E}_{\mathrm{aff}}(r) = \sum_{k>r} \lambda_k, \qquad V_{\mathrm{pop}} = \mathrm{tr}(C) = \sum_{k\geq 1} \lambda_k. \tag{8}$$

An optimal base is $\bar{f}^\star$, and an optimal rank-r variation span is generated by eigenfunctions associated with the largest r eigenvalues.

The family therefore captures $\sum_{k\leq r} \lambda_k$ of the between-client variation. Exact representation is possible once $r \geq \mathrm{rank}(C) \leq N-1$, but low-dimensional personalization is useful only when the spectral tail is already small at a modest rank.

4.2 Estimation cost within the affine family

Suppose the population predictors lie in a fixed d-dimensional subspace $\mathcal{H}_d \subseteq \mathcal{H}$ and client i provides

$$\widehat{f}_i = f_i^\star + \varepsilon_i, \qquad \mathbb{E}\varepsilon_i = 0, \qquad \mathrm{Cov}(\varepsilon_i) = \frac{\sigma^2}{n_i} I, \tag{9}$$

with independent errors across clients (Zaroubi et al., 1994; Li et al., 2021a).

Let $\widehat{f}_{\mathrm{avg}} = \sum_i q_i \widehat{f}_i$. For a deterministic rank-r orthogonal projector P, least squares over an affine family with variation span $P\mathcal{H}_d$ yields

$$\widehat{f}_i^P = \widehat{f}_{\mathrm{avg}} + P(\widehat{f}_i - \widehat{f}_{\mathrm{avg}}) = P\widehat{f}_i + (I-P)\widehat{f}_{\mathrm{avg}}. \tag{10}$$

Proposition 2 (Risk of a fixed affine family)**.** *Under* (9)*, for every client i,*

$$\mathbb{E}\left\|\widehat{f}_i^P - f_i^\star\right\|_{\mathcal{H}}^2 = \underbrace{\|(I-P)\delta_i\|_{\mathcal{H}}^2}_{\textit{population approximation}} + \underbrace{\frac{\sigma^2 r}{n_i}}_{\textit{client-specific estimation}} + \underbrace{\frac{\sigma^2(d-r)}{n_{\mathrm{tot}}}}_{\textit{pooled estimation}}. \tag{11}$$

The three terms are unrepresented population variation, local coefficient estimation, and pooled estimation. Averaging with $q_i = n_i/n_{\mathrm{tot}}$ and choosing the leading population eigenspace gives, for $v := \sigma^2(N-1)/n_{\mathrm{tot}}$,

$$\mathcal{R}(P_r) := \sum_i q_i \mathbb{E}\left\|\widehat{f}_i^{P_r} - f_i^\star\right\|_{\mathcal{H}}^2 = \frac{\sigma^2 d}{n_{\mathrm{tot}}} + \sum_{k>r} \lambda_k + vr. \tag{12}$$

Thus direction r should be personalized exactly when its population energy λ_r exceeds its additional local estimation cost v.

4.3 Capacity and coefficient fitting are distinct

For a fixed span, multiplying direction k by $a_k \in [0,1]$ yields the direction-wise cost $(1-a_k)^2\lambda_k + va_k^2$, minimized by the oracle weight

$$a_k^\star = \frac{\lambda_k}{\lambda_k + v}. \tag{13}$$

Therefore capacity and coefficient regularization are distinct: r controls which variation can be represented, whereas the fitting budgets control how strongly the available coordinates are fitted.

Scope. These are oracle function-space calculations, not classification-risk or neural-learning guarantees. Finite gradient steps need not realize $a_k^\star$, and joint training need not recover the population eigenspace; Section B details these limitations and conditional extensions.

5 Experiments

We evaluate FedAffine on writer-partitioned FEMNIST and six-site GDP-HMM; CIFAR-10/100 and Office-Caltech provide supplementary image benchmarks. Complete splits, optimization settings, inference procedures, and additional results are reported in Section D.

Table 1: **FEMNIST predictive performance under the pFL-Bench protocol.** Mean ± standard deviation over three seeds. Methods retain their native personalization mechanisms; adaptation budgets are selected on participating validation and frozen for held-out clients. Details: Table 7.

	Participating clients			Held-out clients		
Method	Wtd.	Macro	P10	Wtd.	Macro	P10
Reference and generic FL						
Local	$71.32_{\pm 0.48}$	$69.83_{\pm 0.50}$	$58.50_{\pm 0.28}$	$70.75_{\pm 1.29}$	$69.32_{\pm 1.15}$	$57.15_{\pm 2.49}$
FedAvg	$84.45_{\pm 0.56}$	$83.03_{\pm 0.68}$	$69.85_{\pm 3.91}$	$84.80_{\pm 1.17}$	$83.67_{\pm 2.40}$	$71.75_{\pm 4.40}$
FedProx	$83.86_{\pm 1.03}$	$81.81_{\pm 1.63}$	$68.56_{\pm 4.39}$	$84.95_{\pm 1.45}$	$83.40_{\pm 2.29}$	$72.56_{\pm 4.32}$
Full-parameter personalization						
FedAvg-FT	$85.83_{\pm 0.45}$	$84.70_{\pm 0.54}$	$72.28_{\pm 1.11}$	$86.16_{\pm 1.13}$	$85.10_{\pm 2.08}$	$74.91_{\pm 6.34}$
Ditto	$89.48_{\pm 0.23}$	$88.36_{\pm 0.40}$	$79.06_{\pm 1.15}$	$89.41_{\pm 0.46}$	$88.55_{\pm 0.53}$	$81.33_{\pm 2.22}$
pFedMe	$87.81_{\pm 1.31}$	$87.18_{\pm 1.52}$	$76.04_{\pm 4.77}$	$83.47_{\pm 1.43}$	$82.71_{\pm 1.99}$	$70.28_{\pm 4.09}$
APFL	$78.06_{\pm 0.70}$	$76.16_{\pm 0.94}$	$63.54_{\pm 2.84}$	$88.05_{\pm 0.66}$	$86.86_{\pm 1.25}$	$78.56_{\pm 4.48}$
Partial-parameter personalization						
FedBN	$87.26_{\pm 2.19}$	$86.38_{\pm 2.66}$	$75.12_{\pm 4.91}$	$86.00_{\pm 0.68}$	$84.91_{\pm 1.42}$	$76.06_{\pm 3.68}$
FedPer	$64.16_{\pm 1.69}$	$59.98_{\pm 2.65}$	$43.66_{\pm 3.61}$	$68.97_{\pm 1.44}$	$68.49_{\pm 1.47}$	$58.91_{\pm 2.46}$
FedRep	$70.95_{\pm 1.36}$	$69.03_{\pm 1.41}$	$52.22_{\pm 1.92}$	$64.74_{\pm 1.34}$	$64.78_{\pm 1.42}$	$52.58_{\pm 1.51}$
FedBABU	$84.28_{\pm 0.44}$	$82.76_{\pm 0.42}$	$71.58_{\pm 1.08}$	$84.34_{\pm 1.21}$	$82.95_{\pm 1.49}$	$74.30_{\pm 2.55}$
FedRecon	$75.92_{\pm 0.65}$	$74.52_{\pm 0.57}$	$64.29_{\pm 0.41}$	$75.85_{\pm 1.31}$	$74.61_{\pm 1.23}$	$65.07_{\pm 0.63}$
Auxiliary-variable personalization						
pFedHN-PC	$87.40_{\pm 0.42}$	$85.82_{\pm 0.21}$	$74.50_{\pm 3.18}$	$81.90_{\pm 3.56}$	$81.14_{\pm 3.67}$	$69.45_{\pm 7.66}$
FedEM	$84.26_{\pm 0.16}$	$82.52_{\pm 0.11}$	$70.39_{\pm 0.34}$	$84.97_{\pm 1.54}$	$83.54_{\pm 2.70}$	$73.03_{\pm 5.53}$
FLoRAL	$86.22_{\pm 0.84}$	$85.26_{\pm 1.25}$	$75.18_{\pm 1.94}$	$86.46_{\pm 1.10}$	$85.71_{\pm 1.38}$	$68.64_{\pm 4.47}$
Ours						
FedAffine ($r = 3$)	$\mathbf{90.18}_{\pm 0.25}$	$\mathbf{89.32}_{\pm 0.10}$	$\mathbf{80.48}_{\pm 1.00}$	$\mathbf{90.22}_{\pm 0.89}$	$\mathbf{89.43}_{\pm 1.25}$	$81.34_{\pm 4.24}$

Wtd. weights clients by sample count; Macro weights them equally; P10 is the empirical per-client 10th percentile. Held-out clients do not influence model selection. FEDAFFINE reconstructs $r = 3$ coefficients from zero for $T^\star = 200$ steps with the shared map fixed.

5.1 EXPERIMENTAL SETUP

FEMNIST follows the pFL-Bench protocol (Caldas et al., 2019). We report weighted, macro, and lower-tail accuracy for participating and held-out clients over three seeds. FEDAFFINE uses $r = 3$ and 30 from-zero coefficient steps per round. Participating-client validation selects $T^\star = 200$, which is frozen for held-out clients.

GDP-HMM is a radiotherapy dose-prediction dataset in which hospitals define clients (Gao et al., 2025). We report site-macro Metric 1 in Gy on the restored grid. All methods use the same MedNeXt-B backbone and schedule; FEDAFFINE uses $r = 3$ local coefficients.

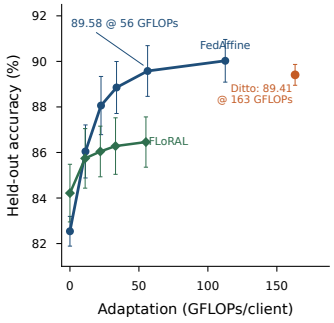

Figure 2: Held-out accuracy versus GFLOPs.

5.2 MAIN RESULTS

FEMNIST. FEDAFFINE has the highest participating-client means and reaches held-out weighted accuracy 90.22 ± 0.89, versus 89.41 ± 0.46 for Ditto. At $T = 50$, it also exceeds Ditto using 56 rather than 163 GFLOP (Figure 2). Participating validation selects $T^\star = 200$; the nearly flat curve from $T = 100$ to 200 indicates convergence rather than an early-stopping optimum. Full-model fine-tuning remains competitive; further details are in Sections D.1 and D.1.2.

GDP-HMM dose prediction. Dose prediction can vary in $p_i(y \mid x)$ because institutions use different planning trade-offs even for comparable anatomy and prescriptions. We apply the affine

Table 2: GDP-HMM test Metric 1 (Gy, ↓). Mean ± standard deviation over seeds $42/43/44$. Macro weights sites equally; micro weights patients equally (n_{test}=12, 9, 30, 7, 7, 33 across sites).

	IMRT			VMAT			All	
Method	S1	S2	S6	S3	S4	S5	Macro	Micro
FEDAFFINE	$\mathbf{3.77}_{\pm.21}$	$\mathbf{3.86}_{\pm.28}$	$3.14_{\pm.13}$	$\mathbf{3.54}_{\pm.28}$	$\mathbf{3.01}_{\pm.07}$	$\mathbf{3.78}_{\pm.26}$	$\mathbf{3.52}_{\pm.15}$	$\mathbf{3.52}_{\pm.10}$
Ditto	$4.12_{\pm.19}$	$4.42_{\pm.28}$	$3.56_{\pm.08}$	$4.00_{\pm.19}$	$3.55_{\pm.19}$	$4.34_{\pm.25}$	$4.00_{\pm.15}$	$4.00_{\pm.11}$
FedAvg-FT	$4.21_{\pm.13}$	$4.29_{\pm.28}$	$3.54_{\pm.09}$	$4.26_{\pm.21}$	$3.52_{\pm.15}$	$4.36_{\pm.28}$	$4.03_{\pm.14}$	$4.02_{\pm.12}$
FedAvg	$4.22_{\pm.16}$	$4.44_{\pm.35}$	$3.60_{\pm.08}$	$4.28_{\pm.20}$	$3.64_{\pm.13}$	$4.44_{\pm.27}$	$4.10_{\pm.15}$	$4.09_{\pm.12}$
FedProx	$4.32_{\pm.34}$	$4.71_{\pm.52}$	$3.76_{\pm.16}$	$4.65_{\pm.38}$	$3.91_{\pm.26}$	$4.52_{\pm.41}$	$4.31_{\pm.29}$	$4.25_{\pm.24}$
Local	$4.25_{\pm.21}$	$4.71_{\pm.20}$	$\mathbf{3.09}_{\pm.15}$	$5.58_{\pm.43}$	$4.67_{\pm.57}$	$3.87_{\pm.28}$	$4.36_{\pm.19}$	$3.93_{\pm.07}$
FedPer	$5.10_{\pm.03}$	$5.88_{\pm.58}$	$3.50_{\pm.16}$	$6.56_{\pm1.4}$	$14.0_{\pm12}$	$4.12_{\pm.29}$	$6.53_{\pm1.8}$	$5.10_{\pm.80}$
FedRep	$5.48_{\pm.24}$	$5.98_{\pm.15}$	$3.64_{\pm.08}$	$9.51_{\pm.35}$	$13.5_{\pm13}$	$4.64_{\pm.14}$	$7.12_{\pm2.2}$	$5.54_{\pm.95}$

factorization to dense 3D regression with $r = 3$ and compare Local, FedAvg, FedAvg-FT, FedProx, FedPer, FedRep, and Ditto under a matched backbone and schedule; detailed implementations and baseline definitions are provided in Section D.2.

At five of the six sites, FEDAFFINE has lower Metric 1 than both Local and FedAvg. On S6, Local is 3.09 ± 0.15 and FEDAFFINE is 3.14 ± 0.13. Per-site differences from FedAvg-FT, patient-level paired comparisons, and a prespecified qualitative slice are reported in Section D.2.

Broader image benchmarks. The CIFAR-10/100 experiments use Dirichlet label skew ($\alpha = 0.1$), 100 clients, ConvNet2, and the reconstructive FEDAFFINE protocol, with $r = 16$ for CIFAR-10 and $r = 32$ for CIFAR-100. Over three seeds, weighted test accuracy is 91.07 ± 0.73 (CIFAR-10) and 60.96 ± 0.60 (CIFAR-100). Ditto reaches 92.01 ± 0.49 and 66.53 ± 1.10, while FedAvg-FT reaches 91.93 ± 0.59 and 66.60 ± 0.72, respectively, under the same splits and checkpoint-selection rule (Table 12). These results assess robustness under label skew, where the full-model personalization baselines remain stronger.

Office-Caltech-10 follows the FedBN four-domain protocol (Amazon, Caltech, DSLR, and Webcam). FEDAFFINE keeps batch-normalization statistics local and uses a persistent $r = 3$ head identity. Over seeds 4–6, it reaches 72.5 ± 0.9 mean test accuracy, compared with 71.4 ± 0.6 for FedBN and 69.0 ± 1.1 for Ditto (Table 11).

5.3 ABLATIONS AND CONTROLLED VALIDATION

Identity capacity and fitting depth. Participating-client validation selects $T^\star = 200$ on every seed. Test accuracy increases with the coefficient budget and is nearly flat from $T = 100$ to $T = 200$ (Figure 4). This is a plateau in predictive performance as coefficient fitting converges, not an interior early-stopping optimum; we therefore do not treat finite-step fitting as a shrinkage knob. Repeating the reconstructive protocol over $r \in \{1, 2, 3, 4, 8\}$ shows that increasing rank does not compensate for insufficient fitting. At $T = 200$, $r \geq 2$ performs comparably to the deployed $r = 3$ operating point (Section D.1.2).

On this ConvNet2, one minibatch step that updates only the three coefficients costs 1.13 GFLOP because it still requires a full forward pass (Figure 2). Held-out Ditto reaches 89.41 with 50 proximal steps (163 GFLOP). FEDAFFINE exceeds that accuracy at $T = 50$ (56 GFLOP, 89.58 ± 1.11) and reaches 90.03 ± 0.94 at $T = 100$ (113 GFLOP). FLoRAL's router adaptation levels off near 86.5, while FedRep remains below 65 at $T = 50$. The FedAvg-FT result is excluded from this step-budget frontier because its evaluation follows a different adaptation procedure (Section D.1.2).

Controlled validation. Gaussian sequence experiments verify the fixed-family risk and rank boundary, while a planted teacher and FEMNIST coefficient-shuffling and random-basis controls show predictive relevance without exact latent-span recovery (Sections D.1.2 and D.5.3).

6 Discussion and limitations

Interpretation and neighboring methods. The population–estimation decomposition does not determine a collaboration action until estimator errors or admissible client predictors are restricted. We choose the latter route: preserve variation within a shared span and pool its complement. Unlike FedRep, the client heads themselves share this variation structure; unlike a general hypernetwork, the predictor is affine in a locally estimated coefficient. Linear generators or output adaptors can nevertheless realize overlapping families, so the contribution is the explicit function-family restriction and its statistical interpretation, not the mere presence of a compact client vector.

Statistical scope. The exact analysis assumes a fixed family and isotropic sequence noise; it does not prove that neural training recovers a persistent population span. Anisotropic estimation error can be mistaken for population structure, and a low-dimensional identity can still memorize a finite client set. The planted teacher establishes client-specific coefficients and partial alignment with a known span, not repeated-sample persistence or exact span recovery.

Capacity versus coefficient fitting. The rank determines which client variation the family can represent, whereas the coefficient budget determines how accurately the available coordinates are fitted. Oracle shrinkage separates these decisions, but finite-step gradient descent need not realize the oracle weights; the FEMNIST sweep shows a fitting plateau rather than an interior early-stopping optimum.

Deployment, privacy, and interpretation. Reconstructive FEMNIST requires labeled local adaptation data, whereas Office-Caltech and GDP-HMM retain selected local state. Keeping coefficients local provides neither differential privacy nor protection from model inversion, and the coordinates are predictive rather than causal. In particular, GDP-HMM improvements do not establish clinical safety, calibration, or deployment readiness. Extended limitations are given in Section E.

7 Conclusion

Personalized federated learning must decide whether observed disagreement reflects persistent population variation or finite-sample estimation error. Existing approaches make this decision using uncertainty information, probabilistic models, auxiliary signals, or structural restrictions. We take the structural route explicitly. Under identity-structured population variation, a high-dimensional shared mechanism can be estimated collaboratively while a short population coordinate remains local. The exact Gaussian analysis yields a direction-selective estimator, a sharp capacity threshold, and a transparent misspecification cost. FedAffine follows this structure by collaborating on the mechanism rather than transmitting the population coordinate. The same shared–local separation admits server aggregation or decentralized mixing of the shared state; Appendix C.4 demonstrates the decentralized extension in the six-site GDP-HMM setting.

The experiments test complementary aspects of this claim rather than universal dominance. On FEMNIST, the method improves participating- and held-out-client accuracy while exposing the cost of coefficient reconstruction; on GDP-HMM, it improves the site-macro dose metric under a common 3D backbone and schedule. The broader image benchmarks show that the affine restriction can help under domain shift but need not outperform full-model personalization under severe label skew. Finally, the planted-span teacher finds client-specific coefficients and partial alignment with a known span, without establishing repeated-sample persistence or recovery of the latent population subspace (Section D.5.3). Together, the results support the affine family as a useful structural bias for personalization, while leaving identification of persistent population directions to future work.

AI USE STATEMENT

We used generative AI tools to assist with language editing, manuscript organization, and LaTeX refactoring. These tools were not used to generate experimental observations or to access patient data. The authors reviewed all AI-assisted text, checked mathematical statements and citations, executed and verified the experimental code, and traced every reported numerical result to the corresponding experiment output. The authors take responsibility for the entire content of the paper.

ETHICS STATEMENT

This study uses existing research datasets and involves no prospective patient recruitment, intervention, or treatment decision. GDP-HMM contains de-identified retrospective radiotherapy planning data distributed for research; we used only the released data and followed its access and data-use conditions. No generative AI system was given access to patient data. The dose models are research prototypes and are not intended for clinical use: their predictions do not establish safety, plan deliverability, or treatment quality without independent dosimetric verification and expert review. Hospital-level performance can also be affected by small and unequal site samples, so we report per-site results and avoid interpreting learned coefficients as causal descriptions of institutional policy. Finally, keeping coefficients local does not provide differential privacy or protection against model inversion.

REPRODUCIBILITY STATEMENT

The method and deployment operator are specified in Algorithm 1, and the statistical assumptions and complete proofs are provided in Section B. The experimental appendix records dataset construction, patient- and client-level splits, preprocessing, architectures, optimization budgets, checkpoint rules, evaluation metrics, seeds, and uncertainty procedures for each benchmark. We report seed-level variation, held-out-client evaluation, and patient-level or site-level results where appropriate. Source code will be released publicly.

REFERENCES

Abdulla Jasem Almansoori, Samuel Horváth, and Martin Takáč. Collaborative and efficient personalization with mixtures of adaptors. In *Conference on Parsimony and Learning*, 2025.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11), 2005.

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

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2019.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=I1hQbx10Kxn.

Huili Chen, Jie Ding, Eric W Tramel, Shuang Wu, Anit Kumar Sahu, Salman Avestimehr, and Tao Zhang. Self-aware personalized federated learning. *Advances in Neural Information Processing Systems*, 35:20675–20688, 2022.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Riqiang Gao et al. Automating radiotherapy planning at scale: High-quality data for AI training. *arXiv preprint arXiv:2501.11803*, 2025.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

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

Wonyong Jeong and Sung Ju Hwang. Factorized-fl: Personalized federated learning with parameter factorization & similarity matching. *Advances in Neural Information Processing Systems*, 35: 35684–35695, 2022.

Salma Kharrat, Marco Canini, and Samuel Horváth. Dpfl: Decentralized personalized federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5086–5094. PMLR, 2025.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

Nikita Kotelevskii, Maxime Vono, Eric Moulines, and Alain Durmus. FedPop: A Bayesian approach for personalised federated learning. In *Advances in Neural Information Processing Systems*, 2022.

Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pp. 963–974, 1982.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021a.

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

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

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. *arXiv preprint arXiv:2001.01523*, 2020.

I-Cheng Lin, Osman Yagan, and Carlee Joe-Wong. FedSPD: A soft-clustering approach for personalized decentralized federated learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2025.

Renpu Liu, Cong Shen, and Jing Yang. Federated representation learning in the under-parameterized regime. In *International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=LIQYhV45D4.

Yingqi Liu, Yifan Shi, Qinglun Li, Baoyuan Wu, Xueqian Wang, and Li Shen. Decentralized directed collaboration for personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems*, 2021.

Connor J McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. *Advances in Neural Information Processing Systems*, 37:77038–77059, 2024.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017.

Konstantin Mishchenko, Rustem Islamov, Eduard Gorbunov, and Samuel Horváth. Partially personalized federated learning: Breaking the curse of data heterogeneity. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=8tMMCf4YYn.

Jaehoon Oh, SangMook Kim, and Se-Young Yun. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=HuaYQfggn5u.

Kaan Ozkara, Antonious M. Girgis, Deepesh Data, and Suhas Diggavi. A statistical framework for personalized federated learning and estimation: Theory, algorithms, and privacy. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FUiDMCr_W4o.

Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pp. 17716–17758. PMLR, 2022.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2023.

Paul Ruvolo and Eric Eaton. Ella: An efficient lifelong learning algorithm. In *International Conference on Machine Learning*, pp. 507–515. PMLR, 2013.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Jonathan Scott, Hossein Zakerinia, and Christoph Lampert. Pefll: Personalized federated learning by learning to learn. In *International Conference on Learning Representations*, volume 2024, pp. 32406–32436, 2024.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, 2021.

Aliaksandra Shysheya, John F Bronskill, Massimiliano Patacchiola, Sebastian Nowozin, and Richard E Turner. Fit: Parameter efficient few-shot transfer learning for personalized and federated image classification. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9aokcgBVIj1.

Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems*, 34:11220–11232, 2021.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 2017.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Mahendra Singh Thapa and Rui Li. Harnessing heterogeneous statistical strength for personalized federated learning via hierarchical bayesian inference. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 59285–59303, 2025.

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *International conference on machine learning*, pp. 37860–37879. PMLR, 2023.

Xianchao Xie, S. C. Kou, and Lawrence D. Brown. SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012. doi: 10.1080/01621459.2012.728154.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *International Conference on Learning Representations*, 2023.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In *International Conference on Machine Learning*, pp. 39801–39817. PMLR, 2023.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Hossein Zakerinia, Jonathan Scott, and Christoph H. Lampert. Federated learning with unlabeled clients: Personalization can happen in low dimensions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wjiw02fqir.

S. Zaroubi, Y. Hoffman, K. B. Fisher, and O. Lahav. Wiener reconstruction of the large scale structure. *arXiv preprint astro-ph/9410080*, 1994.

Chenyu Zhang and Navid Azizan. Personalized collaborative learning with affinity-based variance reduction. In *International Conference on Learning Representations*, volume 2026, pp. 31571–31612, 2026.

Hao Zhang, Chenglin Li, Wenrui Dai, Junni Zou, and Hongkai Xiong. Fedcr: Personalized federated learning based on across-client common representation with conditional mutual information regularization. In *International Conference on Machine Learning*, pp. 41314–41330. PMLR, 2023.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021.

Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, pp. 26293–26310. PMLR, 2022.

APPENDIX CONTENTS

A Connections and Distinctions with Closely Related Methods . . . 15
A.1 Comparison setup . . . 15
A.2 Shared representations, structured heads, and task bases . . . 15
A.3 Hypernetworks, compact codes, and data-dependent model generation . . . 16
A.4 Mixtures, low-rank adaptors, and conditional features . . . 17
A.5 Local estimation, statistical modeling, and guarantees . . . 18
A.6 Communication, resource accounting, and controlled comparisons . . . 19
B Additional Theory and Proofs . . . 20
B.1 An unrestricted observational-equivalence example . . . 20
B.2 Exact Gaussian risk and capacity . . . 20
B.3 Learning the shared basis or filter . . . 23
B.4 Local coefficient estimation and regularization . . . 25
B.5 Complexity of a fixed identity family . . . 25
C Decentralized Extension and Gossip . . . 26
C.1 A one-shot pairwise risk calculation . . . 26
C.2 Weighted consensus in the exact experiment . . . 27
C.3 A neural local-estimation–update–mix protocol . . . 27
C.4 GDP-HMM random-matching experiment . . . 29
D Experimental Details and Additional Results . . . 29
D.1 FEMNIST . . . 30
D.2 GDP-HMM Radiotherapy Dose Prediction . . . 32
D.3 Office-Caltech Domain Generalization and Adaptation . . . 36
D.4 CIFAR Label-Skew Experiments . . . 37
D.5 Controlled Validation of the Statistical Analysis . . . 37
E Extended Discussion and Limitations . . . 40

Table 3: Closest modeling and estimation connections. Each row is expanded below. "Local estimation" refers to obtaining client-specific quantities, not the entire federated training procedure.

Approach	Shared structure and local quantity	Relation to FEDAFFINE
FedRep / structured heads	Shared representation; client-specific head	Our heads lie in a shared affine span. Scalar-output realizations can coincide with a linear-head representation model.
GO-MTL / ELLA	Shared basis tasks; task-specific coefficients	Direct linear-model precedents. Sparsity, sequential updates, and the estimation protocol distinguish the original algorithms.
pFedHN / PeFLL	Shared model generator; client embedding or data descriptor	A linear generator of output heads can realize our family. Data-to-model generation differs from local coefficient optimization.
FLoRAL / FedSA-LoRA	Shared adaptors or factors; local mixture weights or factors	Internal low-rank weight adaptation need not yield an affine prediction function. Output-only specializations can overlap.
FedEM / FedGMM	Shared distribution components; mixture parameters	Fixed-weight output mixtures lie in an affine hull with constrained coordinates; covariate-dependent conditional mixtures need separate treatment.
FLowDUP	Fixed random parameter expansion; data-generated coordinates	Both use low-dimensional coordinates, but the learned object, prediction dependence, and adaptation data differ.
FedRecon	Shared model and an on-demand local-parameter estimator	On-demand estimation is shared. We specify the local affine family and finite-step coefficient procedure.
FedPop / empirical Bayes	Population model and local posterior or personalized estimator	Probabilistic inference can be combined with an affine family; our deployed procedure uses local point estimation.

A CONNECTIONS AND DISTINCTIONS WITH CLOSELY RELATED METHODS

This appendix separates three questions: which client predictors a model can represent, how the client-specific quantities are estimated, and which statistical or optimization guarantees have been established. Two algorithms may represent the same predictor family while using different estimators; conversely, similar parameter counts can conceal different function classes. We first fix the objects being compared and then discuss the nearest structural and statistical precedents.

A.1 COMPARISON SETUP

Write the implemented FEDAFFINE family as

$$f_{\omega,z_i}(x) = b_\omega(x) + G_\omega(x)z_i, \qquad G_\omega(x) = [g_{\omega,1}(x), \dots, g_{\omega,r}(x)], \quad z_i \in \mathbb{R}^r. \tag{14}$$

The parameters ω define shared functions and z_i is a client-level coefficient vector. For regression, f_{ω,z_i} is the predicted output; for classification, it denotes logits, with softmax applied afterwards. With ω fixed, client differences belong to a common function span of dimension at most r. The basis functions themselves can be nonlinear in x. The population hypothesis is approximate: differences outside the chosen family contribute approximation error rather than being classified as sampling noise.

Table 3 summarizes the principal comparisons. Its entries describe modeling objects, not a ranking of expressiveness, accuracy, or statistical validity. In particular, affinity of network weights, affinity of logits, and linear mixing of probabilities are different restrictions.

A.2 SHARED REPRESENTATIONS, STRUCTURED HEADS, AND TASK BASES

Relation to FedRep and partial personalization. FedRep learns a common representation and client-specific heads (Collins et al., 2021). Holding the same representation $\phi_\psi(x) \in \mathbb{R}^m$ fixed,

compare an unrestricted linear head $V_i \in \mathbb{R}^{c\times m}$ with our head construction:

$$f_i(x) = V_i\phi_\psi(x), \qquad V_i = W_0 + \sum_{k=1}^{r} z_{ik}W_k. \tag{15}$$

All W_k are shared. Thus the additional restriction is a common affine span in head space, rather than the fact that the last layer is linear. To see this algebraically, stack the vectorized head deviations as columns:

$$[\mathrm{vec}(V_i - W_0)]_{i=1}^{N} = [\mathrm{vec}(W_1), \ldots, \mathrm{vec}(W_r)][z_1, \ldots, z_N]. \tag{16}$$

The resulting matrix has rank at most r. This is a restriction across clients; it does not require any individual matrix $V_i - W_0$ to have low matrix rank. The functional dimension can be smaller if basis heads are dependent on the support of the shared representation.

This comparison also exposes overlap. For scalar prediction, Equation (14) is a linear head on the augmented representation $(b_\omega, g_{\omega,1}, \ldots, g_{\omega,r})$, with head coefficients $(1, z_i)$. A shared-representation model with this representation and coefficient constraint realizes exactly the same functions. For vector outputs, our coefficients jointly combine vector-valued basis functions across output coordinates. Consequently, a difference in architectural naming is not itself a difference in predictor class. The relevant restriction must be stated relative to a fixed representation, output structure, and parameter budget.

FedPer, FedBABU, FED-ROD, and FedBN provide different shared–local choices (Arivazhagan et al., 2019; Oh et al., 2022; Chen & Chao, 2022; Li et al., 2021b). FedCR, FedPAC, and pFedFDA additionally use representation alignment, classifier combination, or feature-distribution estimates (Zhang et al., 2023; Xu et al., 2023; McLaughlin & Su, 2024). These methods cannot be reduced to a binary layer mask: their objectives and statistical information also matter. FedPAC, for example, estimates client-specific classifier-combination weights using a statistical analysis of prediction risk; our shared basis functions are jointly learned components rather than a prescribed collection of complete client classifiers.

Relation to low-rank multitask learning. Ando et al. (2005) study common predictive structures across tasks. GO-MTL represents task parameter vectors as sparse combinations of latent basis tasks (Kumar & Daume III, 2012); ELLA develops an incremental lifelong-learning algorithm with a shared task basis (Ruvolo & Eaton, 2013). For linear predictors, a model $w_i = w_0 + Lz_i$ gives $f_i(x) = w_0^\top x + z_i^\top L^\top x$, which is precisely a base-plus-basis predictor. This is an algebraic correspondence, not a claim that the algorithms, regularizers, or observation protocols coincide. FEDAFFINE studies the corresponding prediction-space restriction with learned nonlinear base and basis functions, distributed training, and locally estimated coefficients.

FLUTE is especially relevant because it analyzes federated representation learning when the available representation cannot express every ground-truth client model (Liu et al., 2024a). Its sample-complexity and convergence results for linear models explicitly address learning the shared representation in that regime. Our population residual serves a related modeling purpose, but our fixed-family risk calculation is not a substitute for FLUTE's learned-representation analysis.

A.3 HYPERNETWORKS, COMPACT CODES, AND DATA-DEPENDENT MODEL GENERATION

pFedHN maps client embeddings through a shared hypernetwork to generate personalized parameters (Shamsian et al., 2021). If that generator outputs a head through $\mathrm{vec}(V_i) = u_0 + Uz_i$ while a backbone remains shared, its predictions have exactly the form in Equation (15). A nonlinear hypernetwork permits a different, potentially more expressive dependence on z_i. A small embedding alone does not bound the affine dimension of the generated functions: a sufficiently flexible map can associate distinct scalar codes with an arbitrary finite collection of predictors. The function class depends on the generator as well as the code dimension.

Affinity in internal parameters is a separate issue. Even a one-dimensional parameter family can produce non-affine predictions: for $f_\theta(x) = \max\{0, \theta x\}$ and $\theta = z$, the map $z \mapsto f_z(1) = \max\{0, z\}$ is not affine on $\mathbb{R}$. Hence restricting parameters to a linear subspace does not generally impose Equation (14). It does so in special cases, such as a linear final layer on fixed features. This distinction identifies the representation being studied without implying that output-space affinity is always preferable.

PeFLL learns an embedding network that summarizes local data and a hypernetwork that generates client models (Scott et al., 2024). FLowDUP generates low-dimensional personalized parameters from unlabeled data and expands them using a fixed random parameter-space map (Zakerinia et al., 2026). Their data-to-model computations contrast with our local optimization of coefficients against a received shared map. In FLowDUP, low dimensionality is accompanied by an explicit parameter-space restriction; it is not an example of an unrestricted scalar label. Its unlabeled-client objective and transductive PAC-Bayesian analysis also address a different information setting from our supervised local coefficient estimation.

These distinctions suggest two separate comparisons. Holding the generator family fixed isolates optimization-based versus learned data-to-coefficient estimation. Holding the coefficient-estimation procedure fixed isolates affine output conditioning versus nonlinear conditioning or a fixed random parameter map. A comparison that simultaneously changes the backbone, code dimension, and adaptation data cannot attribute an improvement to affinity alone.

A.4 Mixtures, low-rank adaptors, and conditional features

Mixtures are related to, not disjoint from, affine families. FedEM assumes client distributions are mixtures of shared underlying distributions and derives federated EM-like algorithms (Marfoq et al., 2021). For fixed component prediction functions $u_1, \dots, u_K$ and client-level weights π_i in the probability simplex,

$$\sum_{k=1}^{K} \pi_{ik} u_k = u_K + \sum_{k=1}^{K-1} \pi_{ik}(u_k - u_K), \qquad \pi_{ik} \geq 0, \quad \sum_{k=1}^{K-1} \pi_{ik} \leq 1. \tag{17}$$

Thus a fixed-weight mixture is contained in the affine hull of its components. The distinction is the coefficient domain and the interpretation of the component functions, not "mixture versus affine" as mutually exclusive classes. Our regression coefficients are unconstrained real values rather than simplex weights. In classification, mixing logits and then applying softmax is generally different from mixing component probabilities. Equation (17) must be applied to functions of the same type.

A mixture of joint distributions additionally yields input-dependent conditional weights when component input marginals differ. FedGMM explicitly models these marginals using Gaussian mixtures (Wu et al., 2023); it should therefore not be identified with a fixed-coefficient output mixture. This is also why a low-dimensional set of distributional mixture parameters need not imply our fixed affine span of prediction functions.

FLoRAL and parameter factorization. FLoRAL makes task differences themselves collaborative learning objects by combining shared low-rank adaptors with client-specific mixture weights (Almansoori et al., 2025). It is a close precedent both for compact personalization and for weight sharing across the personalized components. In a representative adapted layer its structure is a base weight plus a weighted sum of adaptor matrices. When such adaptors act inside a nonlinear network, the final prediction need not be affine in the mixing weights. When adaptation is restricted to a linear output layer, the families can overlap with Equation (15), subject to adaptor-rank and coefficient constraints.

Factorized-FL separates parameter factors and uses similarity matching for selective aggregation (Jeong & Hwang, 2022). FedSA-LoRA shares the A factor of a LoRA update while retaining a client-specific B_i, so a layer uses an update $B_i A$ (Guo et al., 2025). Our local quantities instead multiply complete shared basis outputs. The distinctions concern the location of personalization, the number and shape of local variables, and the resulting function family. Neither construction is uniformly more expressive under all matched budgets.

FiT uses feature-wise modulation of a pretrained backbone (Shysheya et al., 2023). Intermediate feature conditioning need not impose a shared affine family on the final predictions. A coefficient vector fixed at the client level is also different from a sample-specific routing rule, even if both are represented by short vectors. A comparison should specify both the conditioning location and the adaptation budget.

A.5 LOCAL ESTIMATION, STATISTICAL MODELING, AND GUARANTEES

On-demand local estimation. Federated Reconstruction partitions parameters into shared and local blocks and reconstructs the local block when needed (Singhal et al., 2021). This already supports clients without persistent local parameters and inference for previously unseen clients. Our procedure specifies the local block to be z_i in Equation (14). For a received ω, the implemented estimate is produced by

$$z_i^{(0)} = 0, \qquad z_i^{(t+1)} = z_i^{(t)} - \gamma \nabla_z \widehat{R}_i(f_{\omega, z_i^{(t)}}), \qquad t = 0, \dots, T-1, \tag{18}$$

where $\widehat{R}_i$ is the local empirical risk, T is the number of coefficient-estimation steps, and γ is their step size. Shared parameters remain fixed during these steps. Subsequent local updates jointly modify ω and z_i, but only ω is communicated. The same local estimation procedure is available at deployment. This is an estimator defined by data and the received map, not necessarily an exact minimizer or an amortized inference network. A reproducible stochastic implementation additionally fixes its sampling rule and random seed.

For fixed ω, convexity of the loss in its prediction argument implies convexity in z_i by composition with an affine map. This includes cross-entropy on affine logits. Convexity alone does not guarantee strict convexity or a finite attained unregularized optimum; the actual algorithm is defined by a finite iterate. Shared linear-head methods also possess convex local subproblems for suitable losses. The relevant benefit here is the specified low-dimensional local problem, not an exclusive convexity property. General analyses of simultaneous and alternating shared–local optimization are given by Pillutla et al. (2022); Mishchenko et al. (2025) study an overpersonalized regime in which sufficiently rich local variables permit exact local fitting. Our approximation-aware family need not satisfy that regime.

Population models and uncertainty. Random-effects modeling separates population parameters, individual deviations, and observation errors (Laird & Ware, 1982). Functional principal component methods estimate common covariance eigenfunctions and individual scores from sparse noisy trajectories (Yao et al., 2005). These motivate learning common variation patterns, but our neural basis is optimized for prediction and is not defined as a set of covariance eigenfunctions. Nor is the shared base automatically a population mean unless a centering convention is imposed.

In PFL, pFedBayes uses variational Bayesian inference, FedPop combines common parameters with local random effects, and hierarchical Bayesian models jointly describe global and personalized uncertainty (Zhang et al., 2022; Kotelevskii et al., 2022; Thapa & Li, 2025). Self-FL uses inter- and intra-client uncertainty to guide local updates and aggregation (Chen et al., 2022). Empirical-Bayes formulations provide a statistical framework for personalized estimation and algorithms such as AdaPeD (Ozkara et al., 2023). These approaches go beyond specifying a scalar sharing weight; several include explicit population models and new-client inference. An affine family is compatible with Bayesian inference, so probabilistic modeling and our structural restriction are not opposing categories. Our implemented procedure estimates coefficients directly through the local empirical risk rather than computing a posterior over them.

Scope of the risk benchmark. For a finite-dimensional comparison, take fixed population predictors $f_i^\star \in \mathbb{R}^d$ for $N \geq 2$ clients and observations $\widehat{f}_i = f_i^\star + \varepsilon_i$ with independent centered errors of covariance $(\sigma^2/n_i) I_d$, where $\sigma^2 > 0$. Set $n_{\mathrm{tot}} = \sum_i n_i$, $q_i = n_i / n_{\mathrm{tot}}$, and $\widehat{f}_{\mathrm{avg}} = \sum_i q_i \widehat{f}_i$. For a common deterministic linear filter A, write $\widehat{f}_i^A = \widehat{f}_{\mathrm{avg}} + A(\widehat{f}_i - \widehat{f}_{\mathrm{avg}})$. The weighted squared error $\mathcal{R}(A) = \sum_i q_i \mathbb{E} \|\widehat{f}_i^A - f_i^\star\|_2^2$ satisfies

$$\mathcal{R}(A) = \frac{\sigma^2 d}{n_{\mathrm{tot}}} + \mathrm{tr}\big((I-A) C (I-A)^\top\big) + \frac{\sigma^2 (N-1)}{n_{\mathrm{tot}}} \|A\|_F^2, \tag{19}$$

where $C = \sum_i q_i (f_i^\star - f_{\mathrm{avg}}^\star)(f_i^\star - f_{\mathrm{avg}}^\star)^\top$ and $f_{\mathrm{avg}}^\star = \sum_i q_i f_i^\star$. The terms separate shared estimation, population variation not retained by the filter, and additional local estimation. The optimal linear shrinkage form belongs to established statistical estimation; Ditto and empirical-Bayes PFL already relate optimal personalization to population heterogeneity and estimation noise (Li et al., 2021a; Ozkara et al., 2023). Learning a basis from the same data requires additional analysis beyond this fixed-filter identity.

Table 4: Idealized correspondence, not a derivation. Under a common quadratic idealization of the local objective, each method's solution can be written as a diagonal shrinkage; the table records *in which basis* and *against which quantity* that shrinkage is calibrated. The methods themselves are defined by different objectives, optimizers, and tuning procedures, so the table organizes intuition and does not imply a risk ordering among the implemented algorithms.

Method family	Corresponding shrinkage, in the stated idealization	Local coordinates
FedAvg / Pooled	$a_k \equiv 0$	0
Local / Isolated	$a_k \equiv 1$	d
Ditto, pFedMe	proximal solution $(\widehat{H}_i + \mu I)^{-1}\widehat{H}_i$: shrinkage $h_k/(h_k+\mu)$ in the eigenbasis of the *local information* $\widehat{H}_i$, with one tuned μ	d
FedPer, FedRep, FedBN	$a_k \in \{0,1\}$ on a *pre-assigned* coordinate block	block size
Hard subspace oracle	$a_k \in \{0,1\}$ on a leading *population* eigenspace (Corollary 1)	r
Soft spectral oracle	$a_k = \lambda_k/(\lambda_k + v)$ in the *population* eigenbasis (Equation (28)); not the deployed algorithm	d

This benchmark also differs from a local-optimization filter. For a quadratic loss with positive-definite Hessian H, a ridge penalty $\beta\|z\|_2^2/2$, with $\beta > 0$, produces factors $h/(h+\beta)$ along eigenvectors of H. Thus even one regularization parameter can yield direction-dependent shrinkage. These directions describe local curvature, whereas Equation (19) uses population variation through C. An early-stopped neural coefficient estimator need not reproduce the oracle filter. Likewise, FLoRAL's analysis of stochastic gradients and aggregation, AffPCL's variance reduction for stochastic approximation, and FLUTE's learned-representation guarantees concern distinct objects (Almansoori et al., 2025; Zhang & Azizan, 2026; Liu et al., 2024a).

A.6 Communication, resource accounting, and controlled comparisons

The shared–local separation in FedAffine can be used with server aggregation or decentralized communication of shared parameters. FedEM already provides both communication settings (Marfoq et al., 2021); directed collaboration, DPFL, and FedSPD study personalized decentralized protocols (Liu et al., 2024b; Kharrat et al., 2025; Lin et al., 2025). Our coefficients do not determine a collaboration graph or undergo consensus. This is a specification of which variables are communicated, not a new topology-selection mechanism or a guarantee that arbitrary graphs have identical performance.

Local parameter count, persistent state, and total model size should be reported separately. In Equation (15), r basis heads add rcm shared parameters beyond the base head, even though only r coefficients are personalized. An optional coefficient cache can be eliminated by local estimation, but shared-model storage, optimizer state, cached activations, and adaptation computation remain. PeFLL, Federated Reconstruction, and other partially personalized schemes also support forms of on-demand personalization. Storage saved on a client must not be confused with eliminating state maintained elsewhere in a protocol.

Controlled comparisons should vary one modeling choice at a time: unrestricted versus affine-constrained heads on the same representation; affine versus nonlinear generators at matched budgets; learned versus random bases; and coefficient estimators within the same family. Basis scale matters: replacing G by aG and z_i by z_i/a preserves predictions but changes Euclidean penalties and finite-step updates. Adaptation labels, total local computation, private and shared parameters, and transmitted bytes should be matched or reported explicitly. These controls are needed to attribute gains to the predictor family or its estimator.

An idealized shrinkage comparison. Table 4 organizes the methods under a common quadratic idealization, not as an ordering of their implemented risks. The relevant distinction is not simply scalar versus direction-wise shrinkage: a scalar proximal penalty generally produces different shrinkage coefficients along directions of different local curvature. The oracle in Corollary 1, by contrast, compares population variation with estimation noise. These are different quantities, and neither the oracle population spectrum nor its coefficients are assumed to be available to FedAffine. Whether an implementable approximation improves performance is an empirical question.

B Additional Theory and Proofs

Scope and coordinate conventions. The exact sequence model measures $\|f - f_i^\star\|_{\mathbb{H}}^2$. For squared-error regression, this quantity equals excess risk when $f_i^\star$ is the conditional mean and the norm is defined using client i's input distribution. A common Hilbert norm across clients requires a common design measure or an explicit change-of-measure argument; no analogous equality holds automatically for cross-entropy classification. Accordingly, the exact results below are not relabeled as exact FEMNIST risk identities. The neural predictor is affine in the client coefficient at the logit level, whereas its probabilities after softmax are nonlinear in that coefficient. An orthonormal coordinate system identifies $\mathbb{H}_d$ isometrically with $\mathbb{R}^d$; in these coordinates, the Hilbert–Schmidt norm is the Frobenius norm. Neural parameter vectors ω remain distinct from these function coordinates.

B.1 An unrestricted observational-equivalence example

This counterexample shows why trained-model disagreement alone does not select a collaboration action. It concerns an unrestricted experiment class and permits dependence between client estimation errors.

Theorem 1 (Observational equivalence over unrestricted trained-model experiments)**.** *Fix $d \geq 1$ and $\sigma^2 > 0$. There are two experiments S and H with the same observation law*

$$(Z_i, Z_j) \sim \mathcal{N}(0, \sigma^2 I_d) \otimes \mathcal{N}(0, \sigma^2 I_d), \tag{20}$$

but different unique Bayes-optimal squared-risk actions:

$$\mathsf{S}\ (\textit{shared population, estimation disagreement}): \quad \Theta_i = \Theta_j = \tfrac{1}{2}(Z_i + Z_j), \quad Q_{\mathsf{S}}^\star = \tfrac{1}{2}(Z_i + Z_j), \tag{21}$$

$$\mathsf{H}\ (\textit{heterogeneous populations, precise trained models}): \quad \Theta_i = Z_i,\ \Theta_j = Z_j, \qquad Q_{\mathsf{H}}^\star = Z_i. \tag{22}$$

Consequently, no action based only on the trained models is Bayes optimal in both worlds. More quantitatively, every measurable $Q = Q(Z_i, Z_j)$ satisfies

$$\max_{W \in \{\mathsf{S},\mathsf{H}\}} \left\{ \mathbb{E}_W \left\| Q - \Theta_i \right\|^2 - \inf_{\widetilde{Q}} \mathbb{E}_W \left\| \widetilde{Q} - \Theta_i \right\|^2 \right\} \geq \frac{\sigma^2 d}{8}. \tag{23}$$

Proof of Theorem 1. Let U, V be independent $\mathcal{N}(0, \sigma^2 I_d)$ variables and set $(Z_i, Z_j) = (U, V)$ in both worlds. In S, let $M = (U + V)/2$ and $\Theta_i = \Theta_j = M$. Then the two errors are $\pm(U - V)/2$, centered conditional on M, and the unique squared-risk action is M. In H, set $(\Theta_i, \Theta_j) = (U, V)$, so the unique action for client i is U.

For any $Q = Q(U, V)$, the two excess risks are $\mathbb{E}\left\|Q - M\right\|^2$ and $\mathbb{E}\left\|Q - U\right\|^2$. The parallelogram identity gives

$$\mathbb{E}\left\|Q - M\right\|^2 + \mathbb{E}\left\|Q - U\right\|^2 \geq \tfrac{1}{2}\mathbb{E}\left\|M - U\right\|^2 = \frac{\sigma^2 d}{4}. \tag{24}$$

At least one is therefore at least $\sigma^2 d/8$. □

The construction is deliberately broad. Conditional independence, calibrated error models, repeated estimates, auxiliary validation data, or restrictions on the population predictors can exclude one of the two worlds.

B.2 Exact Gaussian risk and capacity

Throughout this section, the population functions are fixed and expectation is over the independent noises in (9). The filter is deterministic and independent of those noises.

Proof of the population approximation proposition. Let P project onto any rank-r candidate variation span. Optimizing the client coefficients projects each $f_i^\star - b$ onto that span. Since $\sum_i q_i(f_i^\star - \bar{f}^\star) = 0$,

$$\sum_i q_i \left\|(I-P)(f_i^\star - b)\right\|^2 = \mathrm{tr}((I-P)C) + \left\|(I-P)(\bar{f}^\star - b)\right\|^2. \tag{25}$$

The base $b = \bar{f}^\star$ removes the second term. Ky Fan's variational principle then chooses the leading r eigenfunctions of C, leaving $\sum_{k>r} \lambda_k$. At $r = 0$, this is $\mathrm{tr}(C) = V_{\mathrm{pop}}$, proving Proposition 1 including its rank-zero case.

General linear personalization filters. Retain the fixed population functions, sample weights, and second-moment assumptions of Section 4.2, with C defined in (7), and set $s^2 = \sigma^2/n_{\mathrm{tot}}$ and $v = \sigma^2(N-1)/n_{\mathrm{tot}}$. The fixed-projector estimator extends to any common deterministic linear operator A on $\mathcal{H}_d$:

$$\widehat{f}_i^A = \widehat{f}_{\mathrm{avg}} + A(\widehat{f}_i - \widehat{f}_{\mathrm{avg}}), \qquad \mathcal{R}(A) = \sum_i q_i \mathbb{E}\left\|\widehat{f}_i^A - f_i^\star\right\|_{\mathcal{H}}^2. \tag{26}$$

Here $A = 0$ gives pooling, $A = I$ gives local estimation, and $A = P$ gives the fixed affine-family fit. The following calculation specializes classical linear shrinkage to sample-weighted personalization (Zaroubi et al., 1994; Xie et al., 2012).

Proposition 3 (Exact risk of a fixed linear personalization filter)**.** *Let $N \geq 2$, $q_i = n_i/n_{\mathrm{tot}}$, and $\widehat{f}_i = f_i^\star + \varepsilon_i$, where the population functions are fixed and the errors are independent, centered, and have covariance $(\sigma^2/n_i)I$, with $\sigma^2 > 0$. For* (26)*, every common deterministic linear A satisfies*

$$\mathcal{R}(A) = ds^2 + \mathrm{tr}((I-A)C(I-A)^*) + v\left\|A\right\|_{\mathrm{HS}}^2. \tag{27}$$

Here A^* is the adjoint and $\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(AA^*)$. The unique minimizer within this class is $A_\star = C(C+vI)^{-1}$. Gaussianity is not required.

The quantity $\|A\|_{\mathrm{HS}}^2$ measures the variance cost in units of v: it equals r for a rank-r projector and $\sum_k a_k^2$ for a filter with direction-wise coefficients a_k. It is neither the number of trainable parameters nor the usual trace-based degrees of freedom.

Corollary 1 (Oracle capacity and continuous shrinkage)**.** *Let $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ and $u_1, \ldots, u_d$ be the eigenvalues and an orthonormal eigenbasis of C. Among rank-r orthogonal projectors, $P_r = \sum_{k=1}^r u_k \otimes u_k$ minimizes risk, with the risk and marginal difference given by* (12)*. Over all common deterministic linear filters, the optimum is*

$$A_\star = \sum_{k=1}^d a_k^\star u_k \otimes u_k, \qquad a_k^\star = \frac{\lambda_k}{\lambda_k + v}, \qquad \mathcal{R}(A_\star) = ds^2 + \sum_{k=1}^d \frac{\lambda_k v}{\lambda_k + v}. \tag{28}$$

Restricting instead to $\mathrm{rank}(A) \leq r$ gives the optimum risk

$$\mathcal{R}_{\mathrm{soft}}(r) = ds^2 + \sum_{k \leq r} \frac{\lambda_k v}{\lambda_k + v} + \sum_{k>r} \lambda_k. \tag{29}$$

Consequently, increasing the allowed rank from $r-1$ to r reduces the optimally regularized risk by $\lambda_r^2/(\lambda_r + v)$. This oracle risk is nonincreasing: an additional direction can receive coefficient zero. The distinction from the hard-projection threshold in (12) is a consequence of the different estimator classes. Neither calculation identifies the learned neural basis or gives the risk of a filter selected from the same observations. The derivations follow below; dependence from learning the filter is addressed in Section B.3.

B.2.1 Affine least squares as a projection filter

Fix r linearly independent shared basis functions in $\mathcal{H}_d$. Since an invertible change of basis can be absorbed into the unrestricted identity coordinates, choose an orthonormal basis $g_1, \ldots, g_r$ for their span. The associated orthogonal projector is $P = \sum_{k=1}^r g_k \otimes g_k$. In the same notation as the

affine family in Equation (4), consider fitted functions $b + \sum_{k=1}^{r} z_{ik} g_k$, where $b \in \mathcal{H}_d$ is shared and $z_i \in \mathbb{R}^r$ is client-specific. The basis is held fixed throughout this calculation; the base and coordinates are otherwise unrestricted.

We impose $Pb = 0$ to remove redundant representations. This does not restrict the fitted functions: replacing b by $(I - P)b$ and each z_{ik} by $z_{ik} + \langle g_k, b \rangle_{\mathcal{H}}$ leaves every predictor unchanged. This convention separates the shared base from the local coordinates; it does not require the true population base $b^\star$ to be orthogonal to the candidate subspace.

The sample-weighted least-squares problem is

$$(\widehat{b}, \{\widehat{z}_i\}) = \underset{\substack{b \in \mathcal{H}_d,\ Pb=0 \\ z_i \in \mathbb{R}^r,\ i=1,\dots,N}}{\arg\min} \sum_i q_i \left\| \widehat{f}_i - b - \sum_{k=1}^{r} z_{ik} g_k \right\|_{\mathcal{H}}^2 . \tag{30}$$

Orthogonality separates each summand into

$$\left\| (I - P)\widehat{f}_i - b \right\|_{\mathcal{H}}^2 + \left\| P\widehat{f}_i - \sum_{k=1}^{r} z_{ik} g_k \right\|_{\mathcal{H}}^2 . \tag{31}$$

The second term is minimized to zero separately for each client, while the first is minimized by the weighted mean in the orthogonal complement. Therefore,

$$\begin{aligned} \widehat{z}_{ik} &= \left\langle g_k, \widehat{f}_i \right\rangle_{\mathcal{H}}, \qquad \widehat{b} = (I - P)\widehat{f}_{\mathrm{avg}}, \\ \widehat{f}_i^{P} &= \widehat{b} + \sum_{k=1}^{r} \widehat{z}_{ik} g_k = \widehat{f}_{\mathrm{avg}} + P(\widehat{f}_i - \widehat{f}_{\mathrm{avg}}). \end{aligned} \tag{32}$$

This is exactly Equation (26) with $A = P$. The restriction on fitted client differences does not require $f_i^\star - \bar{f}^\star$ to lie in the chosen span. Its component outside that span produces the bias $-(I - P)\delta_i$. The calculation concerns least squares on the Gaussian function observations, not empirical risk minimization on image–label data or joint learning of the neural basis.

B.2.2 The dependence induced by pooling

Let $\bar{\varepsilon} = \sum_j q_j \varepsilon_j$. Because $q_j = n_j / n_{\mathrm{tot}}$ and $\tau_j^2 = \sigma^2 / n_j$,

$$\mathrm{Cov}(\bar{\varepsilon}) = \sum_j q_j^2 \tau_j^2 I = s^2 I, \tag{33}$$

$$\mathrm{Cov}(\varepsilon_i, \bar{\varepsilon}) = q_i \tau_i^2 I = s^2 I, \tag{34}$$

$$\mathrm{Cov}(\varepsilon_i - \bar{\varepsilon}) = (\tau_i^2 - s^2) I, \qquad \mathrm{Cov}(\bar{\varepsilon}, \varepsilon_i - \bar{\varepsilon}) = 0. \tag{35}$$

The last cancellation, not independence between $\widehat{f}_i$ and $\widehat{f}_{\mathrm{avg}}$, drives the risk calculation. Gaussianity additionally implies independence of the pooled noise and the centered local noise, but zero cross-covariance suffices for squared risk.

B.2.3 Proof of the fixed-filter risk proposition

From (26),

$$\widehat{f}_i^{A} - f_i^\star = -(I - A)\delta_i + \bar{\varepsilon} + A(\varepsilon_i - \bar{\varepsilon}). \tag{36}$$

The first term is deterministic, the other two are centered, and their cross-covariance vanishes by (35). Thus, for every client,

$$\mathbb{E} \left\| \widehat{f}_i^{A} - f_i^\star \right\|_{\mathbb{H}}^2 = \|(I - A)\delta_i\|_{\mathbb{H}}^2 + d s^2 + (\tau_i^2 - s^2) \|A\|_{\mathrm{HS}}^2 . \tag{37}$$

The weighted squared bias equals $\mathrm{tr}((I - A)C(I - A)^*)$. Also

$$\sum_i q_i (\tau_i^2 - s^2) = \frac{N\sigma^2}{n_{\mathrm{tot}}} - \frac{\sigma^2}{n_{\mathrm{tot}}} = v. \tag{38}$$

Summing (37) therefore proves (27).

For the minimization, put $M = C + vI \succ 0$. Expanding the trace in an orthonormal coordinate system gives

$$\mathcal{R}(A) - ds^2 = \operatorname{tr}(C) - 2\operatorname{tr}(AC) + \operatorname{tr}(AMA^*). \tag{39}$$

Set $A_\star = CM^{-1}$. Since $A_\star M = C$, completing the square yields

$$\begin{aligned} \mathcal{R}(A) - \mathcal{R}(A_\star) &= \operatorname{tr}((A - A_\star)M(A - A_\star)^*) \\ &= \left\| (A - A_\star)(C + vI)^{1/2} \right\|_{\mathrm{HS}}^2 . \end{aligned} \tag{40}$$

Because M is positive definite, the right-hand side vanishes exactly when $A = A_\star$. This proves uniqueness over all linear operators, not merely diagonal or symmetric ones. Gaussianity can be weakened to independent centered noises with the same covariance structure. No condition that the true functions lie in a rank-r affine family was used: all misspecification remains in the bias term.

B.2.4 Hard capacity and its threshold

For a rank-r orthogonal projector P, $P^2 = P = P^*$ and $\|P\|_{\mathrm{HS}}^2 = r$. Therefore,

$$\mathcal{R}(P) = ds^2 + \operatorname{tr}((I - P)C) + vr. \tag{41}$$

Write $C = \sum_k \lambda_k u_k \otimes u_k$. Then

$$\operatorname{tr}(PC) = \sum_k \lambda_k \|Pu_k\|_{\mathbb{H}}^2 , \quad 0 \le \|Pu_k\|_{\mathbb{H}}^2 \le 1, \quad \sum_k \|Pu_k\|_{\mathbb{H}}^2 = r. \tag{42}$$

With decreasing eigenvalues, this sum is at most $\sum_{k \le r} \lambda_k$: moving weight from a smaller eigenvalue to a larger one never decreases it. The bound is attained by P_r. Substitution gives the first part of (12), and subtraction of adjacent ranks gives $v - \lambda_r$. A tie at $\lambda_r = v$ allows either choice without a risk change.

The per-client form follows directly from (37):

$$\mathbb{E} \left\| \widehat{f}_i^P - f_i^\star \right\|_{\mathbb{H}}^2 = \|(I - P)\delta_i\|_{\mathbb{H}}^2 + \underbrace{\frac{\sigma^2 r}{n_i}}_{\text{local estimation}} + \underbrace{\frac{\sigma^2(d - r)}{n_{\mathrm{tot}}}}_{\text{pooled estimation}} . \tag{43}$$

This makes the meaning of the aggregate additional cost vr explicit.

Diagonalizing C in Proposition 3 gives $A_\star = \sum_k [\lambda_k/(\lambda_k + v)] u_k \otimes u_k$; substituting these coefficients into (27) yields (28). Thus the continuous oracle result follows directly from the fixed-filter risk identity, without a separate comparison to binary filters.

B.2.5 Rank-constrained soft filters

Equation (40) reduces the constraint $\operatorname{rank}(A) \le r$ to the best rank-r approximation of $D = A_\star(C + vI)^{1/2}$. Its singular values are $\lambda_k/\sqrt{\lambda_k + v}$. The Eckart–Young theorem therefore retains the first r oracle coefficients and sets the rest to zero, yielding (29) and

$$\mathcal{R}_{\mathrm{soft}}(r) - \mathcal{R}_{\mathrm{soft}}(r - 1) = -\frac{\lambda_r^2}{\lambda_r + v} \le 0. \tag{44}$$

Thus the nested oracle class has no interior optimal rank; such an optimum for a learned model must arise from basis estimation, optimization, validation noise, or another cost outside this benchmark.

B.3 Learning the shared basis or filter

B.3.1 Covariance shift and the finite-client obstruction

Define the observed covariance operator

$$\widehat{C} = \sum_{i=1}^N q_i \left[(\widehat{f}_i - \widehat{f}_{\mathrm{avg}}) \otimes (\widehat{f}_i - \widehat{f}_{\mathrm{avg}}) \right], \qquad \widehat{f}_{\mathrm{avg}} = \sum_i q_i \widehat{f}_i. \tag{45}$$

Proposition 4 (Expected observed covariance)**.** *In* (9)*,* $\mathbb{E}\widehat{C} = C + vI$*. More generally, with independent centered client noises of covariance* Σ_i*, the covariance shift is* $\sum_i q_i(1-q_i)\Sigma_i$*.*

Proof. Let $\bar{\varepsilon} = \sum_j q_j \varepsilon_j$. Expanding each centered observation gives $\delta_i + (\varepsilon_i - \bar{\varepsilon})$. Signal–noise cross terms vanish in expectation. Since

$$\mathrm{Cov}(\varepsilon_i - \bar{\varepsilon}) = (1-2q_i)\Sigma_i + \sum_j q_j^2 \Sigma_j, \tag{46}$$

the weighted covariance shift is $\sum_i q_i \Sigma_i - \sum_i q_i^2 \Sigma_i$. Under the isotropic sample-size model it becomes

$$\sum_i q_i(1-q_i)\tau_i^2 I = \frac{\sigma^2}{n_{\mathrm{tot}}} \sum_i (1-q_i) I = vI. \tag{47}$$

□

Thus isotropic noise does not rotate the *expected* eigenspaces. Anisotropic noise can add its own coherent directions and change them. The isotropic statement is an assumption-dependent calculation, not a deconvolution result for arbitrary trained models.

There is also a deterministic high-dimensional obstruction. The N centered observations satisfy a nontrivial weighted linear relation, so $\mathrm{rank}(\widehat{C}) \le N-1$. If $N-1 < d$, choose a unit function u in its nullspace. Then

$$\left\| \widehat{C} - (C+vI) \right\|_{\mathrm{op}} \ge \left| \left\langle u, [\widehat{C} - (C+vI)]u \right\rangle \right| = \langle u, Cu \rangle + v \ge v. \tag{48}$$

Therefore no argument can assert an arbitrarily small absolute operator-norm error in this regime at fixed v. A large eigengap, a smaller effective function space, larger local sample sizes, or additional structural assumptions would be needed to make a perturbation bound useful.

B.3.2 A CONDITIONAL SUBSPACE GUARANTEE

Proposition 5 (Perturbation and approximation cost)**.** *Let* $1 \le r < d$*,* $\Delta_r = \lambda_r - \lambda_{r+1} > 0$*, and let* $\widehat{P}_r$ *be a leading rank-*r *projector of* $\widehat{C}$*. On an event where* $\left\| \widehat{C} - (C+vI) \right\|_{\mathrm{op}} \le \varepsilon \le \Delta_r/4$*,*

$$\left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}} \le 2\varepsilon/\Delta_r, \tag{49}$$

$$0 \le \mathrm{tr}((P_r - \widehat{P}_r)C) \le \frac{4r\lambda_1 \varepsilon^2}{\Delta_r^2}. \tag{50}$$

Proof. Adding vI leaves the eigengap and eigenfunctions unchanged. The projector form of the Davis–Kahan inequality under a separated spectrum gives the first bound (Yu et al., 2015). The lower bound follows from optimality of P_r among rank-r projectors. For the upper bound, expand in the eigenbasis of C:

$$\mathrm{tr}((P_r - \widehat{P}_r)C) = \sum_{k \le r} \lambda_k (1 - \left\| \widehat{P}_r u_k \right\|^2) - \sum_{k > r} \lambda_k \left\| \widehat{P}_r u_k \right\|^2 \tag{51}$$

$$\le \lambda_1 \left\| (I - \widehat{P}_r) P_r \right\|_{\mathrm{HS}}^2 \tag{52}$$

$$\le r\lambda_1 \left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}}^2. \tag{53}$$

Substitute the first bound. The result controls only population approximation energy, not the full risk of reusing the same observations for fitting and evaluation. □

This is a deterministic conditional statement. It does not supply the probability of the event, a neural-network recovery rate, or a guarantee that a finite-client covariance is accurate when $N \ll d$. Equation (48) makes one limitation explicit.

B.3.3 WHY A FIXED-FILTER IDENTITY CANNOT BE REUSED CONDITIONALLY ON ITS FIT

A filter learned from $\widehat{f}_1, \ldots, \widehat{f}_N$ depends on the same noise used in (36). Conditioning on that filter can change the noise mean and covariance, so Proposition 3 cannot simply be applied conditionally. An independent pilot would restore the fixed-filter calculation, but the deployed neural algorithm does not use such a split. Thus Proposition 5 controls approximation energy under an assumed covariance error event; it is not a risk guarantee for a basis learned and evaluated on the same client estimates.

B.4 LOCAL COEFFICIENT ESTIMATION AND REGULARIZATION

B.4.1 EXACT EARLY-STOPPING FILTER FOR A FIXED QUADRATIC

Early stopping as spectral regularization is classical (Yao et al., 2007; Raskutti et al., 2014); the calculation here states the fixed-quadratic conditions used to interpret fitting depth.

Proposition 6 (Curvature filter)**.** *Let $J(z) = \frac{1}{2}(z - z_{\min})^\top H(z - z_{\min}) + c_0$ with $H \succ 0$. Gradient descent from $z^{(0)} = 0$ with a constant step size γ gives*

$$z^{(T)} = [I - (I - \gamma H)^T] z_{\min}. \tag{54}$$

If $0 < \gamma \leq 1/\lambda_{\max}(H)$, each coefficient in the eigenbasis of H is $1 - (1 - \gamma h_k)^T \in [0, 1]$ and increases to one.

Proof. Subtracting $z_{\min}$ from one gradient step gives

$$z^{(t+1)} - z_{\min} = (I - \gamma H)(z^{(t)} - z_{\min}). \tag{55}$$

Iterating and substituting $z^{(0)} = 0$ proves (54). Diagonalizing H proves the coordinate expression. The step-size restriction makes $0 \leq 1 - \gamma h_k < 1$. For $1/\lambda_{\max}(H) < \gamma < 2/\lambda_{\max}(H)$, convergence still holds but individual coefficients can oscillate and should not be described as monotone shrinkage. □

This filter uses the local *curvature* eigenbasis. The oracle in (28) uses the between-client *population* eigenbasis. Even if these bases happen to coincide, a single T and γ generally cannot match all oracle coefficients. Cross-entropy further has iterate-dependent curvature. Thus an interior accuracy optimum in T is evidence that fitting depth matters, not a measurement of the oracle filter. Under a fixed quadratic, a scalar ridge penalty similarly gives coefficients $h_k/(h_k + \beta)$, which are direction-dependent whenever the h_k differ.

B.4.2 GAUGE, SCALE, AND INITIALIZATION

The affine family has translation, scale, and basis ambiguities. For invertible Q, replacing G_ω by $G_\omega Q$ and z_i by $Q^{-1} z_i$ preserves predictions. An appropriate shift of all coefficients can also be absorbed into the base. Orthogonalization in a fixed function inner product is a convenient analytical gauge; it is not automatically available as a common federated operation.

A quadratic coefficient penalty or finite Euclidean gradient schedule is not invariant to arbitrary Q. In particular, without controlling the shared basis, the rescaling $G \mapsto aG$, $z_i \mapsto z_i/a$ leaves predictions fixed while reducing the coefficient penalty as $a \to \infty$. Positive ridge still gives a unique *fixed-basis* minimizer, but does not by itself regularize the joint family. Basis weight decay, common normalization, or a fixed scale convention is therefore necessary for an interpretable comparison. These choices must be matched to the actual implementation; one must not insert a function-space orthogonality step into pseudocode without a shared measure and a corresponding implementation.

Initializing every basis readout and every coefficient to zero creates a stationary interaction: the basis gradient is proportional to the coefficient and the coefficient gradient is proportional to the basis. A nonzero basis initialization or another explicit symmetry-breaking step is therefore required.

B.5 COMPLEXITY OF A FIXED IDENTITY FAMILY

B.5.1 WHY IDENTITY DIMENSION NEEDS SMOOTHNESS

Smoothness of the identity-to-function map controls how the predictor varies with identity; it does not by itself bound the complexity of selecting the shared map h. The lemma below isolates the

contribution of an r-dimensional identity when h is fixed. A uniform generalization theorem would require additional entropy or norm conditions on the shared class.

Lemma 1 (Covering cost of a Lipschitz identity family)**.** *Fix a shared map h satisfying*

$$\|h(\cdot, e) - h(\cdot, e')\|_{\mathbb{H}} \le L_z \|e - e'\|_2 \tag{56}$$

for e, e' in the Euclidean ball $B_2^r(B)$. Then, for every $\epsilon > 0$, the induced function family has covering number

$$\log \mathcal{N}\big(\epsilon, \{h(\cdot, e) : e \in B_2^r(B)\}, \|\cdot\|_{\mathbb{H}}\big) \le r \log\left(1 + \frac{2BL_z}{\epsilon}\right). \tag{57}$$

Proof. An (ϵ/L_z)-net of $B_2^r(B)$ maps to an ϵ-net of the function family by Lipschitzness. The standard volumetric bound for a Euclidean ball gives a net of size at most $(1 + 2B/(\epsilon/L_z))^r$. Taking logarithms proves the result. □

Lemma 1 does not bound the complexity of choosing h; that is the shared-class term. Without the Lipschitz condition, no bound depending on r alone is possible.

C Decentralized Extension and Gossip

C.1 A one-shot pairwise risk calculation

We use the function-space notation of Section 4.2. Consider two independent observations

$$\widehat{f}_k = f_k^\star + \varepsilon_k, \qquad \varepsilon_k \sim \mathcal{N}(0, \tau_k^2 I), \qquad k \in \{i, j\}, \tag{58}$$

on $\mathcal{H}_d$, where $\tau_i^2, \tau_j^2 > 0$ and I is the identity on $\mathcal{H}_d$.

Theorem 2 (Population-preserving pairwise operator)**.** *Let P be a fixed rank-r orthogonal projector satisfying $(I - P)(f_j^\star - f_i^\star) = 0$. For $0 \le \rho \le 1$, define*

$$Q_{i \leftarrow j}(\rho) = \widehat{f}_i + \rho(I - P)(\widehat{f}_j - \widehat{f}_i). \tag{59}$$

Then $\mathbb{E}Q_{i \leftarrow j}(\rho) = f_i^\star$ and

$$\mathbb{E}\,\|Q_{i \leftarrow j}(\rho) - f_i^\star\|^2 = r\tau_i^2 + (d - r)[(1 - \rho)^2 \tau_i^2 + \rho^2 \tau_j^2]. \tag{60}$$

When $d > r$, the unique minimizing weight is

$$\rho^\star = \frac{\tau_i^2}{\tau_i^2 + \tau_j^2}, \tag{61}$$

and the receiver's risk reduction relative to local estimation is

$$(d - r)\frac{\tau_i^4}{\tau_i^2 + \tau_j^2}. \tag{62}$$

If $d = r$, communication has no effect and every ρ ties. In contrast, full mixing $\widehat{f}_i + \rho(\widehat{f}_j - \widehat{f}_i)$ has squared bias $\rho^2 \left\|P(f_j^\star - f_i^\star)\right\|^2$.

Proof. The shared-complement condition gives the orthogonal error decomposition

$$Q_{i \leftarrow j}(\rho) - f_i^\star = P\varepsilon_i + (I - P)[(1 - \rho)\varepsilon_i + \rho\varepsilon_j].$$

Centeredness gives unbiasedness; independence and isotropic covariance give Equation (60). For $d > r$, differentiating this strictly convex quadratic gives Equation (61). The minimized complement variance is $(d - r)\tau_i^2\tau_j^2/(\tau_i^2 + \tau_j^2)$; subtracting it from $(d - r)\tau_i^2$ gives Equation (62). Full mixing has bias $\rho(f_j^\star - f_i^\star) = \rho P(f_j^\star - f_i^\star)$, proving the final statement. □

With $\tau_k^2 = \sigma^2/n_k$, the optimum is $\rho^\star = n_j/(n_i + n_j)$, so the communicated complement is the sample-weighted mean of the two observations. This is the two-client instance of Equation (32).

Scope of the calculation. The condition $(I-P)(f_j^\star - f_i^\star) = 0$ is an additional structural assumption, not a consequence of the capacity threshold in Corollary 1: a discarded direction can have small but nonzero population variation. Without this condition, Equation (60) acquires the additional squared bias $\rho^2 \left\|(I-P)(f_j^\star - f_i^\star)\right\|^2$, and Equation (61) need not minimize the risk. The full-mixing comparison identifies a bias cost, not uniform risk dominance over full mixing. After repeated mixing, client states are correlated, so this one-shot independence calculation cannot be applied afresh at each encounter.

C.2 Weighted consensus in the exact experiment

Fix an orthogonal projector P and positive weights q_i summing to one. Let rows of X^0 represent $x_i^0 = (I-P)\widehat{f}_i$ in fixed orthonormal coordinates. Put $J_q = \mathbf{1}q^\top$ and $\|X\|_q^2 = \sum_i q_i \|x_i\|^2$. Consider nonnegative mixing matrices W_t such that $W_t\mathbf{1} = \mathbf{1}$ and $q^\top W_t = q^\top$ almost surely. Let $\mathcal{F}_t$ contain X^0 and all mixing matrices before step t, so X^t is $\mathcal{F}_t$-measurable. Assume the conditional contraction

$$\mathbb{E}[\left\|W_tX^t - J_qX^t\right\|_q^2 \mid \mathcal{F}_t] \le \rho_{\mathrm{net}} \left\|X^t - J_qX^t\right\|_q^2, \qquad 0 \le \rho_{\mathrm{net}} < 1. \tag{63}$$

Proposition 7 (Gossip realizes the fixed-basis estimator)**.** *For $X^{t+1} = W_tX^t$, the weighted average is invariant, and*

$$\mathbb{E}[\left\|X^t - J_qX^0\right\|_q^2 \mid X^0] \le \rho_{\mathrm{net}}^t \left\|X^0 - J_qX^0\right\|_q^2. \tag{64}$$

Thus $P\widehat{f}_i + x_i^t$ converges in mean square to

$$P\widehat{f}_i + (I-P)\widehat{f}_{\mathrm{avg}}, \qquad \widehat{f}_{\mathrm{avg}} = \sum_j q_j \widehat{f}_j, \tag{65}$$

the fixed-basis estimator in Equation (32) *for the same weights.*

Proof. The two invariances imply $W_tJ_q = J_qW_t = J_q$, hence $J_qX^t = J_qX^0$. Apply Equation (63) and the tower property at each step, conditioning on X^0, to obtain Equation (64). Adding the fixed local component gives Equation (65). The conditional bound also yields unconditional mean-square convergence when $\mathbb{E}\left\|X^0\right\|_q^2 < \infty$, as in the exact sequence experiment. □

The component $P\widehat{f}_i$ is unchanged by communication. It is an observed component in the chosen span, not necessarily a noise-free population coordinate; the consensus result alone does not establish that P recovers the population span.

Ordinary doubly stochastic gossip has stationary weights $q_i = 1/N$, so Equation (65) becomes $P\widehat{f}_i + (I-P)N^{-1}\sum_j \widehat{f}_j$. In this case $\|X\|_q^2 = N^{-1}\|X\|_{\mathrm{F}}^2$, giving the same contraction rate in the Frobenius norm. A pairwise update replacing both states by $(q_ix_i + q_jx_j)/(q_i+q_j)$ preserves arbitrary specified q. A connected sampling scheme can then supply contraction; an arbitrary disconnected contact sequence cannot. These weight conventions must not be mixed when comparing to the sample-weighted theory.

Neural implementation boundary. For neural copies ω_i, mixing only the shared parameters expresses a consensus constraint $\omega_1 = \cdots = \omega_N$ while leaving local identities unconstrained. Parameter averaging in this learned network is not function-space orthogonal projection, even though the model is affine in z. Local stochastic gradients, partial participation, coefficient re-estimation, and changing bases all lie outside the preceding fixed-state theorem. The six-site random-matching experiment in Section C.4 evaluates this shared-state protocol but does not establish topology invariance or general nonconvex convergence. Analyses of decentralized SGD address stochastic-gradient and network effects under separate assumptions (Lian et al., 2017; Koloskova et al., 2020).

C.3 A neural local-estimation–update–mix protocol

The following is the decentralized extension of the server prescription. It uses unrestricted identities $z_i \in \mathbb{R}^r$, no explicit identity penalty, and the shared parameterization $f_{\omega_i}(\cdot; z_i)$ from Equation (4).

Algorithm 2 FEDAFFINE-GOSSIP: a specified local-estimation–update–mix extension

Require: shared states $\{\omega_i^0\}$; weights $\{q_i\}$; identity settings (r, T, γ); joint steps K and rates $\alpha_{\omega,t}, \alpha_{z,t}$; rounds L; exogenous matchings $\{\mathcal{M}_t\}$; $0 < \rho < 1$

for $t = 0, \dots, L-1$ **do**
  for all $i = 1, \dots, N$ in parallel **do**
    Set $\omega_i^{t,0} \leftarrow \omega_i^t$.
    Estimate $z_i^{t,0}$ from zero for T steps using (6) with ω_i^t fixed.
    Take K joint steps using (67) on $Nq_i\widehat{R}_i$.
    Set $\widetilde{\omega}_i \leftarrow \omega_i^{t,K}$.
  end for
  for all $(i, j) \in \mathcal{M}_t$ in parallel **do**
    Exchange $\widetilde{\omega}_i, \widetilde{\omega}_j$ only.
    $\omega_i^{t+1} \leftarrow (1-\rho)\widetilde{\omega}_i + \rho\widetilde{\omega}_j$.
    $\omega_j^{t+1} \leftarrow (1-\rho)\widetilde{\omega}_j + \rho\widetilde{\omega}_i$.
  end for
  Unmatched clients set $\omega_i^{t+1} \leftarrow \widetilde{\omega}_i$.
end for
Deployment: each client estimates its coefficient from zero against ω_i^L using local training data.

For fixed positive target weights q_i summing to one, the consensus objective is

$$\min_{\{\omega_i, z_i\}_{i=1}^N} \sum_{i=1}^N q_i \widehat{R}_i(f_{\omega_i}(\cdot; z_i)) \quad \text{subject to} \quad \omega_1 = \cdots = \omega_N. \tag{66}$$

The constraint describes the target shared model, not equality at every iteration; there is no consensus constraint on identities.

Weight convention. Let W_t be nonnegative and doubly stochastic. To represent the target weights in Equation (66), define the local joint-update loss as $\ell_i^{\text{joint}}(\omega, z) = Nq_i\widehat{R}_i(f_\omega(\cdot; z))$, so its uniform average equals the target objective when shared copies agree. The stochastic gradients below are gradients of this scaled loss. Unscaled local mean-loss gradients instead correspond to $q_i = 1/N$ under this convention; the sample-weighted choice uses $Nq_i = Nn_i/n_{\text{tot}}$. This is objective accounting, not a convergence guarantee. It is distinct from preserving a q-weighted average of fixed observations in Proposition 7: the matrices here preserve the uniform parameter average, while the target weights enter through the local losses.

One round. Every client, including those without a communication partner, starts round t by setting $\omega_i^{t,0} = \omega_i^t$ and estimating $z_i^{t,0}$ from zero against that fixed map using the unpenalized local loss and the T-step rule in (6). The coefficient-estimation learning rate is γ; the subsequent joint updates have separate learning rates $\alpha_{\omega,t}, \alpha_{z,t} > 0$. For $k = 0, \dots, K-1$, use

$$\omega_i^{t,k+1} = \omega_i^{t,k} - \alpha_{\omega,t} g_{\omega,i}^{t,k}, \qquad z_i^{t,k+1} = z_i^{t,k} - \alpha_{z,t} g_{z,i}^{t,k}, \tag{67}$$

$$\omega_i^{t+1} = \sum_{j=1}^N [W_t]_{ij} \omega_j^{t,K}. \tag{68}$$

Here the two gradients are evaluated at $(\omega_i^{t,k}, z_i^{t,k})$ using the declared minibatch rule. No projection is applied because the coefficient domain is $\mathbb{R}^r$. Only shared parameters are mixed; a coefficient is never transmitted or used to select a peer. At the next round the coefficient is re-estimated against the new shared copy, rather than updated from the previous round's value.

For an exogenous matching, W_t has block $\left[\begin{smallmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{smallmatrix}\right]$ on each pair and an identity row for each unmatched client. Both assignments use the pre-exchange states. The protocol applies to any matching sequence, but a disconnected sequence does not ensure global consensus; connectivity, mixing, and optimization conditions are additional requirements. The preceding fixed-state theorem does not cover these changing neural parameters and re-estimated coefficients. The update order is coefficient estimation, local joint optimization, and then shared-state mixing.

Table 5: GDP-HMM test Metric 1 (Gy, ↓) for server aggregation versus random-matching FEDAFFINE-GOSSIP. Mean ± standard deviation over seeds $42/43/44$. Site-macro is primary. Server FEDAFFINE and Ditto repeat Table 2 to contextualize the decentralized results.

	IMRT			VMAT			All	
Method	S1	S2	S6	S3	S4	S5	Macro	Micro
FEDAFFINE (server)	$3.77_{\pm.21}$	$3.86_{\pm.28}$	$3.14_{\pm.13}$	$3.54_{\pm.28}$	$3.01_{\pm.07}$	$3.78_{\pm.26}$	$3.52_{\pm.15}$	$3.52_{\pm.10}$
FEDAFFINE-GOSSIP	$3.82_{\pm.19}$	$3.87_{\pm.27}$	$3.23_{\pm.13}$	$3.58_{\pm.28}$	$3.12_{\pm.12}$	$3.98_{\pm.27}$	$3.60_{\pm.16}$	$3.63_{\pm.12}$
Ditto	$4.12_{\pm.19}$	$4.42_{\pm.28}$	$3.56_{\pm.08}$	$4.00_{\pm.19}$	$3.55_{\pm.19}$	$4.34_{\pm.25}$	$4.00_{\pm.15}$	$4.00_{\pm.11}$

Table 6: Shared-state communication on GDP-HMM (N=6, 200 rounds, fp32). Directed η copies count uploads and broadcasts for server methods, and pairwise exchanges for FEDAFFINE-GOSSIP. Identity coordinates are never sent.

Method	Shared params.	Dir. η/round	Id. /round	η bytes
FedAvg / FEDAFFINE (server)	10.86M	12	0	104.2 GB
FEDAFFINE-GOSSIP	10.86M	6	0	52.1 GB

C.4 GDP-HMM RANDOM-MATCHING EXPERIMENT

We evaluate Algorithm 2 on the six-site dose task of Table 2 using random perfect matchings. The experiment evaluates one communication pattern rather than comparing alternative graph topologies. The shared map, identity rank, coefficient budget, optimizer, splits, and warm-start match server FEDAFFINE ($r = 3$ residual fields at `dec_block_1`, $T_e = 30$, ident-reset, FedAvg warm-start, 200 rounds, one local epoch). After the local epoch, a random perfect matching mixes only the shared state η with $\rho = 0.5$; unmatched sites keep their post-local copy. Each site then re-solves e_i from zero against the mixed map. Coefficients are never transmitted or used to choose a peer.

Site-macro Metric 1 is 3.60 ± 0.16 for FEDAFFINE-GOSSIP versus 3.52 ± 0.15 for server FEDAFFINE and 4.00 ± 0.15 for Ditto (Table 5). Gossip is worse than the server affine run by about 0.08 Gy and better than Ditto on every site. Seed 42 re-resolved S1/S2 coefficients against the selected shared map after a pack-level identity write; S3–S6 and seeds $43/44$ use the training-time coefficient. That caveat does not change the comparison: dropping seed 42 leaves gossip at 3.66 site-macro.

Communication is counted as directed copies of the shared state η (fp32), not wall-clock time and not optimizer state. The shared tensor has 10,858,854 floats (41.4 MiB). FedAvg, and server FEDAFFINE, move N uploads and N broadcasts per round (12 directed copies). A perfect matching moves one copy each way on three pairs (6 directed copies). Over 200 rounds that is 1200 versus 2400 directed η transfers, or 52.1 GB versus 104.2 GB of shared-state payload (Table 6). Identity transfers are zero in both affine protocols. Reconstructive training stores no coefficient between rounds; the selected checkpoint keeps three local floats per site. Ditto additionally stores a full private MedNeXt-B; its upload is still a full shared model, so its η byte count is not smaller than FedAvg.

This instance supports the modeling claim that the same shared–local split can run without a server. It does not show that every contact graph matches server aggregation, that the mixing rate equals Equation (64), or that FEMNIST reconstructive gossip has been measured.

D EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

Each dataset-specific section specifies the protocol, evaluation metrics, implementation, inference procedure, and results. Controlled experiments that directly test the statistical analysis are presented separately in Section D.5; they are not conflated with empirical performance on natural federated benchmarks.

Evaluation objectives. The client-specific objective is $R_i(f_i)$, whereas the client-weighted objective is $\sum_i q_i R_i(f_i)$. A pooled-distribution score instead evaluates one predictor under the mixture distribution $\sum_i q_i \mathcal{D}_i$ and is generally different. Personalized evaluation applies client i's predictor to

Table 7: Inference procedures for the pFL-Bench FEMNIST comparison.

Method	Participating clients	Held-out clients
Local	Isolated training	Fresh ConvNet2; isolated early stopping
FedAvg	Shared model	Shared model
FedProx	Shared model, $\mu = 0.1$	Shared model
FedAvg-FT	Released five-step full update	Released five-step full update
Ditto	Stored personalized model	Proximal adaptation, $T^\star$
pFedMe	Stored Moreau model	Moreau local solve
APFL	Stored mixture	Mixture re-estimated from shared checkpoint
FedBN	Local batch-norm state	Statistics on adaptation split
FedPer	Stored classifier	Classifier SGD, $T^\star$
FedRep	Stored representation head	Head SGD, $T^\star$
FedBABU	Classifier SGD, $T^\star$	Classifier SGD, $T^\star$
FedRecon	Reconstructed head	Same reconstruction
pFedHN-PC	Stored embedding; generated classifier	Mean embedding
FedEM	Stored mixture weights	Mixture weights at evaluation
FLoRAL	Router adaptation, $T^\star$	Router adaptation, $T^\star$
FEDAFFINE	$r = 3$, from zero, $T^\star = 200$	Same operator

client i's test distribution and then aggregates across clients, either uniformly or by sample count as specified for each benchmark.

D.1 FEMNIST

FEMNIST is a 62-class handwritten-character dataset derived from LEAF, with writers serving as natural clients (Caldas et al., 2019). All results use the pFL-Bench protocol.

D.1.1 pFL-BENCH PROTOCOL

A writer subsample fraction of 0.05 yields 200 clients with approximately 217 examples per client on average. Each client's examples are divided into $60\%/20\%/20\%$ train, validation, and test splits. Twenty percent of the clients are held out from federated training, and each communication round samples 20% of the remaining client pool. The benchmark uses ConvNet2 with hidden width 2048, batch size 32, and at most 1000 communication rounds, with checkpoint selection based on validation loss. We use the three released seeds. Baselines with released pFL-Bench configurations use their benchmark-selected settings. For methods without a released configuration, we tune method-specific hyperparameters on participating-client validation data and report the best validation-selected configuration. All comparisons use the same data partitions, backbone, client sampling, communication budget, and seeds to ensure a fair comparison.

PeFLL, FedCP, and FLowDUP are excluded because their available results or configurations are not compatible with this protocol. PeFLL uses a 3,597-writer seen/unseen pipeline and a different generated CNN; FedCP has no pFL-Bench FEMNIST-s02 configuration; and FLowDUP lacks a protocol-compatible implementation. A controlled comparison would require architectural adaptation and independent method-specific tuning.

D.1.2 ADDITIONAL pFL-BENCH ANALYSES

Client-specific coefficients and the learned basis. On frozen FEDAFFINE checkpoints, each client estimates its three coefficients from $z_i = 0$ for $T = 200$ steps. Shuffling permutes estimated coefficients among clients within the same split. The random-basis control replaces the learned output basis by a shared i.i.d. $\mathcal{N}(0, 0.01)$ draw and then re-estimates coefficients from zero. Neither control retrains the shared map.

Post-hoc full-model adaptation. Unrestricted post-hoc full-model SGD from the frozen FedAvg checkpoint can reach comparable held-out accuracy. This analysis differs from the released pFL-Bench FedAvg-FT configuration and is therefore reported separately from the primary comparison.

It shows that the FEMNIST result does not imply a universal advantage over unrestricted local adaptation.

Table 8: FEMNIST mechanism analyses under the pFL-Bench protocol. Weighted accuracy (%), mean ± standard deviation over three seeds.

Control	Participating	Held-out
Matched z_i	90.23 ± 0.19	90.27 ± 0.69
Shuffled z_i	77.02 ± 0.60	79.55 ± 2.16
Random basis	81.89 ± 0.42	83.60 ± 0.45

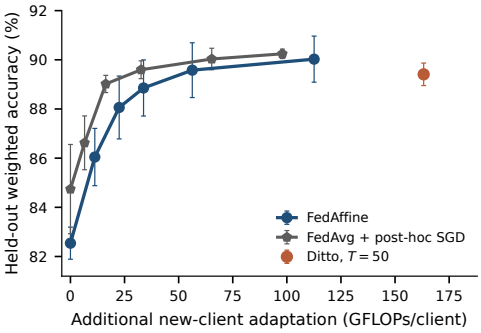

Figure 3: Post-hoc full-model adaptation from the frozen FedAvg checkpoint. The orange marker is Ditto at 89.41% and 163 GFLOPs.

Shuffling reduces participating accuracy by about 13 points and held-out accuracy by about 11 points, showing that the estimated coefficient is client specific. Re-estimation against a random basis performs better than a shuffled assignment but about 8 points worse than the learned basis. These analyses show that the learned basis carries transferable predictive structure; they do not identify it with a ground-truth population subspace.

Capacity–fitting surface. Identity rank r and coefficient budget T are distinct controls. We retrain reconstructive FEDAFFINE on the same pFL-Bench FEMNIST-s02 protocol at $r \in \{1, 2, 3, 4, 8\}$ (three seeds) and then estimate z_i from zero on the frozen shared map at $T \in \{0, 10, 20, 30, 50, 100, 200\}$. The one-dimensional $r = 3$ slice contains the deployed operating point $(r, T) = (3, 200)$. One identity-only minibatch costs 1.13 GFLOP on this ConvNet2 at every r, so GFLOPs scale only with T.

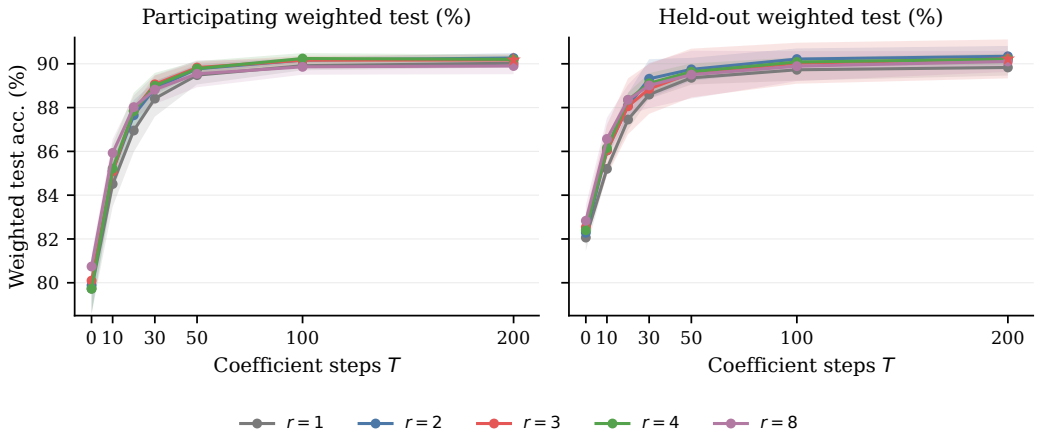

Figure 4: Reconstructive FEMNIST capacity–fitting surface. Each curve is a separately trained identity rank. Markers are mean weighted test accuracy over three seeds; bands are ± one standard deviation. The star is the Table 1 operating point $(r, T) = (3, 200)$.

Participating validation selects $T^\star = 200$ on 14 of 15 rank–seed pairs; one $r = 4$ seed ties $T = 100$ and $T = 200$ on validation and keeps $T = 100$. At $T = 10$, increasing rank from 1 to 8 gains about one point. At $T = 200$, ranks $r \geq 2$ are statistically indistinguishable on both splits, and $r = 8$ is not better than $r = 3$. Extra identity capacity does not replace coefficient fitting, and FEMNIST does

Table 9: Endpoints of Figure 4. Weighted test accuracy (%), mean ± standard deviation over three seeds. $T = 10$ is 11.3 GFLOP; $T = 200$ is 226 GFLOP.

r	T=10 part.	T=10 held-out	T=200 part.	T=200 held-out
1	84.51 ± 1.04	85.20 ± 0.25	90.03 ± 0.20	89.83 ± 0.37
2	85.11 ± 0.77	86.04 ± 0.31	90.26 ± 0.23	90.34 ± 0.47
3	85.09 ± 0.46	86.05 ± 1.16	90.18 ± 0.25	90.22 ± 0.89
4	85.21 ± 1.11	86.15 ± 0.46	90.20 ± 0.21	90.21 ± 0.25
8	85.93 ± 0.62	86.57 ± 0.97	89.90 ± 0.37	90.11 ± 0.48

not reward ranks above the Table 1 choice $r = 3$. The grid isolates the effects of representational capacity and coefficient fitting rather than providing a separate ranking against other methods.

D.2 GDP-HMM Radiotherapy Dose Prediction

D.2.1 Cohort, clients, and patient splits

GDP-HMM is a multi-institutional head-and-neck radiotherapy dose-prediction cohort. A reference dose records one delivered treatment plan rather than a unique deterministic label. For comparable anatomy and prescriptions, institutions may produce different acceptable dose distributions because of planning priorities and delivery techniques. Hospitals therefore define natural clients whose conditional prediction rules may differ.

We define six clients from the source hospitals after merging HMR with CHUS and excluding the mixed-modality remnant `15Ag`/`S215Ag`. Each site is assigned one delivery class, IMRT (`9Ag`/`S29Ag`) or VMAT (`A4Ac`/`S2A4Ac`), and a patient is kept only if that site's plan exists; there is no modality fallback. This retains 487 of 498 challenge patients and drops 11. Each remaining patient contributes one plan. Site membership and modality are fixed across seeds. Within each site, seeds 42, 43, and 44 independently shuffle a $70\%/10\%/20\%$ patient-level train/validation/test split (Table 10). The hospital identifier is not provided as a network input.

Table 10: GDP-HMM client composition and patient-level splits.

Site	S1	S2	S3	S4	S5	S6
Source	TCIA 0522	CHUM	HMR+CHUS	HN other	HNC	HNSCC
Plan	IMRT	IMRT	VMAT	VMAT	VMAT	IMRT
Train	43	31	22	26	113	106
Validation	6	4	3	4	16	15
Test	12	9	7	7	33	30

Volumes stay on the native challenge NPZ voxel grid; we do not resample to a physical millimetre spacing. CT intensities are clipped to $[-1000, 1000]$ HU and divided by 500. Structures are merged into one PTV channel, one optimization-PTV channel, and one OAR channel rather than concatenated per organ. The network input is eight channels: optimization PTV, PTV, OAR, body, CT, beam plate, angle plate, and a spatially tiled prompt (`isVMAT`, number of optimization structures, anatomy site, and a zeroed cohort slot).

Prescriptions come from the challenge `PTV_DICT`: PTV High/Mid/Low prescribed doses in Gy. The reference dose on the original grid is scaled so that PTV-High D_{97} equals the prescribed PTV-High dose, then divided by 10 for loader-space training. Training labels are this loader-space dose multiplied by the body mask.

Geometry is isocenter-centred. Training crops a random box between 0.85 and 1.2 times $(96, 128, 144)$ about the isocenter, applies an in-plane z rotation by a multiple of 40° with probability 0.3 and independent axis flips with probability 0.4, then resizes CT/dose with trilinear interpolation and masks with nearest-neighbour interpolation to $(96, 128, 144)$. Validation and test use a deterministic crop of $(96, 128, 144)$ about the isocenter, with constant padding if the crop exceeds the volume, then the same resize. Because crop size equals network size, the resize is the identity on those splits.

Reported Metric 1 is not computed in crop space. The network output is mapped back to the original NPZ grid by the challenge inference crop: place the $(96, 128, 144)$ prediction in the isocenter-centred window, trim overflow, zero-pad the remainder, and multiply by 10 to recover Gy. The reference dose used for scoring is reloaded from the uncropped NPZ and D97-normalized independently. Predictions that fall outside the restored window are treated as zero dose.

D.2.2 METRICS AND UNCERTAINTY

Let Ω_p denote the body mask for patient p, and let y_p and $\widehat{y}_p$ be the reference and predicted dose in Gy on the original image grid. The challenge metric is

$$\mathrm{Metric1}_p = \frac{\sum_{v \in \Omega_p \cap (\{y_p > 5\,\mathrm{Gy}\} \cup \{\widehat{y}_p > 5\,\mathrm{Gy}\})} |\widehat{y}_p(v) - y_p(v)|}{\sum_{v \in \Omega_p} \mathbf{1}\{y_p(v) > 5\,\mathrm{Gy}\}}. \tag{69}$$

Predictions are uncropped before scoring. A site score is the mean over its test patients. The site-macro score averages the six sites equally and is the primary aggregate. The patient-micro score averages all test patients (n=98 per seed). The worst-site score is the maximum of the six site means. Lower is better.

Negative transfer relative to Local is reported at both levels. At the site level, NT_0 is the fraction of sites whose seed-mean Metric 1 exceeds Local. At the patient level it is the fraction of test plans, pooled over seeds, for which the method is strictly worse than Local. We use no post-hoc clinical tolerance $\epsilon_{\mathrm{clin}} > 0$. DVH endpoints are outside this evaluation because they require a prespecified structure list and do not replace Metric 1 as the challenge primary.

Loader-space scores are used only for checkpoint selection. All numbers in Table 2 and in this appendix are official Gy-valued Metric 1 on the restored image grid. Uncertainty is the sample mean $\pm$ standard deviation over seeds $42/43/44$, matching the main table. Patient-level scatter plots treat plans, not voxels, as the comparison unit.

D.2.3 IMPLEMENTATION AND BASELINES

All methods use the same MedNeXt-B backbone (Roy et al., 2023), voxel-wise ℓ_1 training loss, Adam with learning rate 10^{-4} and cosine decay, batch size 1, and 200 communication rounds (or 200 local epochs for Local) with one local epoch per round. The training crop is $(96, 128, 144)$. Checkpoints are selected by the site-macro mean of loader-space validation Metric 1, then scored with official Gy Metric 1.

The affine dose predictor is

$$h_\eta(x, e) = \sigma\big(y_{0,\eta}(x)\big)s + B_\eta(x)e, \tag{70}$$

with $r = 3$ residual fields taken from `dec_block_1`. Each round resets $e \leftarrow 0$ and takes $T_e = 30$ coefficient steps at learning rate 0.05 on a cache of up to 32 local plans, with the mean identity absorbed into the shared map and no Hessian preconditioner. The shared map is warm-started from the matched-seed FedAvg checkpoint. Only η (including B_η) is aggregated; e_i never leaves the site. Deployment uses the stored site identity from the selected checkpoint. Pairwise mixing of η defines the FEDAFFINE-GOSSIP experiment in Section C.4; it is not part of Table 2.

Baselines share the same splits, backbone, optimizer, batch size, and round budget. FedProx uses proximal weight $\mu = 0.01$. Ditto keeps a second full MedNeXt copy with proximal weight $\lambda = 1$ toward the received global weights. FedPer and FedRep keep the last decoder and readout (`dec_block_0`, `out_0`) private; FedRep uses the two-stage representation-then-head local update. FedAvg-FT fine-tunes the matched-seed FedAvg map independently at each site for 12 epochs with Adam at 2×10^{-5} and cosine decay, selecting the epoch with the best site-level validation Metric 1. pFedHN and mixture methods are omitted because applying them to MedNeXt-B would require method-specific model-generation or mixture architectures beyond a configuration-only rerun.

D.2.4 PER-SITE AND PATIENT-LEVEL RESULTS

Table 2 is the primary ranking. The figures below unpack that table without introducing a second method ranking. Figure 5 shows the per-site change versus unrestricted FedAvg-FT: FEDAFFINE

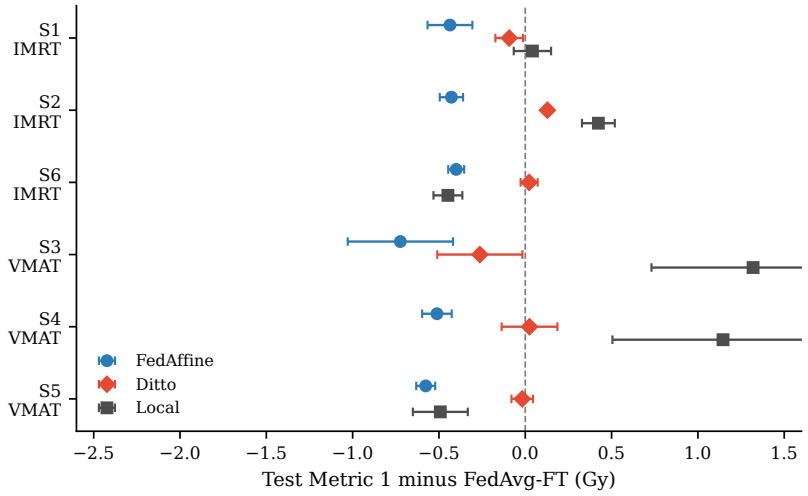

Figure 5: Per-site GDP-HMM test Metric 1 minus FedAvg-FT (Gy). Negative values favor the displayed method. Markers are means over seeds $42/43/44$; bars are sample standard deviations of those three site means.

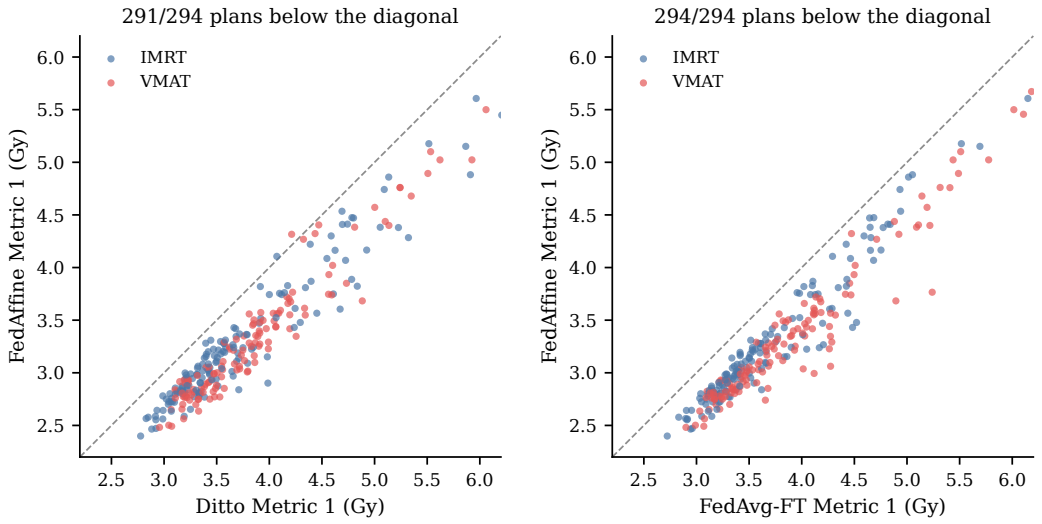

Figure 6: Patient-level test Metric 1 (Gy) on all 294 scored plans (six sites $\times$ three seeds). Each point is one plan. Points below the diagonal favor FEDAFFINE. IMRT and VMAT plans are distinguished only for readability; they are not separate evaluations.

improves every site, including the two small VMAT silos (S3, S4) where Local is unstable, while Ditto is close to FedAvg-FT on several sites. Error bars are the sample standard deviation of the three seed-level site means.

Paired patient-level comparisons (Figure 6) use the same test plans. FEDAFFINE has lower Metric 1 than Ditto on $291/294$ plans and lower Metric 1 than FedAvg-FT on $294/294$ plans, pooled over the three seeds. The three Ditto exceptions are isolated and small; they do not reverse the site ranking. Versus Local, FEDAFFINE is better on $210/294$ plans. The exceptions concentrate on S6, the largest IMRT silo, where Local already fits well (3.09 ± 0.15 versus 3.14 ± 0.13 for FEDAFFINE). Site-level NT_0 versus Local is therefore $1/6$. Worst-site mean Metric 1 (maximum over the six sites, then averaged over seeds) is 3.98 for FEDAFFINE, 4.51 for Ditto, 4.43 for FedAvg-FT, and 5.58 for Local.

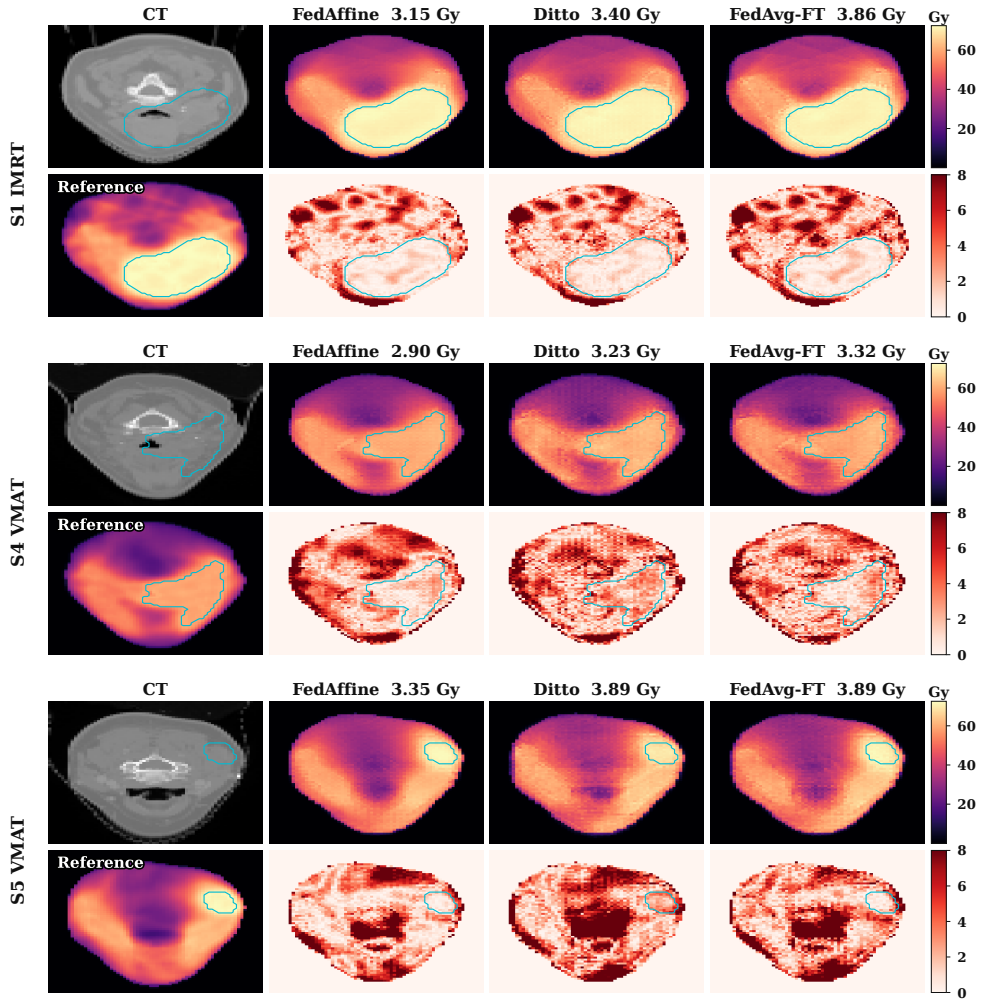

Figure 7: Axial slices through the PTV-High centroid of the median-FEDAFFINE test plan at three sites. First column: CT (top) and reference dose (bottom). Remaining columns: predicted dose (top) and absolute error (bottom). Cyan: PTV High. Metric 1 is printed on each prediction panel.

Figure 7 shows median-error examples rather than selected best-case results. Each pair of rows is one test plan: CT and the reference dose occupy the first column; FEDAFFINE, Ditto, and FedAvg-FT occupy the remaining columns, with the matching absolute-error map underneath. The plan is the median-FEDAFFINE Metric 1 case at S1 (IMRT), S4 (VMAT), and S5 (VMAT). Panel titles report that plan's Metric 1. FedAvg is omitted because it is not the competitive control. Dose panels share a color scale; error panels share a separate scale. The figure supports the corresponding columns of Table 2; it is not evidence about unseen patients or clinical deliverability.

Interpretation and limitations. FedAvg and FedAvg-FT provide shared-only and unrestricted-adaptation controls, while FedPer and FedRep test a private last decoder. Mechanism controls are conducted on FEMNIST rather than MedNeXt-B; consequently, the dose results establish predictive benefit from the affine parameterization but do not isolate recovery of a site-specific mechanism. The learned coefficient is a predictive coordinate, not a validated encoding of institutional planning policy (Section 6).

Table 11: Office-Caltech-10 participating-domain test accuracy (%). Mean$_{\pm \mathrm{sd}}$ over seeds $4/5/6$. Macro: equal domain weight. Worst: minimum of the four domain means. FEDAFFINE uses local batch-norm and a persistent $r = 3$ head identity.

Method	Amazon	Caltech	DSLR	Webcam	Macro	Worst
FEDAFFINE	$\mathbf{63.7}_{\pm 1.5}$	$46.5_{\pm 2.7}$	$\mathbf{87.5}_{\pm 3.1}$	$92.1_{\pm 1.0}$	$\mathbf{72.5}_{\pm 0.9}$	46.5
FedBN	$62.5_{\pm 1.0}$	$\mathbf{47.4}_{\pm 1.1}$	$85.4_{\pm 1.8}$	$90.4_{\pm 1.0}$	$71.4_{\pm 0.6}$	**47.4**
Ditto	$57.1_{\pm 0.8}$	$41.9_{\pm 0.9}$	$84.4_{\pm 3.1}$	$\mathbf{92.7}_{\pm 1.0}$	$69.0_{\pm 1.1}$	41.9
pFedMe	$56.8_{\pm 1.4}$	$41.6_{\pm 0.3}$	$84.4_{\pm 3.1}$	$89.8_{\pm 3.4}$	$68.2_{\pm 2.0}$	41.6
Local	$54.7_{\pm 4.5}$	$36.3_{\pm 3.5}$	$75.0_{\pm 11}$	$81.4_{\pm 16}$	$61.8_{\pm 8.9}$	36.3
FedPer	$53.1_{\pm 2.3}$	$44.9_{\pm 0.9}$	$66.7_{\pm 4.8}$	$81.9_{\pm 2.0}$	$61.7_{\pm 1.8}$	44.9
FedAvg	$51.7_{\pm 4.8}$	$44.7_{\pm 3.0}$	$70.8_{\pm 7.9}$	$78.5_{\pm 12}$	$61.5_{\pm 6.9}$	44.7
FedProx	$52.3_{\pm 5.0}$	$44.6_{\pm 5.1}$	$66.7_{\pm 7.9}$	$77.4_{\pm 13}$	$60.2_{\pm 7.3}$	44.6
FedRep	$49.1_{\pm 0.3}$	$41.3_{\pm 0.8}$	$52.1_{\pm 3.6}$	$80.8_{\pm 1.0}$	$55.8_{\pm 1.2}$	41.3
pFedHN	$38.7_{\pm 8.4}$	$34.2_{\pm 4.2}$	$51.0_{\pm 1.8}$	$63.3_{\pm 14}$	$46.8_{\pm 6.7}$	34.2

D.3 OFFICE-CALTECH DOMAIN GENERALIZATION AND ADAPTATION

D.3.1 PROTOCOL

Office-Caltech-10 treats Amazon, Caltech, DSLR, and Webcam as four domain-defined clients under the FedBN protocol (Li et al., 2021b). All domains share the same ten classes. Each domain uses 62 training, 50 validation, and $192/225/32/59$ test examples (Amazon/Caltech/DSLR/Webcam). The backbone is AlexNet with batch-norm, trained with SGD at learning rate 10^{-2}, batch size 32, one local epoch per round, and 300 rounds. Aggregation is equal-client, not example-weighted. Checkpoint selection uses the domain-macro validation accuracy; reported numbers are the corresponding test accuracies. Seeds are 4, 5, and 6.

FEDAFFINE on this benchmark keeps batch-norm local (the FedBN special case) and uses a persistent head identity of dimension $r = 3$, fitted for 50 steps at learning rate 10^{-2}. The coefficient is not reset between rounds. This is a frozen Office-specific specialization, not the reconstructive from-zero protocol used on FEMNIST and CIFAR. Shared-BN FEDAFFINE (identity only, BN aggregated) serves as a control. Every domain participates in shared training, so this experiment does not evaluate leave-one-domain-out adaptation.

D.3.2 METRICS AND BASELINES

The primary aggregate is the unweighted mean of the four domain test accuracies. We also report the worst domain. Baselines use the same splits, backbone, optimizer, and round budget: Local (single-domain training), FedAvg, FedProx ($\mu = 10^{-2}$), FedBN, FedPer, FedRep, Ditto ($\lambda = 0.05$), pFedMe, and pFedHN.

D.3.3 RESULTS

Table 11 is the complete three-seed participating-domain comparison. FEDAFFINE has the highest domain-macro mean. The gain over FedBN is 1.1 points and is concentrated on Amazon and DSLR; FedBN remains slightly better on Caltech, the worst domain for every method. Private-head FedPer/FedRep and pFedHN are below FedAvg. A shared-BN FEDAFFINE control (seeds 4–5) averages 65.0%, close to FedAvg, so local batch-norm is required for the Office operating point.

D.3.4 OFFICE-CALTECH ABLATIONS

On seed 4, replacing the persistent identity with a reset-to-zero eval-only fit, freezing e during shared training, or raising r to 8 does not improve domain-macro accuracy over the $r = 3$ no-reset operating point (73.5%). An $r = 0$ local-BN run (no identity) is 72.3%. These single-seed ablations assess the contributions of identity persistence, joint training, and rank. Coefficient-shuffling and random-basis analyses are reported on FEMNIST in Section D.1.2. With only four Office clients, this benchmark is not used for population-subspace visualization.

Table 12: CIFAR Dirichlet $\alpha = 0.1$, 100 clients, ConvNet2, three seeds. Weighted / macro / P10 test accuracy (%) at the validation-loss checkpoint. No unseen clients. FEDAFFINE is reconstructive, $r = 16$ (CIFAR-10) and $r = 32$ (CIFAR-100).

	CIFAR-10			CIFAR-100		
Method	Weighted	Macro	P10	Weighted	Macro	P10
Ditto	$\mathbf{92.01}_{\pm.49}$	$89.05_{\pm1.4}$	$76.2_{\pm4.1}$	$66.53_{\pm1.1}$	$66.21_{\pm1.1}$	$56.2_{\pm2.5}$
FedAvg-FT	$91.93_{\pm.59}$	$\mathbf{89.47}_{\pm.94}$	$\mathbf{77.6}_{\pm.39}$	$\mathbf{66.60}_{\pm.72}$	$\mathbf{66.24}_{\pm.72}$	$\mathbf{56.7}_{\pm2.4}$
FedBN	$91.19_{\pm.34}$	$88.06_{\pm1.7}$	$73.8_{\pm4.2}$	$46.00_{\pm2.0}$	$45.77_{\pm1.9}$	$36.7_{\pm2.0}$
FEDAFFINE	$91.07_{\pm.73}$	$88.59_{\pm.64}$	$76.0_{\pm2.1}$	$60.96_{\pm.60}$	$60.48_{\pm.48}$	$50.9_{\pm2.9}$
pFedHN	$89.11_{\pm1.4}$	$85.61_{\pm1.6}$	$69.4_{\pm3.3}$	$38.42_{\pm1.5}$	$38.14_{\pm1.4}$	$26.5_{\pm1.0}$
FedPer	$87.69_{\pm.98}$	$83.73_{\pm1.3}$	$64.0_{\pm4.9}$	$45.23_{\pm1.3}$	$44.80_{\pm1.4}$	$32.9_{\pm.37}$
pFedMe	$87.39_{\pm.55}$	$87.33_{\pm1.1}$	$69.8_{\pm3.9}$	$53.15_{\pm2.6}$	$53.02_{\pm2.8}$	$38.0_{\pm2.4}$
Local	$83.13_{\pm1.2}$	$77.72_{\pm1.1}$	$51.0_{\pm3.4}$	$38.09_{\pm.78}$	$37.57_{\pm.88}$	$25.2_{\pm.95}$
FedRep	$83.12_{\pm1.6}$	$78.14_{\pm.29}$	$55.8_{\pm2.6}$	$38.44_{\pm.31}$	$37.91_{\pm.44}$	$25.3_{\pm.58}$
FedAvg	$65.00_{\pm.45}$	$65.11_{\pm1.1}$	$47.4_{\pm2.4}$	$38.62_{\pm.64}$	$38.54_{\pm.58}$	$28.1_{\pm1.5}$
FedEM	$64.82_{\pm1.5}$	$64.57_{\pm2.5}$	$45.5_{\pm7.8}$	$40.59_{\pm1.0}$	$40.48_{\pm1.0}$	$31.0_{\pm1.6}$

D.4 CIFAR LABEL-SKEW EXPERIMENTS

We evaluate CIFAR under Dirichlet label skew. Because the construction has no known continuous client factor, it serves as a robustness benchmark rather than ground-truth validation of an affine population model. It is not the pFL-Bench CIFAR-10-$\alpha0.1$ protocol (that benchmark uses unseen clients and 1000 rounds). Published pFL-Bench CIFAR numbers are not mixed with the table below.

D.4.1 DIRICHLET PROTOCOL

CIFAR-10 and CIFAR-100 each use 100 clients, per-class Dirichlet concentration $\alpha = 0.1$, and a 4:1:1 train/validation/test split of each client's examples. There are no held-out clients. Participation is 20% per round (20 clients), one local epoch, 400 rounds, batch size 32, ConvNet2 with hidden width 512, and SGD at learning rate 0.1. Checkpoints are selected by weighted validation loss; reported test numbers are from that checkpoint. Seeds are 1, 2, and 3. The primary metric is the sample-weighted test accuracy; we also report client-macro accuracy and the empirical tenth percentile P10 (FederatedScope `test_acc_bottom_decile`).

FEDAFFINE uses the reconstructive FEMNIST protocol: identity rank $r = 16$ on CIFAR-10 and $r = 32$ on CIFAR-100, frozen from a seed-1 capacity sweep; $T_e = 30$ from-zero coefficient steps each round; evaluation reconstructs z_i from zero for 200 steps with the shared map fixed. Ditto uses proximal weight $\lambda = 0.05$ and 5 personalized local steps. FedAvg-FT fine-tunes the full model for 200 steps at learning rate 0.01 before each evaluation. All other methods share the same split, backbone, sampling, and checkpoint rule.

D.4.2 RESULTS

Ditto and FedAvg-FT are the strongest methods on both datasets (Table 12). FEDAFFINE is within about one point of them on CIFAR-10 (91.07 ± 0.73 versus 92.01 ± 0.49 / 91.93 ± 0.59) and about five to six points behind on CIFAR-100 (60.96 ± 0.60 versus 66.53 ± 1.10 / 66.60 ± 0.72). FedBN matches FEDAFFINE on CIFAR-10 and collapses on CIFAR-100. Local and FedRep remain far below full-model personalization. These results do not support a claim that the affine family dominates label-skew classification; the reconstructive configuration is transferred from FEMNIST rather than optimized specifically for CIFAR.

D.5 CONTROLLED VALIDATION OF THE STATISTICAL ANALYSIS

These experiments use known population structure. They test the fixed-family statistical calculations and the ability of a controlled learning problem to recover a useful predictor family. Natural-dataset accuracy is not used as a surrogate for population-subspace recovery.

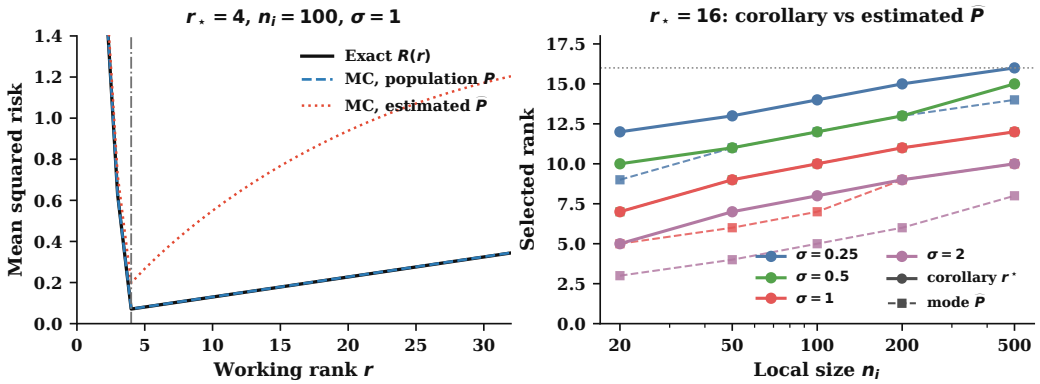

Figure 8: Gaussian sequence validation. Left: exact $R(r)$ versus Monte Carlo risk at $r_\star = 4$, $n_i = 100$, $\sigma = 1$. The population projector tracks the formula; the estimated covariance $\widehat{P}$ pays extra risk past the corollary rank. Right: planted $r_\star = 16$. Solid markers are the corollary rank; dashed markers are the mode of $\widehat{P}$. The dotted line is the planted dimension. Weak eigenvalues fall below the cutoff before $r = 16$ except at the largest n_i and smallest σ.

D.5.1 Exact Gaussian numerical validation

The numerical validation uses seed 20260913. The local size vector repeats $(20, 30, 40, 60, 80, 50)$ four times. With $\sigma^2 = 1$, this gives $n_{\text{tot}} = 1120$, $s^2 = 1/1120$, and $v = 23/1120$. The population eigenvalues, divided by v, are

$$(8, 4, 2, 1, 0.5, 0.25, 0.1, 0.05, 0, 0, 0, 0). \tag{71}$$

The realized population effects are constructed to have exactly this weighted covariance: we project a Gaussian $N \times d$ matrix orthogonally off $\sqrt{q}$, orthonormalize the result, and rescale its columns by the prescribed eigenvalues. For each of 20,000 independent noise draws, risk is evaluated for all hard and rank-constrained soft filters, with Monte Carlo standard errors computed across complete draws. The largest absolute discrepancy between an empirical mean and exact risk is 1.332 Monte Carlo standard errors. A separately generated non-symmetric filter has exact risk 0.557580, empirical mean 0.557425, and standard error 0.000191. This validates the numerical implementation of the fixed-filter formulas, not the learned neural method.

D.5.2 Gaussian phase diagram

This grid tests the sequence estimator $\widehat{\theta}_i(P) = PZ_i + (I - P)\overline{Z}$ against the closed-form risk, not the neural method. Client means follow

$$\theta_i = \mu + Bz_i + \gamma C w_i \tag{72}$$

in $d = 128$ with $N = 40$ equal-sized clients. The planted energies are geometric, $\alpha_k = 4 \cdot 0.55^{k-1}$, and $[B, C]$ is QR-orthonormal. The well-specified grid is $\gamma = 0$ crossed with

$$r_\star \in \{1, 2, 4, 8, 16\}, \quad n_i \in \{20, 50, 100, 200, 500\}, \quad \sigma \in \{0.25, 0.5, 1, 2\}, \tag{73}$$

one hundred cells, one hundred Monte Carlo draws per cell, and consecutive ranks $r = 0, \ldots, 32$. Selected risk curves use five hundred additional draws. Rank is selected by the uncalibrated cutoff $\lambda_k > \sigma^2(N - 1)/n_{\text{tot}}$. The grid uses balanced client sizes; misspecification is evaluated at $(r_\star, n_i, \sigma, \gamma) = (4, 100, 1, 0.5)$ with five hundred draws.

Across the hundred well-specified cells the mean relative discrepancy $|R_{\text{MC}} - R_{\text{exact}}|/R_{\text{exact}}$ at the population projector is 0.0365, and the corollary rank lies in the exact argmin of $R(r)$ on every draw (ties excluded at relative gap 10^{-3}). The Monte Carlo argmin rank using the population projector matches the corollary rank on 94.3% of draws. Using the sample covariance $\widehat{P}$ the match falls to 71.0%: $\widehat{P}$ systematically underselects when several weak eigenvalues sit near the cutoff (Figure 8, right). Subspace overlap at the corollary rank is 0.973 for $r_\star = 4$ and 0.734 for the truncated $r^\star = 10$

Table 13: Selected Gaussian cells. Oracle $r^\star$ is the corollary cutoff. Emp. P / $\widehat{P}$ are Monte Carlo mode ranks using the population projector and the sample covariance. Rel. err. is mean $|R_{\mathrm{MC}} - R_{\mathrm{exact}}|/R_{\mathrm{exact}}$ at the population projector.

$r_\star$	n_i	σ	γ	Oracle $r^\star$	Emp. P / $\widehat{P}$	Rel. err.
4	100	1	0	4	4 / 4	0.038
8	100	1	0	8	8 / 7	0.034
16	100	1	0	10	10 / 7	0.031
16	500	0.25	0	16	16 / 14	0.020
4	100	1	0.5	12	12 / 12	0.026

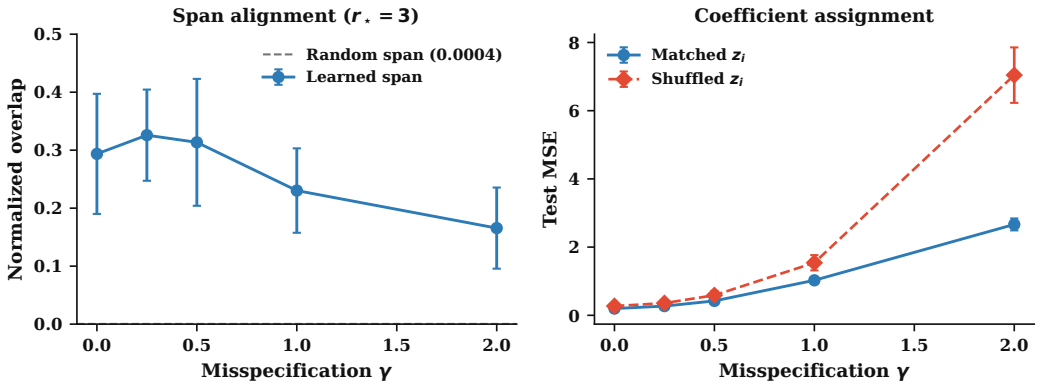

Figure 9: Planted-span teacher. Left: normalized projector overlap versus misspecification γ (mean $\pm$ one standard deviation over six populations). The dashed line is chance overlap 0.0004. Right: test MSE after from-zero coefficient estimation, matched versus shuffled coefficients.

cell at $r_\star = 16$, both at $n_i = 100$, $\sigma = 1$. These numbers validate the fixed-family formulas and the rank boundary; they do not identify a learned neural basis with that population span.

At $r_\star \le 4$ the corollary rank equals the planted dimension on every (n_i, σ) cell. At $r_\star = 16$ it does not: only $(n_i, \sigma) = (500, 0.25)$ retains all sixteen directions (Table 13). Orthogonal leftover variance $\gamma = 0.5$ inflates the selected rank from 4 to 12 in the misspecified setting, which is the expected leftover-population effect rather than a successful recovery of $r_\star$.

D.5.3 Neural teacher with a planted span

The planted-span teacher is $f_i^\star(x) = f_0^\star(x) + G^\star(x)z_i^\star + \gamma r_i(x)$ with $r_\star = 3$, where r_i is the client residual after least-squares removal of $G^\star$ on that client's sample. Twenty clients and six independent populations are used. Inputs are 16-dimensional; the shared student is an MLP encoder of width 32 with a scalar base and a rank-3 basis. Each of 80 rounds samples eight clients, resets the coefficient to zero, takes twelve local coefficient steps, then one shared step, and averages only the encoder, base, and basis. Training uses 256 labeled points per client; the test set uses 400, the step size is 0.05, and the observation noise is 0.05. Overlap $r^{-1}\,\mathrm{tr}(\widehat{P}_r P_{r_\star})$ is evaluated on pooled test inputs; a random three-dimensional subspace in the same sample has overlap 0.0004.

The overlap is above chance and decreases under severe misspecification (Figure 9, left). Shuffling coefficients increases test error at every γ (Table 14), so the estimated coefficient is client specific. The exact-specification overlap is not close to one. Joint learning of the encoder, base, and basis leaves a representation gauge, so a useful predictive family need not recover the teacher basis. This experiment shows partial alignment and client-specific coefficients, not recovery of the population span. Because the feature map and basis gauge are jointly learned, the experiment evaluates predictive family alignment rather than identified basis recovery.

Table 14: Planted-span teacher. Mean ± standard deviation over six populations. Test MSE is after from-zero coefficient estimation.

γ	Subspace overlap	Test MSE	Shuffled MSE
0	0.29 ± 0.10	0.198 ± 0.070	0.272 ± 0.090
0.25	0.33 ± 0.08	0.270 ± 0.054	0.357 ± 0.074
0.5	0.31 ± 0.11	0.422 ± 0.047	0.588 ± 0.129
1	0.23 ± 0.07	1.027 ± 0.054	1.541 ± 0.224
2	0.17 ± 0.07	2.665 ± 0.175	7.043 ± 0.813

D.5.4 Capacity versus coefficient fitting

The pFL-Bench FEMNIST $r \times T$ grid in Section D.1.2 is the neural analogue of the Gaussian rank boundary: identity rank r is representational capacity and T is coefficient fitting. That experiment reports participating and held-out accuracy and adaptation GFLOPs. It does not test correspondence between finite-step coefficient estimates and oracle shrinkage weights; finite-step optimization is not expected to reproduce those weights in general.

E Extended Discussion and Limitations

Relationship to neighboring parameterizations. FedRep assumes a shared input representation with separate client heads. FedAffine instead constrains client predictors to a shared affine family whose basis outputs are mixed by a short coefficient vector re-estimated against the current shared map. Conversely, pFedHN already maps compact client embeddings through a shared hypernetwork. Our distinction is the explicit affine function-family restriction and its fixed-family approximation–estimation analysis, not the mere presence of a compact client variable. These families can overlap for suitable linear generators or output adaptors.

Low dimension and memorization. A finite client set can be indexed by one real number. Identity dimension is therefore the number of locally estimated coordinates, not a complete complexity measure. Smoothness, norm control, cross-client coverage, and the complexity of the shared map remain relevant. The bound in Section B.5 controls only the fixed-map covering cost contributed by the identity.

Anisotropic estimation error and span recovery. The covariance calculation is favorable under isotropic noise, which shifts the spectrum without rotating it. Deep-model estimation error can instead be anisotropic and aligned across clients, confounding empirical cross-client structure with population variation. We do not estimate this effect using bootstrap replicas or repeated neural training samples. The planted teacher measures alignment with a known span across independently generated populations; it does not establish persistence of a learned span under repeated sampling.

Deployment supervision. Our implementations estimate client coefficients by minimizing labeled local loss. Reconstructive FEMNIST therefore requires labeled adaptation data for both participating and held-out clients. Office-Caltech retains local batch normalization and its identity, while GDP-HMM retains the selected site identity. Without labels, $z_i = 0$ provides only a default shared prediction; client-specific inference would require an additional self-supervised, generative, or amortized mechanism.

Security and semantics. Withholding the coefficient does not provide differential privacy or security against model inversion. Learned coordinates are predictive latent variables and should not be interpreted causally as writer style, scanner physics, or planning policy without external validation.

Medical interpretation. GDP-HMM clients are hospitals, but a learned site coefficient is not a validated encoding of institutional policy. Improved own-site Metric 1 does not establish clinical safety, calibration, plan deliverability, or deployment readiness. Patient-level evaluation, dosimetric quality assurance, and human review remain necessary.