CURVATURE TRANSFER: WHEN DOES COLLABORATION HELP PERSONALIZED LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized Federated Learning (PFL) improves client-specific models by transferring information across heterogeneous clients, yet existing methods provide limited insight into what recipient-specific information is actually needed for transfer and when collaboration adds value beyond local learning. We characterize this information through local transfer statistics that determine how population information should be adapted to a recipient. For least-squares heads over a shared representation, the characterization is exact and reduces to a curvature matrix and a feature–objective cross-moment, without requiring the full client distribution or client-to-population density ratios. This characterization shows that collaboration is useful when other clients improve estimation of the recipient's transfer statistics through similarity or exploitable structure. We develop pooled and structured estimators for these regimes and provide finite-time guarantees that characterize when collaboration improves personalized transfer. Experiments recover the predicted collaboration regimes and examine whether the transfer principle extends to nonlinear personalized heads. Under scarce homogeneous CIFAR features, pooled estimation reduces error by 10–19% relative to the strongest non-oracle comparator, while structured estimation reduces error by 4–21% under structured CIFAR-10/100 heterogeneity. The results also identify settings in which strong local estimation remains preferable, particularly under higher heterogeneity and on FEMNIST features. Overall, our results characterize when cross-client information improves personalized learning through better estimation of recipient-specific transfer information.

1 INTRODUCTION

Personalized federated learning (PFL) seeks to improve each client's model using information from other clients while allowing clients to retain distinct predictors (Smith et al., 2017; T Dinh et al., 2020; Li et al., 2021). A common approach separates a shared representation from client-specific prediction layers (Arivazhagan et al., 2019; Collins et al., 2021), a setting that is increasingly relevant as pretrained and foundation models provide representations that can be adapted to downstream tasks (Woisetschläger et al., 2024). Although the personalized component may be small, estimating it reliably can remain difficult when local data are limited. Information from other clients can reduce this estimation uncertainty, but distribution heterogeneity can make that information mismatched to the recipient's personalized learning problem.

This creates a fundamental tension between statistical gain and personalization. Pooling information across clients can reduce estimation variance, but under heterogeneous distributions the resulting population direction can be biased for a particular recipient. Existing approaches address heterogeneity through regularization, shared representations, model mixtures, or collaborator selection (T Dinh et al., 2020; Li et al., 2021; Deng et al., 2020; Collins et al., 2021; Even et al., 2022). For methods that transfer stochastic update directions, distribution mismatch can instead be corrected at the sample level using client-to-population density ratios (Zhang & Azizan, 2026). Such correction is general, but it raises a more basic question: *how much information about the recipient's distribution is actually required to adapt population information to the recipient?*

Our key observation is that personalized transfer does not necessarily require characterizing the full distributional mismatch between the population and the recipient. We make this precise for least-

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.

squares personalization over a shared, fixed representation, which isolates client-specific adaptation from representation learning. In this setting, the mean direction required by client i depends on its sampling distribution only through two finite-dimensional quantities: the curvature H_i of its local objective and a feature–objective cross-moment F_i. We refer to (H_i, F_i) as the client's *local transfer statistics*. These quantities compress the distributional information relevant to adapting a population direction to the recipient's learning problem, without requiring sample-level density-ratio correction.

The statistical benefit of collaboration then depends on how these client-specific transfer statistics are estimated. Estimating (H_i, F_i) locally is limited by the client's sample size. Information from other clients can reduce this estimation uncertainty when it is informative about the recipient's transfer statistics. Under homogeneous client distributions, pooling estimates across clients reduces estimation variance. Under heterogeneous distributions, collaboration can remain beneficial when variation in the transfer statistics has a lower-dimensional structure, allowing the shared component to be estimated across clients while only the client-specific variation is estimated locally. This perspective shifts the role of collaboration from transferring population updates to improving the estimation of the information required for personalized adaptation.

We develop local, pooled, and structured estimators of the transfer statistics and analyze their effect on personalized learning. Our finite-time analysis separates the error due to client–population mismatch, transfer-statistic estimation, and estimation of the population reference. Under homogeneous client distributions, pooling yields a $1/n$ improvement in the transfer-estimation term by reducing the uncertainty in the estimated transfer statistics. Under structured heterogeneity, collaboration remains beneficial when client-specific variation can be represented by a lower-dimensional component. Our experiments evaluate these regimes and show that the benefit of collaboration depends on the information other clients provide about the recipient's transfer statistics. We additionally evaluate nonlinear personalized heads to examine the transfer principle beyond the least-squares setting in which the characterization is exact. Our contributions are threefold.

- We characterize the information required for personalized transfer in PFL and identify two objective-specific local transfer statistics, (H_i, F_i), that adapt a population update to the recipient's learning problem without characterizing the full distributional mismatch.
- We develop local, pooled, and structured estimators of the transfer statistics and provide a theoretical analysis connecting their estimation error to personalized learning. The analysis identifies when cross-client information improves estimation, yielding a $1/n$ gain under homogeneous distributions and corresponding gains under structured heterogeneity.
- We empirically characterize the collaboration regimes predicted by the analysis, identifying settings in which transfer improves personalized learning and those in which local estimation remains preferable. We further evaluate nonlinear personalized heads to examine the transfer principle beyond least squares.

2 Problem formulation and transferability

Personalized learning under heterogeneous clients: We consider a personalized learning setting with n clients, where client i observes samples $s \sim \mu_i$ and seeks a model that minimizes its own prediction risk. Each client fits a linear head $x \in \mathbb{R}^d$ on a shared representation $\phi(s)$, such as the output of a frozen backbone or penultimate layer. Its least-squares objective is $L_i(x) = \frac{1}{2}\mathbb{E}_{\mu_i}[(\phi(s)^\top x - y_i(s))^2]$. Writing $A(s) := \phi(s)\phi(s)^\top$ and $b_i(s) := \phi(s)y_i(s)$, the stochastic gradient takes the affine form $g^i(x; s) = A(s)x - b_i(s)$. The map $A(s)$ is shared across clients, while clients may differ in their sampling distributions μ_i and targets y_i. Taking expectation under μ_i gives $\nabla L_i(x) = H_i x - q_i$. The personalized solution therefore satisfies

$$H_i x_i^\star = q_i, \qquad H_i := \mathbb{E}_{\mu_i}[A(s)], \quad q_i := \mathbb{E}_{\mu_i}[b_i(s)]. \tag{1}$$

We call H_i the *curvature* of client i. For least squares, it is exactly the Hessian of L_i and captures the feature geometry induced by the client's sampling distribution. The analysis below requires only the affine gradient form $A(s)x - b_i(s)$. Least squares gives $A(s) = \phi(s)\phi(s)^\top$, while the synthetic systems of Section 5 allow a general $A(s)$.

To separate heterogeneity in the sampling environment from heterogeneity in the prediction objective, we parameterize the targets using a second shared feature map $\psi(s) \in \mathbb{R}^m$, with $y_i(s) = \psi(s)^\top \theta_i^\star$. Setting $\Phi(s) := \phi(s)\psi(s)^\top$ and $F_i := \mathbb{E}_{\mu_i}[\Phi(s)]$ gives $b_i(s) = \Phi(s)\theta_i^\star$ and $q_i = F_i\theta_i^\star$.

The pair (H_i, F_i) describes how client i's sampling distribution acts on the shared feature maps, while $\theta_i^\star$ specifies its objective. Thus, the personalized solution $x_i^\star = H_i^{-1} F_i \theta_i^\star$ depends on both sampling and objective heterogeneity. This separation follows the linear framework of Zhang & Azizan (2026). Clients communicate through a server and perform one stochastic update per communication round. Notation is summarized in Appendix A.

What makes a population update transferable? Independent learning, $x_{t+1}^i = x_t^i - \alpha_t g_t^i(x_t^i)$ with $g_t^i(x) := A(s_t^i)x - b_i(s_t^i)$, already follows client i's mean field $H_i x_t^i - q_i$, but its stochastic gradients can have high variance. Collaboration can reduce this variance by constructing population reference quantities from information aggregated across clients. We maintain $\bar{x}_t$ and $\bar{\theta}_t$ as stochastic approximations to the solutions of $H_0 \bar{x} = \bar{q}$ and $F_0 \bar{\theta} = \bar{q}$, where $H_0 := \frac{1}{n}\sum_j H_j$, $F_0 := \frac{1}{n}\sum_j F_j$, and $\bar{q} := \frac{1}{n}\sum_j q_j$. Their updates average the stochastic residuals $A(s_t^i)\bar{x}_t - b_i(s_t^i)$ and $\Phi(s_t^i)\bar{\theta}_t - b_i(s_t^i)$ across clients, reducing the variance of the reference updates.

Client i uses these references through

$$x_{t+1}^i = x_t^i - \alpha_t \Big(g_t^i(x_t^i) + u_t^i - r_t^i \Big), \qquad r_t^i := A(s_t^i)\bar{x}_t - \Phi(s_t^i)\bar{\theta}_t. \tag{2}$$

Because g_t^i and r_t^i are evaluated on the same sample, subtracting r_t^i acts as a control variate that removes correlated noise. It also removes the conditional mean of r_t^i. To preserve client i's mean field, u_t^i must restore this mean, which gives the following characterization.

Proposition 1 (Transfer target)**.** *Let $\mathcal{F}_t$ be the sigma-field generated by all samples drawn before round t. Suppose u_t^i, $\bar{x}_t$, and $\bar{\theta}_t$ are $\mathcal{F}_t$-measurable and the required moments exist. The update* (2) *has conditional mean $H_i x_t^i - q_i$, the mean field of client i's own objective, if and only if*

$$\mathbb{E}[u_t^i \mid \mathcal{F}_t] = \mathbb{E}_{\mu_i}\big[A(s)\bar{x}_t - \Phi(s)\bar{\theta}_t\big] = H_i \bar{x}_t - F_i \bar{\theta}_t. \tag{3}$$

The proof is given in Appendix F.2. For fixed reference quantities, Proposition 1 shows that the sampling distribution μ_i enters the required transfer correction only through (H_i, F_i), which we call client i's *local transfer statistics*. This characterization concerns the transfer correction rather than the entire personalized problem. The solution $x_i^\star$ also depends on the objective parameter $\theta_i^\star$ through $q_i = F_i \theta_i^\star$. The transfer statistics themselves are expectations of shared maps and do not depend on $\theta_i^\star$. Client i can therefore estimate them directly from its own inputs,

$$\widehat{H}_t^i = \frac{1}{t}\sum_{\tau=0}^{t-1} A(s_\tau^i), \qquad \widehat{F}_t^i = \frac{1}{t}\sum_{\tau=0}^{t-1} \Phi(s_\tau^i), \tag{4}$$

without labels or knowledge of its sampling density. The transfer target can therefore be identified locally, although limited local data may prevent accurate estimation of the transfer statistics.

Relation to sample-level correction: When all clients share the same sampling distribution, their transfer statistics agree and the population average $u_t^i = \frac{1}{n}\sum_j r_t^j$ satisfies (3), with the usual $1/n$ variance reduction from averaging across clients (Chayti et al., 2021; Zhang & Azizan, 2026). Under heterogeneous sampling, the same average targets $H_0\bar{x}_t - F_0\bar{\theta}_t$ and is generally biased for client i.

A sample-level alternative is importance weighting. Let $\mu_0 := \frac{1}{n}\sum_j \mu_j$ denote the population mixture and $\rho_i(s) := \frac{d\mu_i}{d\mu_0}(s)$. Then $u_t^i = \frac{1}{n}\sum_j \rho_i(s_t^j) r_t^j$ has conditional expectation $H_i\bar{x}_t - F_i\bar{\theta}_t$ and therefore recovers the same client-specific target (Zhang & Azizan, 2026). This correction uses current-round samples and is not $\mathcal{F}_t$-measurable. It also requires sufficient information at the server to evaluate the density ratios and can increase variance under distribution mismatch.

Proposition 1 shows that distribution-level correction is sufficient but not intrinsic to the transfer problem, since the required mean depends on μ_i only through (H_i, F_i). This separates two questions. Transferability asks which client-specific quantities are needed to adapt a population direction, while collaborative estimation asks whether other clients can help estimate those quantities more accurately than local data alone. Section 3 addresses the second.

3 CURVATURE TRANSFER

Proposition 1 reduces transfer to estimating the local statistics (H_i, F_i). We consider three ways to estimate them: locally, by pooling information across clients, or through shared structure under heterogeneous environments. Each construction applies to H_i and, analogously, to F_i. To state their estimation errors, let V denote a uniform upper bound on the Frobenius variance of $A(s)$ and $\Phi(s)$ around their client-specific means; the formal regularity conditions are given in Assumption 1. The bounds summarized below are stated formally in Proposition 3 and proved in Appendix F.5.

Local transfer: Each client maintains the estimates in (4) and constructs the transferred direction:

$$u_t^i := \widehat{H}_t^i \, \bar{x}_t - \widehat{F}_t^i \, \bar{\theta}_t. \tag{5}$$

Nothing else in (2) changes. The server receives only the residuals used to maintain the reference quantities and broadcasts $(\bar{x}_t, \bar{\theta}_t)$. No raw samples leave the clients. The resulting direction satisfies (3) up to the estimation error of the running moments. Since each estimate averages t independent samples with Frobenius variance at most V, it is unbiased with mean-squared error at most V/t; see Proposition 3(a). Communication costs are reported in Table 5.

Federated curvature: Local estimation does not improve with the number of clients. When clients have similar transfer statistics, however, pooling can reduce estimation error. Under exact agreement, averaging the n independent local estimates reduces the Frobenius mean-squared error from V/t to $V/(nt)$. Away from exact agreement, pooling trades sampling variance against cross-client bias. We therefore use an entry-wise empirical-Bayes shrinkage estimator in the spirit of James–Stein (Efron & Morris, 1973),

$$\widehat{H}_t^{i,\mathrm{fed}} = \widehat{H}_t^0 + \Lambda_t^i \odot (\widehat{H}_t^i - \widehat{H}_t^0), \qquad \Lambda_t^i = \frac{\hat{\tau}_t^2}{\hat{\tau}_t^2 + \hat{\sigma}_{i,t}^2}. \tag{6}$$

Here $\widehat{H}_t^0 := \frac{1}{n}\sum_{j=1}^n \widehat{H}_t^j$ is the population estimate, $\hat{\sigma}_{i,t}^2$ estimates the sampling variance of client i's running mean, and $\hat{\tau}_t^2$ estimates the across-client variation in the corresponding transfer statistic. The same construction is applied separately to $\widehat{H}$ and $\widehat{F}$. When cross-client variation is small relative to sampling noise, the estimator shrinks toward the population mean; as heterogeneity grows, it approaches the local estimate.

To quantify this tradeoff, let

$$T^2 := \max\left\{ \frac{1}{n}\sum_{j=1}^n \|H_j - H_0\|_F^2 \,,\; \frac{1}{n}\sum_{j=1}^n \|F_j - F_0\|_F^2 \right\},$$

where $H_0 := \frac{1}{n}\sum_j H_j$ and $F_0 := \frac{1}{n}\sum_j F_j$, so that T^2 measures the across-client spread of the transfer statistics. Let $\bar{\varepsilon}(t)$ denote the corresponding agent-averaged estimation error. Proposition 3(b) characterizes the resulting tradeoff between pooling bias and local estimation variance as $\bar{\varepsilon}(t) \le \frac{V}{nt} + \min\left\{T^2, \frac{V}{t}\right\}$. Thus, pooling achieves $V/(nt)$ when $T^2 = 0$ and approaches the local V/t scale as heterogeneity increases. The bias–variance analysis and plug-in guarantee are in Appendix F.5, with implementation and communication details in Appendix C and Table 5.

Structured environment shift. Pooling gives a $1/n$ gain when clients share their transfer statistics. When they differ systematically, other clients can still help if the variation has exploitable low-dimensional structure. We describe the construction for H_i; the same argument applies to F_i. Suppose $H_i = \mathcal{A}(H_{\mathrm{sh}}, \nu_i)$, where $\mathcal{A}$ is a known reconstruction map, H_{sh} is a shared high-dimensional component, and $\nu_i \in \mathbb{R}^k$ is a client-specific environment parameter with $k \ll d^2$. The shared component is estimated across clients, while ν_i is estimated locally.

Suppose $\mathcal{A}$ is Lipschitz in its shared and local arguments with constants 1 and L_i, respectively. If the shared component is estimated with mean-squared error at most $V/(nt)$ and each coordinate of the local nuisance estimator has mean-squared error at most V_ν/t, then Proposition 3(c) gives

$$\mathbb{E}\left\|\widehat{H}_t^i - H_i\right\|_F^2 \le \frac{2V}{nt} + \frac{2L_i^2 k V_\nu}{t}.$$

Thus collaboration improves estimation of the high-dimensional shared component, while only the low-dimensional client-specific shift must be learned locally.

Under *label skew*,$H_i = \sum_c p_i(c) H_c$, where the class-conditional curvatures H_c are shared and estimated across clients, while client i estimates only its class frequencies p_i. Under *mean shift*, $H_i = (I + \epsilon(I + m_i m_i^\top))A_{\text{base}}$, where A_{base} is shared and the shift vector m_i is estimated locally. Appendix D gives these constructions and their additional regularity conditions.

Collaboration gate. Transfer need not improve finite-sample performance when the estimated correction provides little useful variance reduction. We therefore optionally scale the correction $c_t^i := u_t^i - r_t^i$ and use the update direction $g_t^i(x_t^i) + \beta_i c_t^i, \qquad \beta_i \in [0,1]$. For a scalar coefficient β_i, minimizing the conditional variance of this direction gives

$$\beta_i^\star = -\frac{\mathbb{E}\left[\left\langle g_t^i - \mathbb{E}[g_t^i \mid \mathcal{F}_t], c_t^i - \mathbb{E}[c_t^i \mid \mathcal{F}_t]\right\rangle \,\middle|\, \mathcal{F}_t\right]}{\mathbb{E}\left[\left\|c_t^i - \mathbb{E}[c_t^i \mid \mathcal{F}_t]\right\|^2 \,\middle|\, \mathcal{F}_t\right]}.$$

We estimate $\beta_i^\star$ online and clip it to $[0,1]$, so that $\beta_i = 1$ applies the full transfer correction, while $\beta_i = 0$ recovers independent learning. Thus, the gate suppresses collaboration when the correction provides little useful variance reduction. It applies to all three estimators above, but does not address error from estimating a high-dimensional local moment with too few samples. Appendix C gives the online estimator, an additional optional gate, and implementation details.

4 Analysis

The following conditions are used for the finite-time analysis, but not for Proposition 1.

Assumption 1. *Let $m = d$. (i) $\mathrm{sym}(H_i) \succeq \lambda_x I$ for all i and $\mathrm{sym}(F_0) \succeq \lambda_\theta I$, with $\lambda_x, \lambda_\theta > 0$ and $\lambda := \min\{\lambda_x, \lambda_\theta\}$. (ii) $\max\{\|H_i\|, \operatorname{ess\,sup}_s \|A(s) - H_i\|\} \le G_A$ and $\max\{\|F_i\|, \operatorname{ess\,sup}_s \|\Phi(s) - F_i\|\} \le G_\Phi$, while $\|x_i^\star\|, \|\bar{x}^\star\| \le G_x$ and $\|\theta_i^\star\|, \|\bar{\theta}^\star\| \le G_b$. (iii) $\mathbb{E}_{\mu_i}\|A(s) - H_i\|_F^2 \le V$ and $\mathbb{E}_{\mu_i}\|\Phi(s) - F_i\|_F^2 \le V$. Define $L := \max\{G_A, G_\Phi\}$, $\sigma := 2\max\{G_A G_x, G_\Phi G_b\}$, $\kappa := \sigma/\lambda$, and $G^2 := G_x^2 + G_b^2$, with $G_x, G_b \ge \frac{1}{2}$ so that $L \le \sigma$.*

Here G_A and G_Φ bound the transfer operators and their fluctuations, G_x and G_b bound the solution norms, and V bounds the transfer-statistic variance. Condition (iii) does not require independent matrix entries. For least squares, H_i is symmetric and (i) reduces to $H_i \succeq \lambda_x I$; the symmetric-part formulation also covers the general affine systems used in the synthetic experiments. The analysis assumes $m = d$; the $m < d$ implementation is discussed in Appendix F.9 and is not covered by Theorem 1.

Let $\mathcal{F}_t$ be the sigma-field generated by all samples drawn before round t. The iterates x_t^i, $\bar{x}_t$, $\bar{\theta}_t$ and the history-based transferred direction u_t^i, constructed from samples s_τ^i with $\tau < t$, are $\mathcal{F}_t$-measurable, while s_t^i is independent of $\mathcal{F}_t$. No sample splitting is used unless stated otherwise. For the analysis, the reference iterates are projected onto balls of radius $2G_x$ and $2G_b$ after each averaging step; the reported runs are unprojected and remain within these balls. Let $\bar{g}_i(x) := H_i x - q_i$ denote client i's mean field and $e_t^i := u_t^i - (H_i \bar{x}_t - F_i \bar{\theta}_t)$ the error in estimating its transfer target.

Proposition 2 (Decomposition). *Conditional on $\mathcal{F}_t$, the update $\tilde{g}_t^i = g_t^i(x_t^i) + u_t^i - r_t^i$ has mean $\bar{g}_i(x_t^i) + e_t^i$ and variance at most $2G_A^2 \left\|x_t^i - \bar{x}_t\right\|^2 + 2G_\Phi^2 \left\|\theta_i^\star - \bar{\theta}_t\right\|^2$.*

Proposition 2 separates the two quantities that govern the analysis. The variance of the corrected update depends on the client's mismatch with the reference quantities, while e_t^i captures error in estimating the client-specific transfer target. Because the history-based correction is fixed before the current sample is drawn, it introduces no additional fresh-sample variance; its error enters instead through the $\mathcal{F}_t$-measurable term e_t^i. The proof is given in Appendix F.3. Proposition 2 reduces the effect of transfer estimation to e_t^i; the next result relates this error to estimation of the transfer statistics themselves.

Proposition 3 (Transfer-statistic estimation error). *For the transfer-statistic estimator used by client i, denoted generically by $(\widehat{H}_t^i, \widehat{F}_t^i)$, let $\varepsilon_i(t) := \max\left\{\mathbb{E}\left\|\widehat{H}_t^i - H_i\right\|_F^2, \mathbb{E}\left\|\widehat{F}_t^i - F_i\right\|_F^2\right\}$,*

and, for agent-averaged error, $\bar{\varepsilon}(t) := \max\left\{\frac{1}{n}\sum_i \mathbb{E}\left\|\widehat{H}_t^i - H_i\right\|_F^2, \frac{1}{n}\sum_i \mathbb{E}\left\|\widehat{F}_t^i - F_i\right\|_F^2\right\}$. *Then* $\mathbb{E}\left\|e_t^i\right\|^2 \le 8G^2\varepsilon_i(t)$, $\frac{1}{n}\sum_i \mathbb{E}\left\|e_t^i\right\|^2 \le 8G^2\bar{\varepsilon}(t)$, *without requiring independence between the moment estimates and the reference iterates. The estimators of Section 3 satisfy the following bounds.*

(a) Local transfer: *Running local means satisfy* $\varepsilon_i(t) \le \frac{V}{t}$.

(b) Federated curvature: *Let* T^2 *be the larger of the across-client spreads* $\frac{1}{n}\sum_j \|H_j - H_0\|_F^2$ *and* $\frac{1}{n}\sum_j \|F_j - F_0\|_F^2$. *Under common entry-wise sampling variance across clients, the oracle empirical-Bayes estimator of Section 3 satisfies* $\bar{\varepsilon}(t) \le \frac{V}{nt} + \min\left\{T^2, \frac{V}{t}\right\}$. *When the transfer statistics agree, population pooling satisfies* $\varepsilon_i(t) \le V/(nt)$ *for every client.*

(c) Structured environment shift: *Suppose* H_i *and* F_i *admit the structured representations described in Section 3, with reconstruction maps Lipschitz in their shared and local arguments with constants* 1 *and* L_i*, respectively. If the shared components are estimated with mean-squared error at most* $V/(nt)$ *and the* k*-dimensional local nuisance estimate has per-coordinate mean-squared error at most* V_ν/t*, then* $\varepsilon_i(t) \le \frac{2V}{nt} + \frac{2L_i^2 k V_\nu}{t}$.

The proof is given in Appendix F.3, including the exact empirical-Bayes bias–variance risk and the derivation of the pooled bound. Proposition 4 separately gives an order-level guarantee for plug-in empirical-Bayes weights estimated from an independent split.

Write $D_t^i := \mathbb{E}\left\|x_t^i - x_i^\star\right\|^2$ for the optimization error. To quantify client-to-reference mismatch, define

$$\Delta_i := \frac{G_A^2\left\|x_i^\star - \bar{x}^\star\right\|^2 + G_\Phi^2\left\|\theta_i^\star - \bar{\theta}^\star\right\|^2}{\sigma^2}.$$

We normalize the transfer-statistic estimation error as

$$\zeta_i(t) := \frac{G^2 t\,\varepsilon_i(t)}{\sigma^2}, \qquad \zeta_i^{\max}(t) := \sup_{1\le\tau\le t}\zeta_i(\tau),$$

and, for agent-averaged bounds,

$$\zeta_{\mathrm{avg}}(t) := \frac{G^2 t\,\bar{\varepsilon}(t)}{\sigma^2}, \qquad \zeta_{\mathrm{avg}}^{\max}(t) := \sup_{1\le\tau\le t}\zeta_{\mathrm{avg}}(\tau).$$

The pointwise quantities measure transfer-statistic estimation error at round t, while the maxima account for errors accumulated over previous rounds in the finite-time recursion.

Theorem 1 (Rate of CC-PCL)**.** *Under Assumption 1, let the projected reference model be maintained by federated averaging and use step size* $\alpha_t = c/(\lambda(t+t_0))$*, where* $c \ge 4$ *is a fixed numerical constant and* $t_0 \ge 12c\sigma^2/\lambda^2$*. Then every client satisfies*

$$D_t^i = \tilde{O}\left(\frac{\kappa^2}{t}\left[\underbrace{\Delta_i}_{\text{reference mismatch}} + \underbrace{\zeta_i^{\max}(t)}_{\text{transfer estimation}}\right]\right) + \underbrace{\tilde{O}\left(\frac{\kappa^4}{nt^2}\right)}_{\text{reference estimation}}. \tag{7}$$

Here $\tilde{O}$ *hides numerical constants and logarithmic factors. In addition, the initialization transient is* $\left(2t_0/(t+t_0)\right)^{c/2} D_{t_0}^i$.

The bound separates two leading effects. The term Δ_i measures the mismatch between client i and the population reference quantities and therefore controls the residual variance of the population control variate. The term $\zeta_i^{\max}(t)$ captures the accumulated effect of estimating the client-specific transfer statistics. Local running means have $\varepsilon_i(t) = O(1/t)$ and hence $\zeta_i^{\max}(t) = O(1)$, whereas an estimator satisfying $\varepsilon_i(t) = O(1/(nt))$ uniformly over rounds gives $\zeta_i^{\max}(t) = O(1/n)$. The reference-model error contributes the lower-order $\tilde{O}(\kappa^4/(nt^2))$ term. Thus, local transfer retains the $1/t$ order of independent learning, with a different constant. Collaboration can improve the transfer-estimation contribution when pooling or shared structure estimates the transfer statistics

more accurately than local data alone; the remaining benefit is governed by the client-to-reference mismatch Δ_i. Appendix F.9 relates Δ_i to the total-variation heterogeneity measures of Zhang & Azizan (2026) and discusses the corresponding comparison with importance weighting.

For the pooled estimator, Proposition 3(b) controls the agent-averaged transfer-statistic error $\bar{\varepsilon}(t)$. Averaging the per-client recursion before unrolling it gives

$$D_t^{\mathrm{avg}} = \tilde{O}\left(\frac{\kappa^2}{t}\left[\Delta_{\mathrm{avg}} + \zeta_{\mathrm{avg}}^{\max}(t)\right]\right) + \tilde{O}\left(\frac{\kappa^4}{nt^2}\right), \tag{8}$$

where $D_t^{\mathrm{avg}} := \frac{1}{n}\sum_i D_t^i$ and $\Delta_{\mathrm{avg}} := \frac{1}{n}\sum_i \Delta_i$. Thus, the general empirical-Bayes guarantee is agent averaged. When the transfer statistics agree, Proposition 3(b) gives $\varepsilon_i(t) = O(1/(nt))$ for every client, yielding the corresponding per-client rate.

The theorem covers local transfer and pooled transfer with fixed oracle empirical-Bayes weights constructed from past samples; no sample splitting is required for the moment estimates themselves. Proposition 4 separately controls plug-in weights estimated from an independent split. Same-sample plug-in weights, current-sample moment updates, the rectangular objective reference, and the online collaboration gate are discussed in Appendix F.9.

Corollary 1. *Suppressing the common factor G^2V/σ^2, the estimators of Proposition 3 yield the following transfer-estimation regimes.*

(a) Local transfer: *Local running means give $\zeta_i(t) = O(1), \qquad \zeta_i^{\max}(t) = O(1)$, and therefore retain the $1/t$ order of independent learning, with a different constant.*

(b) Federated curvature: *For the oracle pooled estimator, $\zeta_{\mathrm{avg}}(t) \le \frac{1}{n} + \min\left\{1, \frac{T^2 t}{V}\right\}$, where $\zeta_{\mathrm{avg}}(t) := \frac{G^2 t\,\bar{\varepsilon}(t)}{\sigma^2}$. Thus, when the transfer statistics agree, pooling gives an $O(1/n)$ transfer-estimation term. Under fixed nonzero heterogeneity, the bound guarantees an improvement while T^2t/V is small; once t reaches order V/T^2, the heterogeneity contribution reaches the local $O(1)$ scale. Defining $\zeta_{\mathrm{avg}}^{\max}(t) := \sup_{\tau \le t}\zeta_{\mathrm{avg}}(\tau)$, the monotonicity of the displayed upper bound gives the same order for $\zeta_{\mathrm{avg}}^{\max}(t)$. When the transfer statistics agree, Proposition 3(b) gives $\zeta_i(t) \le \frac{1}{n}, \qquad \zeta_i^{\max}(t) \le \frac{1}{n}$ for every client.*

(c) Structured environment shift. *Structured estimators give $\zeta_i(t) \le 2\left[\frac{1}{n} + \delta_i^{\mathrm{curv}}\right], \qquad \delta_i^{\mathrm{curv}} := \frac{L_i^2 k V_\nu}{V}$. The bound is uniform in t, so $\zeta_i^{\max}(t) \le 2\left[\frac{1}{n} + \delta_i^{\mathrm{curv}}\right]$. Consequently,* (7) *has leading-order dependence $\tilde{O}\left(\frac{\kappa^2}{t}\left[\Delta_i + \frac{1}{n} + \delta_i^{\mathrm{curv}}\right]\right)$.*

Corollary 1 characterizes when collaboration improves transfer estimation. Local transfer identifies the client-specific correction but provides no n-dependent gain. Pooling provides such a gain when transfer statistics agree or their heterogeneity is small relative to the sample size. Structured estimation can retain an n-dependent shared component under heterogeneity by separating shared variation from a low-dimensional client-specific shift. The overall rate also depends on the reference mismatch Δ_i. Moreover, the Frobenius variance V may scale with dimension, making local estimation costly in high-dimensional, data-scarce regimes, as examined in Section 5.2.

5 Experiments

The characterization leads to four empirical questions. *(Q1)* When a client can accurately estimate its own transfer statistics, does transfer provide information beyond a strong local estimator? *(Q2)* When local estimation is sample-starved and clients have similar transfer statistics, does pooling improve estimation? *(Q3)* Under structured heterogeneity, can structured estimation retain the benefit of collaboration after unstructured pooling stops helping? *(Q4)* What happens when neither similarity nor the chosen structure is sufficiently informative?

Experimental setup: We first use the controlled linear system of Zhang & Azizan (2026) with $n = 20$ clients and $d = 5$, varying sampling-distribution and objective heterogeneity through δ_{env} and δ_{obj}, respectively. Real-feature experiments use frozen ImageNet ResNet-18 representations reduced by PCA to $d = 64$ for CIFAR-10/100 and frozen LeNet features with $d = 84$ for FEMNIST. Full data-generation and optimization details are provided in Appendix H.

Table 1: Controlled linear-system benchmark ($n = 20$, $d = 5$; 10 runs). Polyak-averaged $\mathrm{MSE}_0 \times 10^{-4}$.

Step	$(\delta_{\mathrm{env}}, \delta_{\mathrm{obj}})$	Local	FedAvg	Ditto	pFedMe	IFCA	All-1	IW-none	IW-exact	CC-local	CC-PCL	CC-struct
Const.	$(0, 0)$	4.27	0.22	0.24	35.90	0.22	0.22	0.22	0.22	4.34	0.60	**0.17**
	$(.05, .05)$	4.27	22.83	5.73	63.28	22.80	15.51	0.50	0.25	4.32	1.02	**0.24**
	$(.2, .2)$	4.19	383.82	33.12	98.73	383.67	14.50	43.88	0.68	4.32	4.66	**0.84**
	$(.5, .5)$	3.79	2457.91	176.16	242.78	2005.16	6.77	430.14	2.44	4.81	4.88	**2.29**
Decay	$(0, 0)$	2.66	0.17	0.19	2.96	0.17	0.17	0.17	0.17	2.84	0.38	**0.09**
	$(.05, .05)$	2.66	22.79	4.20	7.13	22.79	15.55	0.45	0.18	2.84	0.77	**0.14**
	$(.2, .2)$	2.60	383.80	31.78	34.60	383.59	13.59	43.87	0.33	2.82	2.95	0.50
	$(.5, .5)$	2.38	2459.61	174.79	159.07	1978.11	3.70	431.93	1.17	2.91	2.92	1.23

Table 2: Frozen CIFAR representations: excess test MSE $\times 10^{-3}$ over 5 seeds. Q2 has homogeneous sampling and Q3 structured label skew.

Q	Data	Setting	Local	FedAvg	FT	APFL	Ditto	pFedMe	CC-local	CC-PCL	CC-class	IW-exact
Q2	C10	$n_{\mathrm{tr}} = 64$	60.7	44.9	46.6	55.4	43.6	70.1	78.3	**39.1**	48.5	32.0
	C10	$n_{\mathrm{tr}} = 128$	51.9	41.3	42.8	54.7	37.8	64.8	76.2	**31.6**	38.6	27.1
	C10	$n_{\mathrm{tr}} = 256$	40.9	39.3	41.6	53.0	32.7	55.6	72.0	**28.6**	32.3	24.9
	C10	$n_{\mathrm{tr}} = 512$	34.7	38.5	39.2	50.1	29.8	47.0	62.5	**26.1**	27.9	23.9
	C100	$n_{\mathrm{tr}} = 128$	39.5	30.9	33.2	39.7	30.5	54.8	55.3	**24.7**	34.6	21.4
Q3	C10	Dir(1)	63.7	136.9	76.7	66.4	84.0	152.0	81.7	71.4	**51.8**	44.2
	C10	Dir(.3)	56.1	212.2	77.7	63.5	106.7	214.4	76.4	75.3	**53.6**	48.6
	C10	$\delta_{\mathrm{env}} = .5$	55.7	128.8	62.3	53.9	68.4	142.9	72.0	60.7	**42.7**	35.8
	C100	Dir(1)	46.2	52.2	46.2	45.7	44.9	75.6	58.3	40.8	**39.1**	27.4
	C100	Dir(.3)	50.0	81.8	57.4	51.1	61.6	105.0	59.1	54.5	**43.7**	32.9
	C100	$\delta_{\mathrm{env}} = .5$	47.0	57.0	50.0	47.7	48.6	81.0	59.1	43.9	**39.6**	28.4

Baselines: We compare with strong local estimators (LOCAL-LS/RIDGE and INDEP-AVG), FedAvg (McMahan et al., 2017), FedAvg-FT (Collins et al., 2022), APFL (Deng et al., 2020), Ditto (Li et al., 2021), and pFedMe (T Dinh et al., 2020). The synthetic benchmark additionally includes IFCA (Ghosh et al., 2020) and All-for-one (Even et al., 2022). For sample-level transfer, IW-EXACT and IW-NONE use the method of Zhang & Azizan (2026) with and without exact client-to-population density ratios. Our variants estimate the transfer statistics locally (CC-LOCAL), by empirical-Bayes pooling (CC-PCL), or using the structured models of Section 3 (CC-STRUCT/CC-CLASS). Hyperparameter selection, evaluation protocols, and additional metrics are reported in Appendix H.

5.1 Q1: WHEN DOES TRANSFER ADD INFORMATION BEYOND LOCAL ESTIMATION?

Table 1 isolates the information supplied by transfer. CC-LOCAL closely tracks the client's local solution across both step schedules, consistent with Proposition 1. When the transfer statistics are estimated only from the recipient's samples, the correction therefore provides little information beyond the local problem itself. Collaboration becomes useful when it improves their estimation. Pooling is effective at low heterogeneity but loses its advantage as the client distributions and objectives separate, whereas CC-STRUCT remains effective by sharing the common component while estimating the client-specific shift locally. IW-EXACT remains a strong oracle comparator supplied with exact density ratios, while IW-NONE deteriorates as the sampling distributions separate.

5.2 Q2–Q3: COLLABORATION WITH SCARCE REAL-FEATURE DATA

We next test the estimation regimes on frozen CIFAR representations with 50 clients. The metric is excess test MSE relative to each client's optimal linear predictor.

Q2: homogeneous clients. With only $O(d)$ local samples, estimating full matrix-valued transfer statistics locally is difficult, and CC-LOCAL performs poorly (Table 2). Pooling reverses this behavior. Across the homogeneous CIFAR settings, CC-PCL reduces excess MSE by 10.4–19.1% relative to the strongest conventional PFL baseline. This supports the predicted regime in which collaboration helps by reducing uncertainty in the recipient's transfer statistics.

Q3: structured heterogeneity. Under label skew, unstructured pooling loses its advantage, whereas CC-CLASS shares class-conditional transfer statistics while retaining client-specific class frequen-

Table 3: Boundary regimes under extreme CIFAR label skew and FEMNIST. Excess test MSE $\times 10^{-3}$.

Data	Setting	Indep.	Local	FedAvg	FT	APFL	CC-local	CC-PCL	CC-class	IW-exact
CIFAR-10	$\delta_{\mathrm{env}} = .9$	38.6	36.2	250.2	50.0	39.3	50.2	52.1	**36.1**	33.1
CIFAR-10	Dir(.1)	47.4	**39.0**	258.4	61.3	45.9	55.9	57.6	42.1	38.8
CIFAR-100	$\delta_{\mathrm{env}} = .9$	57.6	51.7	103.9	67.9	57.7	64.4	66.1	**48.2**	39.0
CIFAR-100	Dir(.1)	51.3	**44.7**	123.4	60.6	49.9	59.0	61.8	45.7	36.7
FEMNIST	$\delta = 0,\ n_{\mathrm{tr}} = 128$	15.8	**11.6**	15.8	15.2	16.0	17.0	14.9	16.2	14.9
FEMNIST	$\delta = .5,\ n_{\mathrm{tr}} = 128$	17.0	**10.8**	21.9	17.7	18.1	17.9	17.3	17.2	15.9
FEMNIST	Dir(.3), $n_{\mathrm{tr}} = 128$	17.7	**10.5**	27.4	19.6	19.1	18.9	18.9	18.3	17.2
FEMNIST	$\delta = .5,\ n_{\mathrm{tr}} = 512$	11.4	**4.1**	18.3	12.2	12.0	12.7	12.7	12.3	12.3
FEMNIST	100 real writers	35.2	**31.2**	42.9	36.2	35.6	36.4	36.0	37.4	37.2

cies. Across the six skewed CIFAR-10/100 settings, it improves on the strongest conventional deployable baseline by 4.5–20.8%. The oracle IW-EXACT remains lower in all six cases, indicating that the structured estimator recovers part of the benefit available with exact density ratios.

5.3 Q4: WHERE DOES COLLABORATION STOP HELPING?

Under extreme CIFAR skew, structured transfer reaches its boundary. At $\delta_{\mathrm{env}} = 0.9$, CC-CLASS is essentially tied with the local solution on CIFAR-10 and improves CIFAR-100 by 6.8%, but loses to local estimation under Dirichlet(0.1) (Table 3). FEMNIST gives a stronger negative result: no tested transfer method improves on LOCAL-RIDGE in any displayed setting, including 100 real writers. Thus local sample scarcity alone is insufficient for collaboration to help; other clients must also provide informative structure for estimating the recipient's transfer statistics.

Beyond linear heads. The characterization is exact for least-squares heads. We additionally evaluate a first-order nonlinear extension in Appendix E.4. These experiments provide a scope check rather than validation of the exact theory and show mixed gains across CIFAR and FEMNIST.

6 DISCUSSION AND CONCLUSION

This work characterizes the information required to transfer population updates to a personalized client. For least-squares heads over a fixed shared representation, this information reduces exactly to two local transfer statistics. Estimating these statistics from the recipient's data identifies the required correction but provides no n-dependent statistical gain. Collaboration becomes useful when other clients improve their estimation, either because the transfer statistics are sufficiently similar or because their variation has exploitable structure. The finite-time analysis separates this estimation benefit from reference mismatch and population-reference error, while the experiments identify both regimes where collaboration helps and where strong local estimation remains preferable.

The exact characterization applies to least-squares personalized heads over a fixed representation, where the affine gradient structure yields an exact transfer identity and enables the finite-time analysis. For nonlinear heads, the same principle leads to a local first-order transfer approximation, and our experiments examine its behavior beyond the exact least-squares setting. Extending the characterization to jointly learned representations and more general nonlinear models is an important direction for future work. Additional technical scope of the analysis and implementation is detailed in Appendix F.9.

The experiments also leave open why CIFAR and FEMNIST fall into different collaboration regimes. One plausible explanation is that the FEMNIST representation supports substantially stronger local estimation, leaving less estimation error for collaboration to reduce. This is consistent with the strong local closed-form results on FEMNIST and with our finding that changing the SGD step size does not remove the gap. More generally, identifying which properties of the representation, data distribution, and transfer statistics predict when collaboration will help remains an important direction for future work. Full curvature estimates also require $O(d^2)$ storage per client, motivating structured or factored approximations at larger scale.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and critique of the manuscript's organization. We have not used generative AI tools to generate research ideas, and the remaining disclosure categories are not applicable to this work. All AI-assisted text was reviewed and revised by the authors to ensure the accuracy of the technical content and claims. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Sections 2–4, and the linear proofs are given in Appendices F. Appendix F.9 states the scope of the theory, including all-samples local and oracle-pooled moment estimation, the independent-split requirement for the plug-in-weight guarantee, and the implementation extensions not covered by the theorems. Appendix C gives the pooled plug-in formulas, pseudocode, gates and communication accounting; Appendix D gives the structured estimators; Appendix E gives additional scaling, partial-participation, classification, nonlinear-scope and ablation experiments; Appendix G gives the complete supporting tables for the main comparisons; and Appendix H gives the remaining settings, hyperparameter grids and compute details. The synthetic system is seeded and deterministic; the code and runner scripts reproducing the reported numbers will be provided as supplementary material.

REFERENCES

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

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

El Mahdi Chayti, Sai Praneeth Karimireddy, Sebastian U. Stich, Nicolas Flammarion, and Martin Jaggi. Linear speedup in personalized collaborative learning. *arXiv preprint arXiv:2111.05968*, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning (ICML)*, 2021.

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

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

Canh T. Dinh, Nguyen H. Tran, Tuan Dung Nguyen, Wei Bao, Amir Rezaei Balef, Bing B. Zhou, and Albert Y. Zomaya. DONE: Distributed approximate Newton-type method for federated edge learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2648–2660, 2022.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 212–225. Curran Associates, Inc., 2022. doi: 10.52202/068431-0016.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory (COLT)*, 2015.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19586–19597. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e32cc80bf07915058ce90722ee17bb71-Paper.pdf.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning (ICML)*, 2021.

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.

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 (NeurIPS)*, 2021.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, 2015.

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

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making Newton-type methods applicable to federated learning. In *International Conference on Machine Learning (ICML)*, 2022.

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.

Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *International Conference on Machine Learning (ICML)*, 2014.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-IID data. *arXiv preprint arXiv:1910.07796*, 2019.

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

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

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

Herbert Woisetschläger, Alexander Erben, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. A survey on efficient federated learning methods for foundation model training. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8317–8325. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/919. URL https://doi.org/10.24963/ijcai.2024/919. Survey Track.

Chenyu Zhang and Navid Azizan. Personalized collaborative learning with affinity-based variance reduction. In *International Conference on Learning Representations (ICLR)*, 2026.

A Notation

Table 4: Notation.

$n,\ d,\ t$	number of agents, dimension, round index
$\mu_i,\ \mu_0$	sampling distribution of agent i; population mixture $\frac{1}{n}\sum_i \mu_i$
$A(s),\ \Phi(s)$	shared feature matrices; $b_i(s) = \Phi(s)\theta_i^\star$ the agent-specific target
$H_i = \mathbb{E}_{\mu_i} A,\ F_i = \mathbb{E}_{\mu_i}\Phi$	curvature and objective-feature moment of agent i; H_0, F_0 their population means
$x_i^\star,\ \theta_i^\star$	personalized solution and objective parameter of agent i
$\bar{x}_t,\ \bar{\theta}_t$	reference model (federated averages); limits $\bar{x}^\star, \bar{\theta}^\star$
$g_t^i(x) = A(s_t^i)x - b_i(s_t^i)$	stochastic residual of agent i; $\bar{g}_i(x) = H_i x - q_i$ its mean field
$r_t^i,\ u_t^i$	agent's own residual at the reference; transferred population direction
$\widehat{H}_t^i,\ \widehat{F}_t^i$	estimates of the transfer statistics (local, pooled or structured)
$\lambda,\ \sigma,\ \kappa = \sigma/\lambda,\ V$	strong monotonicity, noise scale, condition number, Frobenius variance (total over entries; V_ν is per-coordinate)
$\Delta_i,\ \zeta_i(t)$	reference mismatch; transfer-statistic estimation term of Theorem 1
$D_t^i = \mathbb{E}\left\|x_t^i - x_i^\star\right\|^2$	mean-squared error of agent i

B Related Work

Personalized federated learning. PFL addresses client heterogeneity through regularization toward a shared model (T Dinh et al., 2020; Li et al., 2021; Hanzely & Richtárik, 2020), client clustering (Ghosh et al., 2020; Sattler et al., 2020), shared representations (Arivazhagan et al., 2019; Collins et al., 2021; Liang et al., 2020), mixtures of global and local models (Deng et al., 2020), meta-learning (Fallah et al., 2020), and multi-task formulations (Smith et al., 2017; Marfoq et al., 2021). These approaches primarily specify how models, representations, or optimization objectives are shared across clients. Our focus is different. We ask what information about a client's distribution is required to adapt transferred population information to that client's learning objective. In the least-squares setting, this information is captured exactly by the transfer statistics (H_i, F_i). Our pooled estimator consequently shrinks these statistics toward population estimates rather than shrinking the personalized model itself.

Personalized collaboration and transfer. A related line of work studies when information from other clients can improve personalized learning. Chayti et al. (2021) and Even et al. (2022) exploit relationships between clients to select or weight useful collaborators, while Zhang & Azizan (2026) uses bias and importance correction to adapt collaboration under heterogeneous environments and objectives. Our characterization is complementary. Rather than beginning with a particular weighting or collaborator-selection rule, we derive the client-specific mean that a transferred direction must reproduce. For personalized least squares, the required distributional information reduces to the finite-dimensional pair (H_i, F_i).

Distribution shift and importance weighting. Importance weighting corrects expectations under distribution shift using density ratios between source and target distributions (Shimodaira, 2000; Huang et al., 2007; Sugiyama et al., 2012). This provides a general sample-level correction but can require estimating a high-dimensional density ratio. Our result addresses a more specific question of what distributional information is required to reproduce the mean update for a given personalized learning problem. For least-squares heads, the answer is not the full density ratio but two objective-dependent second moments. This does not replace importance weighting for general distribution shift. It identifies a setting in which the information required for personalized transfer has a lower-dimensional, objective-dependent structure.

Control variates and second-order information. Our update uses a population reference as a control variate, relating it to variance-reduced stochastic optimization (Johnson & Zhang, 2013; Frostig et al., 2015). In federated optimization, SCAFFOLD uses control variates to correct client drift toward a common global objective (Karimireddy et al., 2020). Here the objective is personalized, and the transfer statistics determine how a population reference must be adapted to the receiving client. Second-order information has also been used in federated learning to precondition optimization (Shamir et al., 2014; Safaryan et al., 2022; Dinh et al., 2022; Martens & Grosse, 2015) or to regularize local models through Fisher information (Shoham et al., 2019). In our setting, the local second moments serve neither role. They characterize the mean of the client-specific transfer correction and become statistical quantities whose estimation can itself benefit from collaboration.

C Implementation details

C.1 Empirical-Bayes pooling

Equation (6) is implemented entry-wise and separately for $\widehat{H}$ and $\widehat{F}$. For client i, we estimate the sampling variance of each entry of the running mean from its running first and second moments. For $\widehat{H}$, let $\widehat{A^{\odot 2}}_t^i$ denote the entry-wise second moment. We use

$$\hat{\sigma}_{i,t}^2 = \frac{\left[\widehat{A^{\odot 2}}_t^i - \widehat{H}_t^i \odot \widehat{H}_t^i\right]_+}{t},$$

with the analogous estimator for $\widehat{F}$.

The across-client variance of the underlying transfer statistic is estimated by removing the estimated sampling variance from the observed cross-client variation,

$$\hat{\tau}_t^2 = \left[\frac{1}{n}\sum_j (\widehat{H}_t^j - \widehat{H}_t^0)^{\odot 2} - \frac{1}{n}\sum_j \hat{\sigma}_{j,t}^2\right]_+.$$

The resulting plug-in pooling weight is

$$\hat{\Lambda}_t^i = \frac{\hat{\tau}_t^2}{\hat{\tau}_t^2 + \hat{\sigma}_{i,t}^2 + \epsilon},$$

where $\epsilon > 0$ provides numerical stability. Small estimated cross-client variation shrinks the client estimate toward the population statistic, whereas large variation relative to sampling noise favors the local estimate.

Proposition 3(b) analyzes the corresponding oracle empirical-Bayes weight under common entry-wise sampling variance. Proposition 4 gives an order-level guarantee when the plug-in weight is estimated from an independent split. The experiments use the practical same-sample estimator above, whose dependence is outside the scope of the theorem as discussed in Appendix F.9.

Algorithm 1 CC-PCL: federated implementation.

Input: step sizes $\{\alpha_t\}$, synchronization period R, gate coefficients $\{\beta_i\}$
Initialize $\{x_0^i\}_{i=1}^n$, references $(\bar{x}_0, \bar{\theta}_0)$, and transfer statistics $(\widehat{H}_0^{i,\mathrm{fed}}, \widehat{F}_0^{i,\mathrm{fed}})$
for $t = 0, 1, \ldots, T-1$ **do**
  Clients, in parallel:
  for each client i **do**
    Sample a mini-batch and compute A_t^i, b_t^i, and Φ_t^i
    Update the running local moments $(\widehat{H}_t^i, \widehat{F}_t^i)$
    Form

$$u_t^i = \widehat{H}_t^{i,\mathrm{fed}} \bar{x}_t - \widehat{F}_t^{i,\mathrm{fed}} \bar{\theta}_t$$

    Form

$$c_t^i = u_t^i - \left(A_t^i \bar{x}_t - \Phi_t^i \bar{\theta}_t\right)$$

    Update

$$x_{t+1}^i = x_t^i - \alpha_t \left(g_t^i(x_t^i) + \beta_i c_t^i\right)$$

    Compute

$$r_{x,t}^i = A_t^i \bar{x}_t - b_t^i, \qquad r_{\theta,t}^i = \Phi_t^i \bar{\theta}_t - b_t^i$$

    Send $(r_{x,t}^i, r_{\theta,t}^i)$ to the server, together with $(\widehat{H}_t^i, \widehat{F}_t^i)$ on synchronization rounds
  end for
  Server:
  Update

$$\bar{x}_{t+1} = \bar{x}_t - \frac{\alpha_t}{n} \sum_i r_{x,t}^i, \qquad \bar{\theta}_{t+1} = \bar{\theta}_t - \frac{\alpha_t^\theta}{n} \sum_i r_{\theta,t}^i$$

  if $t \bmod R = 0$ **then**
    Update the pooled transfer statistics using Eq. (6)
  end if
  Broadcast the references and any refreshed transfer statistics for round $t+1$
end for
Output: $\{x_T^i\}_{i=1}^n$

C.2 Federated implementation

Algorithm 1 summarizes the square $m = d$ setting analyzed by Theorem 1. Each client maintains local estimates of (H_i, F_i) and periodically communicates them to the server for pooling. Between synchronization rounds, clients continue to update their local statistics and personalized models. The population references are updated from aggregated reference residuals.

The real-feature experiments use $m < d$. In this case, the objective reference is maintained using the stochastic normal-equation residual described in Appendix F.9. This rectangular implementation is used experimentally but is not covered by Theorem 1.

C.3 Communication cost

Each round communicates the residual/reference quantities required to update the population references, contributing $4d$ floats per client in the square synthetic setting. Every R rounds, each client additionally uploads its $d \times d$ and $d \times m$ transfer-statistic estimates. The amortized communication of CC-PCLis therefore $4d + \frac{d^2+dm}{R}$ floats per client per round.

Table 5 evaluates this cost using the synthetic communication benchmark with $m = 2$. Under its communication model, IW-EXACT communicates $7d$ floats per client per round, whereas CC-

Table 5: Communication in floats per client per round for the synthetic benchmark ($m = 2$), including uploads and broadcasts. Parentheses give the ratio to IW-EXACT.

	IW	CC-LOCAL	CC-PCL, $R = 100$			CC-PCL, $R = d$		
d	$7d$	$4d$	residual	periodic	total	residual	periodic	total
20	140	80 (0.57)	80	4	84 (0.60)	80	22	102 (0.73)
64	448	256 (0.57)	256	42	298 (0.67)	256	66	322 (0.72)
128	896	512 (0.57)	512	166	678 (0.76)	512	130	642 (0.72)
256	1792	1024 (0.57)	1024	660	1684 (0.94)	1024	258	1282 (0.72)
512	3584	2048 (0.57)	2048	2632	4680 (1.31)	2048	514	2562 (0.71)
1024	7168	4096 (0.57)	4096	10506	14602 (2.04)	4096	1026	5122 (0.71)

LOCAL communicates $4d$. For CC-PCL, the additional full-moment communication depends on both R and d.

Table 5 makes the scaling tradeoff explicit. CC-LOCAL uses 43% fewer communicated floats than IW-EXACT under this benchmark. For CC-PCL, the advantage depends on R and d. With fixed $R = 100$, the quadratic moment upload eventually dominates the communication cost. Setting $R = d$ keeps the total at approximately 71–73% of the IW-EXACT communication over the dimensions shown. Full moment communication is therefore not intended as a scalable high-dimensional implementation. Structured, low-rank, or factored representations of the transfer statistics are natural alternatives at larger d. In the synthetic experiments, pooling begins after a 200-round warm-up and is disabled when fewer than four clients are available.

D STRUCTURED TRANSFER ESTIMATORS

Structured estimation is useful when client-specific transfer statistics vary through a lower-dimensional parametrization. For example, write

$$H_i = \mathcal{A}(H_{\mathrm{sh}}, \nu_i), \qquad \nu_i \in \mathbb{R}^k, \quad k \ll d^2,$$

with an analogous representation for F_i when applicable. The shared component is estimated across clients, while the lower-dimensional parameter ν_i is estimated locally. This yields the error decomposition underlying Section 3: the shared component benefits from cross-client estimation at rate $O(V/(nt))$, while the client-specific component contributes an $O(k/t)$ term whose magnitude depends on the client-specific shift. We consider two concrete structures.

Label skew. Suppose clients differ through their class frequencies $p_i(c)$ while the class-conditional distributions are shared. Then

$$H_i = \sum_{c=1}^{C} p_i(c) H_c,$$

where the class-conditional statistics $H_{c\,c=1}^{C}$ are shared across clients and estimated using samples from all clients. Client i only needs to estimate its C class frequencies $p_i = (p_i(1), \ldots, p_i(C))$. Thus the client-specific variation is parameterized by a C-dimensional quantity rather than a full $d \times d$ matrix. The same construction is applied to F_i when its class-conditional decomposition is available. No client-to-population density ratio is required.

Mean shift. For the controlled mean-shift model,

$$H_i = \left(I + \epsilon(I + m_i m_i^\top)\right) A_{\mathrm{base}},$$

where A_{base} is shared and $m_i \in \mathbb{R}^d$ is client-specific. Client i estimates m_i locally and forms

$$\widehat{A}_{\mathrm{base},i} = \left(I + \epsilon(I + \hat{m}_i \hat{m}_i^\top)\right)^{-1} \widehat{H}_t^i.$$

Table 6: Speedup over independent learning on the (n, δ) grid with a constant step ($T = 2000$, 3 runs).

δ	method	$n = 2$	$n = 5$	$n = 10$	$n = 20$	$n = 50$
0.02	IW-EXACT	2.3	8.1	9.2	21	156
0.02	CC-GATE	19	34	22	31	72
0.02	CC-PCL	18	**47**	**64**	**112**	**466**
0.2	IW-EXACT	2.2	3.8	4.1	9.9	22
0.2	CC-PCL	**22**	**18**	**18**	**19**	20
0.5	IW-EXACT	1.5	1.5	1.5	2.8	3.0
0.5	CC-PCL	**18**	**4.2**	**5.3**	**4.4**	**3.9**

The server averages these estimates to obtain the shared base, after which each client reconstructs its curvature using the shared estimate and its local $\hat{m}_i$. The client-specific component therefore has dimension d rather than d^2.

Finite local datasets. When a finite local dataset is revisited after a complete pass, its empirical transfer statistics are deterministic functions of that dataset. The streaming estimation error considered in the online analysis then disappears after the full empirical moments have been computed. This case removes streaming uncertainty but does not remove finite-sample error relative to the population transfer statistics.

E ADDITIONAL EXPERIMENTAL RESULTS

This appendix provides additional robustness and scope checks for Section 5. We study scaling with the number of clients, the optimization regime, partial participation, and personalized classification with linear heads.

E.1 SCALING WITH THE NUMBER OF CLIENTS AND HETEROGENEITY

Proposition 3(b) predicts that pooling is most useful when transfer statistics are similar and that its estimation benefit increases with the number of collaborating clients. Table 6 varies both n and $\delta = \delta_{\text{env}} = \delta_{\text{obj}}$. At low heterogeneity, CC-PCLimproves as n increases, while the local estimator has no corresponding cross-client benefit. As heterogeneity increases, the gain from unstructured pooling saturates, consistent with the heterogeneity term in Proposition 3(b).

E.2 OPTIMIZATION REGIME AND STRUCTURED ESTIMATION

We repeat the controlled experiment using the decaying step size and Polyak averaging assumed most closely by the analysis. In this regime, fresh sampling noise is progressively averaged and IW-EXACT becomes a stronger comparator. Table 7 shows that unstructured pooling loses much of its advantage as heterogeneity increases. In contrast, CC-STRUCT shares the common component while retaining the client-specific shift and remains effective across the heterogeneous settings. This behavior is consistent with the structured regime of Proposition 3(c).

E.3 PARTIAL PARTICIPATION

We next activate five of the $n = 20$ clients uniformly at random in each round. An inactive client draws no sample and performs no update. Table 8 reports results after $T = 2400$ rounds and after $T = 9600$ rounds, where each client has observed approximately the same number of samples as under full participation.

Partial participation reduces the number of fresh samples available for sample-level transfer, while the transfer-statistic estimators retain information accumulated across previous rounds. At $T = 9600$, CC-LOCAL approaches its full-participation behavior because each client has observed a comparable number of samples. Pooling gives its largest gains at low heterogeneity, where intermittent participation leaves the local statistics poorly estimated. As heterogeneity increases, the local and

Table 7: Speedup over independent learning with $\alpha_t = 0.03/(1 + t/200)$ and Polyak averaging over the second half ($T = 4000$, 3 runs).

δ	method	$n = 2$	$n = 5$	$n = 10$	$n = 20$	$n = 50$
0.02	IW-EXACT	3.9	6.3	10	25	49
0.02	CC-PCL	0.7	3.7	6.3	12	35
0.02	CC-STRUCT	1.0	**8.0**	**19**	**49**	**97**
0.2	IW-EXACT	3.0	4.0	4.9	**11**	**14**
0.2	CC-PCL	0.3	1.5	1.3	1.4	1.8
0.2	CC-STRUCT	1.0	**6.1**	**7.8**	9.1	8.9
0.5	IW-EXACT	1.6	1.6	2.2	2.8	**3.4**
0.5	CC-PCL	0.7	1.7	1.3	1.3	1.4
0.5	CC-STRUCT	0.9	**4.2**	**3.9**	**2.9**	2.6

Table 8: Partial participation with 5 of 20 clients active per round. Final MSE_0 under a constant step, averaged over 10 runs. Oracle Λ uses the population empirical-Bayes weights of Proposition 3(b).

T	$(\delta_{\mathrm{env}}, \delta_{\mathrm{obj}})$	Indep.	IW-NONE	IW-EXACT	Own-mean	CC-LOCAL	CC-PCL	CC-PCL, oracle Λ
2400	$(0, 0)$	.0122	.0023	.0023	.0033	.0015	**.00027**	.00012
	$(0.05, 0.05)$	.0121	.0023	.0024	.0032	.0014	**.00035**	.00016
	$(0.2, 0.2)$	.0122	.0056	.0085	.0032	**.0016**	.0016	.0015
	$(0.5, 0.5)$	.0128	.0342	.0973	.0048	**.0037**	.0038	.0038
9600	$(0, 0)$	.0114	.0021	.0021	.0031	.00038	**.00010**	.00007
$(0.05, 0.05)$	.0118	.0021	.0024	.0032	.00039	**.00015**	.00010	
	$(0.2, 0.2)$	.0128	.0053	.0084	.0033	**.00064**	.00066	.00065
	$(0.5, 0.5)$	.0135	.0322	.1090	.0050	**.0033**	.0033	.0033

pooled estimators become nearly indistinguishable. The oracle and plug-in empirical-Bayes weights show the same qualitative transition.

E.4 NONLINEAR HEADS

Proposition 1 gives an exact finite-dimensional transfer characterization because the least-squares head has an affine gradient. We examine whether the same transfer principle remains useful when the personalized head is nonlinear. Let $h(\phi(s); x)$ denote the client head and define

$$\ell_i(x; s) = \frac{1}{2}\big(h(\phi(s); x) - y_i(s)\big)^2, \qquad g^i(x; s) = \nabla_x \ell_i(x; s),$$

with mean gradient $\bar{g}_i(x) = \mathbb{E}_{\mu_i}[g^i(x; s)]$. The transfer target at reference $\bar{x}_t$ is now $\bar{g}_i(\bar{x}_t)$, which no longer admits the exact finite-dimensional representation of the linear case.

Local first-order extension. Because historical gradients are evaluated at historical references $\bar{x}_\tau$, we maintain

$$m_t^i = \sum_{\tau<t} w_{t,\tau} g^i(\bar{x}_\tau; s_\tau^i), \qquad \hat{\bar{x}}_t = \sum_{\tau<t} w_{t,\tau} \bar{x}_\tau.$$

Here m_t^i estimates the client's mean gradient near the averaged reference $\hat{\bar{x}}_t$. We transport this estimate to the current reference using

$$u_t^i = m_t^i + \widehat{G}_t^i(\bar{x}_t - \hat{\bar{x}}_t), \tag{9}$$

where $\widehat{G}_t^i$ estimates the Gauss–Newton matrix

$$G_i(x) = \mathbb{E}_{\mu_i}\left[J(s; x) J(s; x)^\top\right], \qquad J(s; x) = \partial_x h(\phi(s); x).$$

This correction is approximate because

Table 9: Nonlinear personalized heads: test MSE above the per-client LM reference ($\times 10^{-3}$), averaged over three seeds and the final 500 of 6000 rounds. † uses $\alpha = 0.1$; all other methods use $\alpha = 0.3$.

	δ_{env}	Indep.	FedAvg	Ditto	FedPer	APFL	Pop. avg.†	IW-exact†	Own mean	CC	CC-class
CIFAR-10	0	8.7	16.2	15.6	6.4	8.9	31.9	32.0	4.1	**3.7**	**3.7**
	0.5	6.9	18.9	16.4	6.2	7.0	70.6	58.1	3.3	**2.9**	**2.9**
	0.9	5.7	31.0	17.0	6.5	5.8	74.6	53.5	4.8	**4.7**	**4.7**
FEMNIST	0	35.4	22.5	45.3	26.6	33.4	117.8	117.3	20.8	**19.8**	**19.8**
	0.5	36.0	29.6	49.0	28.3	33.3	133.7	106.2	22.5	**21.9**	22.0
	0.9	31.5	38.7	51.2	26.5	30.7	108.7	65.9	23.4	**22.9**	23.0

Table 10: Nonlinear personalized classification. Entries report test MSE / accuracy (%) / F1 (%), averaged over three seeds and the final 500 of 6000 rounds. † uses $\alpha = 0.1$; all other methods use $\alpha = 0.3$. The LM head is a training-set reference.

	δ_{env}	LM head	Indep.	FedAvg	Ditto	FedPer	APFL	IW-exact†	CC
CIFAR-10	0	.138/87.0/86.3	.152/79.5/77.2	.224/64.0/63.9	.228/65.0/56.3	.149/80.7/78.5	.153/80.0/77.1	.340/72.3/66.2	**.139/81.9/80.1**
	0.5	.117/90.0/88.0	.113/85.8/81.5	.215/65.9/64.2	.177/75.2/65.2	.121/85.4/80.9	.116/85.7/81.1	.402/78.4/70.5	**.107/86.9/82.9**
	0.9	.060/96.1/87.4	.050/94.7/77.4	.204/68.1/55.7	.098/88.1/65.8	.062/93.9/74.9	.054/94.2/75.8	.121/90.9/74.1	**.049/94.9/78.1**
FEMNIST	0	.178/74.8/74.0	.290/54.1/39.5	.251/52.1/41.0	.293/52.2/33.6	.280/55.5/43.7	.289/54.3/40.1	.369/53.9/42.4	**.244/58.9/51.7**
	0.5	.146/80.7/79.4	.265/59.5/50.3	.249/53.3/51.5	.279/55.3/41.4	.250/62.0/55.7	.261/60.1/51.3	.389/59.1/49.0	**.223/65.6/60.9**
	0.9	.084/90.3/89.2	.214/68.9/58.7	.248/55.0/40.0	.262/59.8/39.1	.194/72.5/65.4	.202/70.5/61.3	.432/65.4/55.0	**.175/75.2/70.1**
FEMNIST writers	–	–	.254/63.4/51.7	.247/56.1/40.4	.281/56.4/37.3	.237/66.6/57.3	.243/65.4/54.3	div.	**.230/67.1/58.0**

$$\nabla \bar{g}_i(x) = G_i(x) + \mathbb{E}_{\mu_i}\left[\big(h(\phi(s); x) - y_i(s)\big)\nabla_x^2 h(\phi(s); x)\right].$$

The second term vanishes for a linear head, recovering the setting of the main analysis. For nonlinear heads, Eq. (9) is instead a local first-order approximation whose accuracy depends on the Gauss–Newton approximation and the distance between $\hat{\bar{x}}_t$ and $\bar{x}_t$. We therefore use the following experiments as a scope check rather than as validation of the exact theory.

Experimental setup. We use the frozen CIFAR-10 and FEMNIST representations from Section 5 and replace the linear personalized predictor with a two-layer tanh MLP of width 32, giving $p = 2113$ parameters on CIFAR-10 and $p = 2753$ on FEMNIST. Experiments use $n = 20$ clients, batch size 1, 6000 rounds, and three seeds under the client heterogeneity constructions used in the main experiments.

For continuous prediction, we report test MSE relative to a per-client Levenberg–Marquardt (LM) reference fitted on the client's full training set. We compare independent learning, FedAvg, Ditto, FedPer, APFL, uncorrected population-gradient transfer, IW-EXACT, the client's own running-mean transfer, and the Gauss–Newton-corrected estimator in Eq. (9) (CC). The collaboration gate is enabled for the control-variate methods. All methods use $\alpha = 0.3$ except the two sample-transfer methods, which use $\alpha = 0.1$ because they are unstable at $\alpha = 0.3$ on CIFAR-10.

Continuous prediction. Table 9 shows that the nonlinear extension behaves differently from the exact linear setting. The sample-level transfer variants are unstable at the common step size on CIFAR-10 and remain worse than independent learning at the smaller stable step size, including when exact density ratios are supplied. Thus exact importance correction alone is insufficient for this nonlinear implementation.

In contrast, estimating the transfer direction from the client's own gradient history improves over independent learning in all six settings. The Gauss–Newton correction gives a further, generally smaller improvement, while class-conditioning the curvature changes the results little. This behavior is consistent with Eq. (9): the curvature term only corrects for the reference lag $\bar{x}_t - \hat{\bar{x}}_t$, whereas the running mean provides the primary estimate of the client's transfer direction.

Personalized classification. We additionally evaluate the nonlinear heads on personalized classification tasks with discrete labels. Table 10 reports test MSE, accuracy, and F1. The LM head is fitted separately to each client's training data and is included as a reference rather than as a performance bound.

The classification results show a similar pattern. Relative to independent learning, CC improves accuracy in all six controlled CIFAR-10 and FEMNIST settings, with larger gains on FEMNIST. F1

Table 11: Component ablation. Synthetic benchmark: final MSE_0 at $T = 2400$ with a constant step (10 runs), and tail-averaged MSE_0 under the decaying step of Table 7 (10 runs). Nonlinear columns report excess test MSE ($\times 10^{-3}$) for the 2-layer MLP head at $\delta_{\mathrm{env}} = 0.5$ (3 seeds). "div." denotes divergence at the common step size. The gate is active in the nonlinear runs.

	synthetic, const. $(0.2, 0.2)$	synthetic, const. $(0.5, 0.5)$	synth., decay $(0.2, 0.2)$	CIFAR-10 MSE	FEMNIST MSE
Independent learning	.0116	.0118	.00038	6.90	36.0
Sample transfer, no ratios (IW-NONE)	.0052	.0446	.00439	div.	153
Sample transfer, exact ratios (IW-EXACT)	.0014	.0053	**.00003**	div.	124
Own running mean, no curvature	.0032	.0048	.00037	3.28	22.5
+ local transfer statistics (CC-LOCAL)	.0006	.0029	.00028	**2.86**	**21.9**
+ reference-vs-moment gate (CC-GATE)	**.0006**	**.0029**	.00027	–	–
+ empirical-Bayes pooling (CC-PCL)	.0006	.0029	.00030	2.87	**21.9**
+ structure (mean shift)	.0002	–	.00002	2.87	22.0

also improves in all six settings. On the real FEMNIST writers, CC increases accuracy from 63.4% to 67.1% and F1 from 51.7% to 58.0%.

FedPer is competitive in several nonlinear settings, and the relative ordering of the personalized methods varies across datasets and heterogeneity levels. We therefore do not interpret these experiments as establishing general superiority over nonlinear PFL methods. Instead, they test whether the transfer construction retains useful behavior once the affine-gradient assumption is removed.

The nonlinear experiments support a limited extension of the transfer principle rather than an extension of the main theorem. Estimating the client's mean transfer direction from its gradient history remains useful in the tested settings, while the Gauss–Newton term provides a smaller correction for reference drift. The exact finite-dimensional characterization remains specific to the affine-gradient setting.

F Proofs for the linear case

F.1 Conventions

Throughout, $\mathcal{F}_t$ is the sigma-field generated by all samples $s_\tau^j :, j \le n,\ \tau < t$. The personalized iterate x_t^i, the reference iterates $\bar{x}_t, \bar{\theta}_t$, and the transferred direction u_t^i are $\mathcal{F}_t$-measurable, while the fresh sample s_t^i is independent of $\mathcal{F}_t$. Thus, the moment estimates used in the analysis are formed from samples observed before round t.

The reference iterates are projected after each averaging step onto balls of radii $2G_x$ and $2G_b$. Since $\bar{x}^\star$ and $\bar{\theta}^\star$ lie in these balls by Assumption 1(ii), projection does not increase their distance to the limiting references and ensures $|\bar{x}_t| \le 2G_x$ and $|\bar{\theta}_t| \le 2G_b$ deterministically. Assumption 1(ii) further gives $|A(s) - H_i| \le G_A$, $|\Phi(s) - F_i| \le G_\Phi$, $|H_i| \le G_A$, and $|F_i| \le G_\Phi$. With $L = \max G_A, G_\Phi$ and $G_x, G_b \ge \frac{1}{2}$, we have $L \le \sigma$.

Recall

$$\bar{g}_i(x) := H_i x - q_i, \qquad e_t^i := u_t^i - (H_i \bar{x}_t - F_i \bar{\theta}_t), \qquad D_t^i := \mathbb{E}\|x_t^i - x_i^\star\|^2.$$

Since $\bar{g}_i(x_i^\star) = 0$, Assumption 1 implies

$$\langle \bar{g}_i(x), x - x_i^\star \rangle \ge \lambda |x - x_i^\star|^2, \qquad |\bar{g}_i(x)| \le \sigma |x - x_i^\star|. \tag{10}$$

We repeatedly use $|a+b|^2 \le 2|a|^2 + 2|b|^2$, Young's inequality $2\langle a, b\rangle \le \eta|a|^2 + \eta^{-1}|b|^2$ for $\eta > 0$, and $|Ex| \le |E|_F |x|$.

Implementation timing. The implementation incorporates the current sample into the running moment estimates before forming the transferred direction, whereas the analysis uses the past-only estimates above to preserve $\mathcal{F}_t$-measurability. The resulting same-sample perturbation is $O(1/N_t)$,

where N_t is client i's number of previously observed samples, and contributes only lower-order terms to the stochastic-approximation recursion. The theorem therefore establishes rate-level equivalence rather than analyzing the implemented iterate sequence verbatim.

F.2 Proof of Proposition 1

Proof. Conditioning on $\mathcal{F}_t$ and using the independence of s_t^i from the history gives

$$\mathbb{E}[g_t^i(x_t^i) \mid \mathcal{F}_t] = H_i x_t^i - q_i, \qquad \mathbb{E}[r_t^i \mid \mathcal{F}_t] = H_i \bar{x}_t - F_i \bar{\theta}_t.$$

Therefore,

$$\mathbb{E}[g_t^i(x_t^i) + u_t^i - r_t^i \mid \mathcal{F}_t] = H_i x_t^i - q_i + \mathbb{E}[u_t^i \mid \mathcal{F}_t] - (H_i \bar{x}_t - F_i \bar{\theta}_t).$$

The conditional mean equals $H_i x_t^i - q_i$ if and only if

$$\mathbb{E}[u_t^i \mid \mathcal{F}_t] = H_i \bar{x}_t - F_i \bar{\theta}_t.$$

□

F.3 Proof of Proposition 2

Proof. Substitute $b_i(s) = \Phi(s)\theta_i^\star$ into $\tilde{g}_t^i = g_t^i(x_t^i) - r_t^i + u_t^i$ and regroup:

$$\begin{aligned} \tilde{g}_t^i &= \left[A(s_t^i)x_t^i - \Phi(s_t^i)\theta_i^\star\right] - \left[A(s_t^i)\bar{x}_t - \Phi(s_t^i)\bar{\theta}_t\right] + u_t^i \\ &= A(s_t^i)(x_t^i - \bar{x}_t) - \Phi(s_t^i)(\theta_i^\star - \bar{\theta}_t) + u_t^i. \end{aligned} \tag{11}$$

Mean: Since s_t^i is independent of $\mathcal{F}_t$, while x_t^i, $\bar{x}_t$, $\bar{\theta}_t$, and u_t^i are $\mathcal{F}_t$-measurable,

$$\begin{aligned} \mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t] &= H_i(x_t^i - \bar{x}_t) - F_i(\theta_i^\star - \bar{\theta}_t) + u_t^i \\ &= \underbrace{\left(H_i x_t^i - F_i \theta_i^\star\right)}_{=\bar{g}_i(x_t^i)} + \underbrace{\left[u_t^i - \left(H_i \bar{x}_t - F_i \bar{\theta}_t\right)\right]}_{=e_t^i}, \end{aligned}$$

where $F_i\theta_i^\star = q_i$. Hence

$$\mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t] = \bar{g}_i(x_t^i) + e_t^i.$$

Variance: Subtracting the conditional mean from (11) gives

$$\tilde{g}_t^i - \mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t] = \big(A(s_t^i) - H_i\big)(x_t^i - \bar{x}_t) - \big(\Phi(s_t^i) - F_i\big)(\theta_i^\star - \bar{\theta}_t).$$

Using $\|Mv\| \le \|M\|\,\|v\|$, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and Assumption 1,

$$\begin{aligned} &\mathbb{E}\Big[\big\|\tilde{g}_t^i - \mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t]\big\|^2 \,\Big|\, \mathcal{F}_t\Big] \\ &\qquad \le 2G_A^2 \left\|x_t^i - \bar{x}_t\right\|^2 + 2G_\Phi^2 \left\|\theta_i^\star - \bar{\theta}_t\right\|^2. \end{aligned}$$

□

F.4 Reference-model convergence

Lemma 1 (Federated averaging)**.** *Let $\bar{x}^\star$ and $\bar{\theta}^\star$ solve $H_0\bar{x} = \bar{q}$ and $F_0\bar{\theta} = \bar{q}$, respectively. Under Assumption 1 and the step size of Theorem 1, there is a numerical constant C_0 such that*

$$\mathbb{E}\,\|\bar{x}_t - \bar{x}^\star\|^2 \le C_0\kappa^2\frac{1+\ln t}{nt}, \qquad \mathbb{E}\,\left\|\bar{\theta}_t - \bar{\theta}^\star\right\|^2 \le C_0\kappa_\theta^2\frac{1+\ln t}{nt},$$

where $\kappa_\theta := \sigma/\lambda_\theta \le \kappa$.

Proof. Each reference iterate is a federated stochastic-approximation iteration for a strongly monotone linear system. Indeed, $\mathrm{sym}(H_0) \succeq \lambda_x I$ because $H_0 = \frac{1}{n}\sum_i H_i$, while $\mathrm{sym}(F_0) \succeq \lambda_\theta I$ by Assumption 1(i). The averaged stochastic residuals are unbiased and, on the projection balls, have conditional variance bounded by $C\sigma^2/n$ for a numerical constant C. The standard federated stochastic-approximation rate therefore gives the stated bounds; see Zhang & Azizan (2026, Proposition 1). Projection onto a ball containing the corresponding solution does not increase its distance to that solution. Finally, $\lambda = \min\{\lambda_x, \lambda_\theta\}$ implies $\kappa_\theta \le \kappa$. □

F.5 Proof of Proposition 3

Proof. Transfer-direction error. By definition,

$$e_t^i = (\widehat{H}_t^i - H_i)\bar{x}_t - (\widehat{F}_t^i - F_i)\bar{\theta}_t.$$

Using $\|Ex\| \le \|E\|_F \|x\|$, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and the pathwise projection bounds $\|\bar{x}_t\| \le 2G_x$ and $\|\bar{\theta}_t\| \le 2G_b$,

$$\left\|e_t^i\right\|^2 \le 8G_x^2 \left\|\widehat{H}_t^i - H_i\right\|_F^2 + 8G_b^2 \left\|\widehat{F}_t^i - F_i\right\|_F^2.$$

Hence

$$\mathbb{E}\left\|e_t^i\right\|^2 \le 8(G_x^2 + G_b^2)\varepsilon_i(t) = 8G^2\varepsilon_i(t).$$

Averaging the same inequality over clients gives

$$\frac{1}{n}\sum_i \mathbb{E}\left\|e_t^i\right\|^2 \le 8G_x^2\bar{\varepsilon}_H(t) + 8G_b^2\bar{\varepsilon}_F(t) \le 8G^2\bar{\varepsilon}(t),$$

where $\bar{\varepsilon}_H$ and $\bar{\varepsilon}_F$ denote the agent-averaged errors for $\widehat{H}$ and $\widehat{F}$, respectively. No independence between the transfer-statistic estimates and the reference iterates is required.

(a) Local transfer. For the running local mean,

$$\widehat{H}_t^i - H_i = \frac{1}{t}\sum_{\tau=0}^{t-1}\left(A(s_\tau^i) - H_i\right).$$

The summands are independent and centered, so

$$\mathbb{E}\left\|\widehat{H}_t^i - H_i\right\|_F^2 = \frac{1}{t^2}\sum_{\tau=0}^{t-1}\mathbb{E}\left\|A(s_\tau^i) - H_i\right\|_F^2 \le \frac{V}{t}$$

by Assumption 1(iii). The same argument applies to $\widehat{F}_t^i$, proving $\varepsilon_i(t) \le V/t$.

(b) Federated curvature. Fix an entry e and write

$$\delta_j := (H_j - H_0)_e, \qquad \epsilon_j := (\widehat{H}_t^j - H_j)_e, \qquad \bar{\epsilon} := \frac{1}{n}\sum_j \epsilon_j.$$

Here δ_j is deterministic, while the ϵ_j are independent across clients, centered, and assumed to have common variance v_e. Moreover,

$$\sum_e v_e \le \frac{V}{t}, \qquad \frac{1}{n}\sum_j \delta_j = 0.$$

Thus $\widehat{H}_e^0 = H_{0,e} + \bar{\epsilon}$. For a shrinkage weight $\Lambda \in [0,1]$, the pooled estimate

$$\widehat{H}_e^{i,\mathrm{fed}} = \widehat{H}_e^0 + \Lambda(\widehat{H}_{t,e}^i - \widehat{H}_e^0)$$

satisfies

$$\widehat{H}_e^{i,\mathrm{fed}} - H_{i,e} = -(1-\Lambda)\delta_i + \Lambda\epsilon_i + (1-\Lambda)\bar{\epsilon}.$$

Because $\bar{\epsilon}$ contains ϵ_i with weight $1/n$,

$$(\Lambda\epsilon_i + (1-\Lambda)\bar{\epsilon}) = v_e\left[\Lambda^2 + \frac{1-\Lambda^2}{n}\right].$$

Hence the exact per-client risk is

$$R_i(\Lambda) = (1-\Lambda)^2\delta_i^2 + v_e\left[\Lambda^2 + \frac{1-\Lambda^2}{n}\right]. \tag{12}$$

Let

$$\tau_e^2 := \frac{1}{n}\sum_j \delta_j^2, \qquad \Lambda_e^{\mathrm{EB}} := \frac{\tau_e^2}{\tau_e^2 + v_e}.$$

The agent-averaged risk is

$$\frac{1}{n}\sum_j R_j(\Lambda) = (1-\Lambda)^2\tau_e^2 + v_e\left[\Lambda^2 + \frac{1-\Lambda^2}{n}\right].$$

Its exact minimizer is

$$\Lambda^\star = \frac{\tau_e^2}{\tau_e^2 + v_e(1-1/n)}.$$

The oracle empirical-Bayes weight Λ_e^{EB} differs only by the finite-n correction and satisfies

$$\frac{1}{n}\sum_j R_j(\Lambda_e^{\mathrm{EB}}) = \frac{\tau_e^2 v_e}{\tau_e^2 + v_e} + \frac{v_e\big(1-(\Lambda_e^{\mathrm{EB}})^2\big)}{n} \le \min\{\tau_e^2, v_e\} + \frac{v_e}{n}.$$

Summing over entries, with

$$T_H^2 := \sum_e \tau_e^2,$$

and using

$$\sum_e \min\{\tau_e^2, v_e\} \le \min\left\{T_H^2, \frac{V}{t}\right\},$$

gives

$$\bar{\varepsilon}_H(t) \le \frac{V}{nt} + \min\left\{T_H^2, \frac{V}{t}\right\}.$$

The same argument for $\widehat{F}$ gives the corresponding bound with T_F^2. Therefore, with

$$T^2 := \max\{T_H^2, T_F^2\},$$

we obtain

$$\bar{\varepsilon}(t) \le \frac{V}{nt} + \min\left\{T^2, \frac{V}{t}\right\}.$$

When $T^2 = 0$, all client transfer statistics agree and $\Lambda_e^{\mathrm{EB}} = 0$. Each client then uses the population mean, and (12) gives $R_i = v_e/n$ entry-wise. Summing over entries yields

$$\varepsilon_i(t) \le \frac{V}{nt}$$

for every client.

(c) Structured environment shift. Let

$$\hat{\mathcal{A}} = \mathcal{A}(\hat{H}_{\mathrm{sh}}, \hat{\nu}_i).$$

By the assumed Lipschitz properties of $\mathcal{A}$,

$$\left\|\hat{\mathcal{A}} - H_i\right\|_F \le \left\|\hat{H}_{\mathrm{sh}} - H_{\mathrm{sh}}\right\|_F + L_i\left\|\hat{\nu}_i - \nu_i\right\|.$$

Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the assumed mean-squared error bounds,

$$\mathbb{E}\left\|\hat{\mathcal{A}} - H_i\right\|_F^2 \le \frac{2V}{nt} + \frac{2L_i^2 k V_\nu}{t}.$$

The same argument applies to the corresponding structured estimator of F_i, proving part (c). □

Remark 1 (Instantiating Proposition 3(c)). The two models of Section 3 instantiate the structured bound under additional regularity. In the mean-shift model,

$$\mathcal{A}(A_{\mathrm{base}}, m) = \big(I + \epsilon(I + mm^\top)\big)A_{\mathrm{base}},$$

and, on a bounded set of m,

$$L_i \le 2\epsilon\,\|m_i\|\,\|A_{\mathrm{base}}\|_F, \qquad k = d.$$

Thus the client-specific contribution in Proposition 3(c) scales with the dimension and sensitivity of the mean-shift parameter. Estimating the shared component additionally requires the inverse map associated with $I + \epsilon(I + mm^\top)$ to be uniformly well conditioned on the parameter set.

For label skew,

$$\mathcal{A}(\{H_c\}, p_i) = \sum_c p_i(c) H_c, \qquad k = C.$$

The per-class shared components can be estimated from samples pooled across clients. Obtaining the $O(1/(nt))$ shared-estimation term requires each class to have population frequency bounded below, $p_0(c) \geq p_{\min} > 0$; the resulting constant depends on $1/p_{\min}$. We do not track these model-specific constants further.

Plug-in weights. Proposition 3(b) uses the oracle weight $\Lambda_e^{\mathrm{EB}} = \tau_e^2/(\tau_e^2 + v_e)$. In practice, both quantities are estimated. We use

$$\hat{\Lambda}_e = \frac{\hat{\tau}_e^2}{\hat{\tau}_e^2 + \hat{\sigma}_{i,e}^2}, \qquad \hat{\tau}_e^2 = \left[\frac{1}{n} \sum_j (\widehat{H}^j - \widehat{H}^0)_e^2 - \frac{1}{n} \sum_j \hat{\sigma}_{j,e}^2 \right]_+ ,$$

where $\hat{\sigma}_{i,e}^2$ estimates the sampling variance of the local running mean. The following result quantifies the additional error from estimating the pooling weight under an independent-split construction.

Proposition 4 (Plug-in pooling weights)**.** *Fix one entry and write*

$$a := \tau_e^2, \qquad b := v_e, \qquad \Lambda^{\mathrm{EB}} := \frac{a}{a+b},$$

with $\Lambda^{\mathrm{EB}} = 0$ when $a + b = 0$. Suppose the corresponding entries of $A(s)$ and $\Phi(s)$ have uniformly bounded fourth moments, and let $\hat{\Lambda}$ be computed from a sample split independent of the means being pooled using the plug-in estimates above. Then the agent-averaged plug-in risk satisfies

$$R_{\mathrm{plug}} = O\left(\frac{b}{n} + \frac{ab}{a+b} + \frac{b^2}{a+b} \left[\frac{1}{n} + \frac{1}{t} \right] \right),$$

where the hidden constant depends only on the normalized fourth-moment bound. Thus the plug-in estimator retains the oracle risk terms $b/n + ab/(a+b)$, with an additional contribution due to estimation of the pooling weight.

Proof sketch. Write

$$\hat{\Lambda} = \Lambda^{\mathrm{EB}} + \delta.$$

Relative to the oracle-weight estimator, replacing Λ^{EB} by $\hat{\Lambda}$ introduces a perturbation proportional to

$$\delta(\delta_j + \epsilon_j - \bar{\epsilon}).$$

Averaging over clients and using the independence supplied by the sample split bounds the resulting excess risk, up to numerical constants, by

$$\frac{1}{n} \frac{ab}{a+b} + \mathbb{E}[\delta^2](a+b).$$

Under the fourth-moment assumption, the plug-in estimates of a and b satisfy the standard variance bounds. Applying these bounds to $(a, b) \mapsto a/(a+b)$ gives

$$\mathbb{E}[\delta^2] = O\left(\frac{b^2}{(a+b)^2} \left[\frac{1}{n} + \frac{1}{t} \right] \right).$$

Substitution, together with the oracle risk bound

$$R_{\mathrm{oracle}} = O\left(\frac{b}{n} + \frac{ab}{a+b} \right),$$

gives the stated result. When $a + b = 0$, the entry is deterministic and the risk is zero. □

F.6 ONE-STEP RECURSION

Lemma 2. *If $\alpha_t \le \lambda/(12\sigma^2)$, then*

$$D_{t+1}^i \le \left(1 - \frac{\alpha_t\lambda}{2}\right) D_t^i + \frac{4\alpha_t}{\lambda}\mathbb{E}\left\|e_t^i\right\|^2 + 4\alpha_t^2\left(G_A^2\mathbb{E}\left\|x_i^\star - \bar{x}_t\right\|^2 + G_\Phi^2\mathbb{E}\left\|\theta_i^\star - \bar{\theta}_t\right\|^2\right).$$

Proof. Write $\Delta x := x_t^i - x_i^\star$. Since $x_{t+1}^i = x_t^i - \alpha_t\tilde{g}_t^i$, expanding the square and conditioning on $\mathcal{F}_t$ gives

$$\mathbb{E}\left[\left\|x_{t+1}^i - x_i^\star\right\|^2 \,\middle|\, \mathcal{F}_t\right] = \left\|\Delta x\right\|^2 - 2\alpha_t\left\langle\mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t], \Delta x\right\rangle + \alpha_t^2\mathbb{E}\left[\left\|\tilde{g}_t^i\right\|^2 \,\middle|\, \mathcal{F}_t\right]. \tag{13}$$

Cross term. By Proposition 2,

$$\mathbb{E}[\tilde{g}_t^i \mid \mathcal{F}_t] = \bar{g}_i(x_t^i) + e_t^i.$$

Since $\bar{g}_i(x_i^\star) = 0$ and $\mathrm{sym}(H_i) \succeq \lambda I$,

$$\left\langle\bar{g}_i(x_t^i), \Delta x\right\rangle = \left\langle H_i\Delta x, \Delta x\right\rangle \ge \lambda\left\|\Delta x\right\|^2.$$

Young's inequality then gives

$$\begin{aligned}-2\alpha_t\left\langle\bar{g}_i(x_t^i) + e_t^i, \Delta x\right\rangle &\le -2\alpha_t\lambda\left\|\Delta x\right\|^2 + \alpha_t\lambda\left\|\Delta x\right\|^2 + \frac{\alpha_t}{\lambda}\left\|e_t^i\right\|^2 \\ &= -\alpha_t\lambda\left\|\Delta x\right\|^2 + \frac{\alpha_t}{\lambda}\left\|e_t^i\right\|^2.\end{aligned}$$

Square term. Using conditional second moment = conditional variance + squared conditional mean and Proposition 2,

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{g}_t^i\right\|^2 \,\middle|\, \mathcal{F}_t\right] &\le 2G_A^2\left\|x_t^i - \bar{x}_t\right\|^2 + 2G_\Phi^2\left\|\theta_i^\star - \bar{\theta}_t\right\|^2 \\ &\quad + 2\left\|\bar{g}_i(x_t^i)\right\|^2 + 2\left\|e_t^i\right\|^2.\end{aligned}$$

Moreover,

$$x_t^i - \bar{x}_t = \Delta x + (x_i^\star - \bar{x}_t),$$

so

$$\left\|x_t^i - \bar{x}_t\right\|^2 \le 2\left\|\Delta x\right\|^2 + 2\left\|x_i^\star - \bar{x}_t\right\|^2.$$

Since $\bar{g}_i(x_t^i) = H_i\Delta x$,

$$\left\|\bar{g}_i(x_t^i)\right\| \le G_A\left\|\Delta x\right\| \le \sigma\left\|\Delta x\right\|.$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{g}_t^i\right\|^2 \,\middle|\, \mathcal{F}_t\right] &\le (4G_A^2 + 2\sigma^2)\left\|\Delta x\right\|^2 + 2\left\|e_t^i\right\|^2 \\ &\quad + 4G_A^2\left\|x_i^\star - \bar{x}_t\right\|^2 + 2G_\Phi^2\left\|\theta_i^\star - \bar{\theta}_t\right\|^2 \\ &\le 6\sigma^2\left\|\Delta x\right\|^2 + 2\left\|e_t^i\right\|^2 \\ &\quad + 4G_A^2\left\|x_i^\star - \bar{x}_t\right\|^2 + 2G_\Phi^2\left\|\theta_i^\star - \bar{\theta}_t\right\|^2.\end{aligned}$$

Collecting terms. Substituting the preceding bounds into (13), the coefficient of $\left\|\Delta x\right\|^2$ satisfies

$$1 - \alpha_t\lambda + 6\alpha_t^2\sigma^2 \le 1 - \frac{\alpha_t\lambda}{2},$$

because $\alpha_t \le \lambda/(12\sigma^2)$.

For the transfer-estimation term, Assumption 1 gives $\lambda \le \sigma$, and hence

$$\alpha_t \le \frac{\lambda}{12\sigma^2} \le \frac{1}{12\lambda}.$$

Thus

$$\frac{\alpha_t}{\lambda} + 2\alpha_t^2 \le \frac{\alpha_t}{\lambda} + \frac{\alpha_t}{6\lambda} \le \frac{4\alpha_t}{\lambda}.$$

Consequently,

$$\begin{aligned}\mathbb{E}\left[\left\|x_{t+1}^i - x_i^\star\right\|^2 \,\middle|\, \mathcal{F}_t\right] &\le \left(1 - \frac{\alpha_t\lambda}{2}\right)\left\|\Delta x\right\|^2 + \frac{4\alpha_t}{\lambda}\left\|e_t^i\right\|^2 \\ &\quad + 4\alpha_t^2G_A^2\left\|x_i^\star - \bar{x}_t\right\|^2 + 2\alpha_t^2G_\Phi^2\left\|\theta_i^\star - \bar{\theta}_t\right\|^2.\end{aligned}$$

Taking full expectations and enlarging the final coefficient from 2 to 4 proves the result. □

F.7 Unrolling the $1/t$ recursion

Lemma 3. *Let*

$$u_{t+1} \le \left(1 - \frac{a}{t+t_0}\right) u_t + \frac{b_t}{(t+t_0)^{1+r}}, \qquad t \ge 0,$$

where $r \in \{1,2\}$, $a \ge r$, $t_0 \ge a$, and $b_t \le B(1+\ln(t+1))$. Then there is a constant $C_{a,r}$ depending only on (a,r) such that

$$u_t \le \frac{C_{a,r} B(1+\ln(t+1))(1+\ln(t+t_0))^{\mathbf{1}\{a=r\}}}{(t+t_0)^r} + \left(\frac{t_0}{t+t_0}\right)^a u_0.$$

Thus the recursion decays as $\tilde{O}(t^{-r})$, with one additional logarithmic factor in the boundary case $a = r$.

Proof. Since $t_0 \ge a$, all contraction factors are nonnegative. For $0 \le k < t$,

$$\prod_{\tau=k+1}^{t-1} \left(1 - \frac{a}{\tau+t_0}\right) \le \exp\left(-a \sum_{\tau=k+1}^{t-1} \frac{1}{\tau+t_0}\right) \le \left(\frac{k+1+t_0}{t+t_0}\right)^a.$$

Unrolling the recursion therefore gives

$$u_t \le \left(\frac{t_0}{t+t_0}\right)^a u_0 + \sum_{k=0}^{t-1} \frac{b_k}{(k+t_0)^{1+r}} \left(\frac{k+1+t_0}{t+t_0}\right)^a.$$

Since $t_0 \ge 1$,

$$k+1+t_0 \le 2(k+t_0),$$

and hence

$$u_t \le \left(\frac{t_0}{t+t_0}\right)^a u_0 + \frac{2^a B(1+\ln(t+1))}{(t+t_0)^a} \sum_{k=0}^{t-1} (k+t_0)^{a-1-r}.$$

If $a > r$, the power sum satisfies

$$\sum_{k=0}^{t-1} (k+t_0)^{a-1-r} \le C_{a,r} (t+t_0)^{a-r},$$

and therefore

$$u_t \le \frac{C_{a,r} B(1+\ln(t+1))}{(t+t_0)^r} + \left(\frac{t_0}{t+t_0}\right)^a u_0.$$

If $a = r$, then

$$\sum_{k=0}^{t-1} (k+t_0)^{-1} \le 1 + \ln(t+t_0),$$

which introduces one additional logarithmic factor and gives the stated bound. □

The recursion in Lemma 2 is linear in D_t^i with nonnegative forcing terms. We therefore apply Lemma 3 separately with $r = 1$ to the transfer-estimation and reference-mismatch terms, and with $r = 2$ to the reference-estimation term.

F.8 Proof of Theorem 1

Proof. Fix client i. Set

$$\alpha_t = \frac{c}{\lambda(t+t_0)}, \qquad c \ge 4, \qquad t_0 \ge \frac{12c\sigma^2}{\lambda^2}.$$

Then

$$\alpha_t \le \frac{c}{\lambda t_0} \le \frac{\lambda}{12\sigma^2},$$

so Lemma 2 applies with contraction factor $1 - \frac{c/2}{t+t_0}$. We first bound its forcing terms for $t \geq t_0$, for which $t + t_0 \leq 2t$.

Transfer-estimation term. By Proposition 3 and $\varepsilon_i(t) = \zeta_i(t)\sigma^2/(G^2 t)$,

$$\begin{aligned}\frac{4\alpha_t}{\lambda}\mathbb{E}\left\|e_t^i\right\|^2 &\leq \frac{4c}{\lambda^2(t+t_0)}\, 8G^2\varepsilon_i(t) \\ &= \frac{32c\sigma^2}{\lambda^2}\frac{\zeta_i(t)}{t(t+t_0)} \\ &\leq \frac{64c\kappa^2\zeta_i(t)}{(t+t_0)^2},\end{aligned}$$

where $\kappa^2 = \sigma^2/\lambda^2$.

Reference-mismatch term. Using

$$\mathbb{E}\left\|x_i^\star - \bar{x}_t\right\|^2 \leq 2\left\|x_i^\star - \bar{x}^\star\right\|^2 + 2\mathbb{E}\left\|\bar{x}_t - \bar{x}^\star\right\|^2$$

and the analogous inequality for θ, the deterministic parts of the final term in Lemma 2 give

$$\begin{aligned}&8\alpha_t^2\left(G_A^2\left\|x_i^\star - \bar{x}^\star\right\|^2 + G_\Phi^2\left\|\theta_i^\star - \bar{\theta}^\star\right\|^2\right) \\ &\quad = \frac{8c^2\kappa^2\Delta_i}{(t+t_0)^2},\end{aligned}$$

by the definition of Δ_i.

Reference-estimation term. The stochastic parts satisfy, by Lemma 1, $\kappa_\theta \leq \kappa$, and $G_A^2 + G_\Phi^2 \leq 2\sigma^2$,

$$\begin{aligned}&8\alpha_t^2\left(G_A^2\mathbb{E}\left\|\bar{x}_t - \bar{x}^\star\right\|^2 + G_\Phi^2\mathbb{E}\left\|\bar{\theta}_t - \bar{\theta}^\star\right\|^2\right) \\ &\quad \leq \frac{16c^2C_0\kappa^4}{(t+t_0)^2}\frac{1+\ln t}{nt} \\ &\quad \leq \frac{32c^2C_0\kappa^4(1+\ln t)}{n(t+t_0)^3}.\end{aligned}$$

The additional factor $1/t$ from the reference-model rate makes this term lower order.

Recursion. Combining the preceding bounds with Lemma 2, for $t \geq t_0$,

$$D_{t+1}^i \leq \left(1 - \frac{c/2}{t+t_0}\right)D_t^i + \frac{\kappa^2\left[64c\,\zeta_i(t) + 8c^2\Delta_i\right]}{(t+t_0)^2} + \frac{32c^2C_0\kappa^4(1+\ln t)}{n(t+t_0)^3}.$$

We now unroll this recursion from round t_0. Applying Lemma 3 to the shifted recursion with $a = c/2 \geq 2$, using $r = 1$ for the first forcing term and $r = 2$ for the reference-estimation term, gives

$$D_t^i \leq C\kappa^2\frac{(1+\ln t)\left[\Delta_i + \zeta_i^{\max}(t)\right]}{t} + C'\kappa^4\frac{(1+\ln t)^2}{nt^2} + \left(\frac{2t_0}{t+t_0}\right)^{c/2}D_{t_0}^i,$$

for numerical constants C, C', where

$$\zeta_i^{\max}(t) := \sup_{\tau \leq t}\zeta_i(\tau).$$

When $c = 4$, the $r = 2$ term is the boundary case $a = r = 2$ of Lemma 3, which accounts for the additional logarithmic factor.

Since $c/2 \geq 2$, the transient decays at least as $O(t^{-2})$ for fixed t_0. Absorbing logarithmic factors and numerical constants into $\tilde{O}$ therefore yields

$$D_t^i = \tilde{O}\left(\frac{\kappa^2}{t}\left[\Delta_i + \zeta_i^{\max}(t)\right]\right) + \tilde{O}\left(\frac{\kappa^4}{nt^2}\right),$$

up to the displayed initialization transient, which proves (7). □

F.9 SCOPE OF THE ANALYSIS

Relation to AffPCL heterogeneity. Items (f) and (j) of Lemma D.1 of Zhang & Azizan (2026) give

$$\|\theta_i^\star - \theta_c^\star\| \le 2\lambda^{-1}\sigma(\delta_{\mathrm{env}} + \delta_{\mathrm{obj}}), \qquad \|x_i^\star - x_c^\star\| \le 2\lambda^{-1}\sigma(\delta_{\mathrm{env}} + \delta_{\mathrm{obj}})$$

for their central client c. With $G_A, G_\Phi \le \sigma$, this implies

$$\Delta_i = O\big(\kappa^2(\delta_{\mathrm{env}} + \delta_{\mathrm{obj}})^2\big)\,.$$

Because the definitions of λ and σ differ slightly between the two analyses, we use this relation only to compare the order of the heterogeneity dependence.

Dependence of moment estimates and reference iterates. The transfer-estimation error is bounded using

$$\big\|e_t^i\big\| \le \Big\|\widehat{H}_t^i - H_i\Big\|_F \|\bar{x}_t\| + \Big\|\widehat{F}_t^i - F_i\Big\|_F \big\|\bar{\theta}_t\big\|\,,$$

together with the projection bounds on the reference iterates. The subsequent analysis uses pathwise inequalities and the $\mathcal{F}_t$-measurability of u_t^i, and does not require the moment estimates to be independent of the reference iterates. Theorem 1 therefore covers local moments formed from all past samples and pooled moments formed from all past samples when the pooling weights are fixed.

Plug-in pooling weights. Proposition 4 controls $\hat{\Lambda}$ when the variance estimates used to construct the pooling weights are computed from a sample split independent of the means being pooled. In the practical implementation, the weights and pooled means are estimated from the same running moments. The resulting dependence between the estimated weights and sampling noise is not covered by Proposition 4 or Theorem 1.

Rectangular objective reference. Theorem 1 analyzes the square case $m = d$. In the real-feature experiments, $m < d$ and the objective reference is updated using the stochastic normal-equation residual

$$\frac{1}{n}\sum_i F_i^\top(F_i\bar{\theta} - q_i),$$

estimated from each mini-batch. The corresponding population system uses $\frac{1}{n}\sum_i F_i^\top F_i$ in its strong-monotonicity condition. Because the mini-batch estimator of $F_i^\top F_i$ has an $O(1/\mathrm{batch})$ product bias, this rectangular implementation is an empirical extension and is not covered by Theorem 1.

G SUPPORTING EXPERIMENTS FOR THE MAIN RESULTS

This section provides additional results supporting the comparisons in Section 5. We report the full Dirichlet label-skew grid and an optimization check for the FEMNIST scarce-data experiments. Additional scaling, partial-participation, and classification results are given in Appendix E.

Baselines and evaluation. LOCAL-LS solves the empirical normal equations from all local samples observed by the client. LOCAL-RIDGE solves the local training problem with the ridge parameter selected by five-fold cross-validation. INDEP-AVG denotes independent SGD with Polyak–Ruppert averaging over the second half of the run. The same averaging procedure is applied to the corresponding averaged variants of the transfer methods.

Scarce-data protocol. The scarce-data experiments use a single pass without replacement over each client's n_{train} samples, with batch size one and $T = n_{\mathrm{train}}$. Each client has $n_{\mathrm{test}} = 1000$ test samples. For Dirichlet label skew, $p_i \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_C)$, and IW-EXACT uses the class-level density ratio $p_i(c)/\bar{p}(c)$.

Abundant-data check. We additionally evaluated the linear-head methods with 6000 local training samples per client. In this regime, the strong local closed-form estimator is competitive with or better

Table 12: Dirichlet label skew in the scarce single-pass regime (50 clients, excess test MSE $\times 10^{-3}$, 5 seeds).

features	n_{train}	α	INDEP-AVG	LOCAL-RIDGE	IW-NONE	IW-EXACT	CC-LOCAL	CC-PCL	CC-CLASS
CIFAR-10	64	0.1	59.9,±2.5	48.9,±2.4	237.8,±9.9	51.3,±1.9	72.7,±2.8	76.2,±2.3	57.9,±3.5
	64	0.3	88.2,±5.7	80.0,±4.7	199.7,±5.5	67.5,±4.0	101.2,±4.8	105.0,±4.3	76.7,±3.0
	64	1.0	81.9,±3.1	85.7,±4.9	123.6,±3.3	58.3,±2.2	96.7,±5.6	91.0,±4.5	70.3,±3.4
CIFAR-10	128	0.1	47.4,±3.3	39.0,±2.9	236.9,±8.1	38.8,±1.4	55.9,±3.3	57.6,±3.1	42.1,±1.4
	128	0.3	64.0,±3.2	56.1,±3.5	191.0,±5.0	48.6,±1.7	76.4,±2.4	75.3,±2.3	53.6,±1.5
	128	1.0	66.2,±3.0	63.7,±3.2	120.6,±3.3	44.2,±1.6	81.7,±2.2	71.4,±1.6	51.8,±1.5
CIFAR-10	256	0.1	30.8,±2.2	28.2,±2.5	228.5,±4.8	26.5,±1.5	36.1,±2.4	37.1,±2.6	28.0,±1.4
	256	0.3	45.6,±2.6	42.5,±2.0	189.4,±7.8	34.9,±1.7	54.9,±2.0	54.4,±1.4	37.3,±1.8
	256	1.0	54.3,±1.1	49.6,±1.0	119.6,±8.4	35.7,±1.0	68.0,±1.6	58.9,±1.1	39.7,±0.9
CIFAR-100	128	0.1	51.3,±0.6	44.7,±0.8	109.4,±2.9	36.7,±0.5	59.0,±0.8	61.8,±1.7	45.7,±0.6
	128	0.3	50.4,±1.5	50.0,±0.8	70.4,±1.4	32.9,±1.1	59.1,±1.3	54.5,±1.0	43.7,±1.0
	128	1.0	46.0,±0.6	46.2,±1.5	41.6,±0.9	27.4,±0.8	58.3,±1.4	40.8,±1.0	39.1,±1.5
CIFAR-100	256	0.1	36.4,±0.8	33.5,±0.8	102.1,±2.4	26.6,±0.9	43.5,±1.0	44.2,±0.5	32.1,±0.5
	256	0.3	39.8,±0.6	36.5,±0.6	63.7,±1.3	25.6,±0.7	49.6,±0.8	43.5,±0.9	32.8,±1.2
	256	1.0	40.3,±1.1	36.0,±1.0	37.2,±1.2	21.9,±0.7	50.7,±1.9	34.2,±1.2	29.9,±1.0
FEMNIST	64	0.1	21.4,±1.3	13.3,±0.5	32.0,±0.9	20.1,±0.9	21.6,±1.2	22.4,±0.9	21.1,±1.3
	64	0.3	22.7,±0.4	17.4,±0.6	25.8,±0.6	21.0,±0.5	23.9,±0.6	23.6,±0.5	23.2,±0.5
	64	1.0	21.0,±0.7	17.1,±0.6	20.6,±0.4	18.8,±0.5	21.7,±0.7	20.6,±0.5	20.9,±0.8
FEMNIST	128	0.1	16.9,±0.6	7.7,±0.4	28.0,±1.9	16.4,±0.7	17.4,±0.7	17.6,±0.8	16.8,±0.7
	128	0.3	17.7,±0.3	10.5,±0.3	23.4,±1.1	17.2,±0.2	18.9,±0.4	18.9,±0.4	18.3,±0.4
	128	1.0	17.7,±0.6	11.3,±0.3	18.2,±0.7	16.5,±0.6	18.7,±0.5	17.7,±0.5	17.9,±0.6
FEMNIST	256	0.1	12.8,±0.7	4.1,±0.1	24.5,±1.0	13.4,±0.7	13.9,±0.6	14.0,±0.6	13.5,±0.6
	256	0.3	14.7,±0.6	5.8,±0.2	19.9,±1.4	14.8,±0.8	15.6,±0.8	15.7,±0.7	15.1,±0.8
	256	1.0	14.1,±0.4	6.4,±0.3	15.8,±0.2	13.8,±0.2	14.9,±0.3	14.6,±0.3	14.3,±0.3

Table 13: Optimization check on FEMNIST features with 50 clients, 128 samples per client, and Dirichlet(0.3) label skew. Entries report excess test MSE ($\times 10^{-3}$), averaged over 5 seeds.

step α	passes	Indep	Local-LS	Local-ridge	IW-EXACT	CC-LOCAL	CC-CLASS
0.3	1	23.2,±0.9	127.5,±3.7	10.5,±0.3	17.2,±0.2	18.9,±0.4	18.3,±0.4
0.1	1	21.9,±0.6	127.5,±3.7	10.5,±0.3	20.0,±0.3	20.9,±0.4	20.8,±0.4
0.03	1	27.0,±0.8	127.5,±3.7	10.5,±0.3	26.4,±0.7	26.9,±0.7	26.8,±0.7
0.1	3	17.5,±0.4	9.8,±0.4	10.5,±0.3	15.4,±0.3	16.6,±0.3	16.1,±0.3
0.03	3	20.9,±0.5	9.8,±0.4	10.5,±0.3	20.2,±0.4	20.8,±0.5	20.7,±0.4

than the transfer-statistic methods across the tested settings, consistent with the main finding that collaboration provides limited benefit when the client-specific predictor can already be estimated accurately from local data.

The full Dirichlet grid reinforces the transition observed in the main experiments. On CIFAR-10 and CIFAR-100, class-conditional transfer is most useful when the class structure provides informative shared statistics. Under stronger skew, its advantage decreases and strong local estimators become competitive. On FEMNIST, LOCAL-RIDGE remains substantially better across the displayed sample sizes and concentration parameters, providing a negative regime in which the tested transfer estimators do not improve on strong local estimation.

Table 13 shows that the FEMNIST result is not reversed by the tested reductions in step size. With three passes, LOCAL-LS approaches LOCAL-RIDGE, while the SGD-based transfer methods remain worse than the local closed-form estimators. This supports treating FEMNIST as a negative transfer regime rather than attributing the result solely to the default step size.

H REPRODUCIBILITY AND HYPERPARAMETERS

H.1 SYNTHETIC BENCHMARK

Data generation. We follow the controlled linear-system construction of Zhang & Azizan (2026). We consider $n = 20$ clients with $d = 5$ and assign client i the sampling distribution $\mu_i = \mathcal{N}(m_i, I_d)$. Its personalized objective is defined through $b_i(s) = \Phi(s)\theta_i^\star$. The stochastic matrices are generated as

$$A(s) = \left(I_d + \epsilon_A s s^\top\right) A_{\text{base}}, \qquad \Phi(s) = \left(I_d + \epsilon_b s s^\top\right) \Phi_{\text{base}},$$

where A_{base} and Φ_{base} are randomly generated positive-definite matrices with $O(1)$ condition numbers. We use $\epsilon_A = 1$ and $\epsilon_b = 0.5$. Reference personalized solutions $x_i^\star$ are estimated using 5,000 Monte Carlo samples.

Sampling-distribution heterogeneity is controlled through

$$m_i = \delta_{\text{env}} C_A v_i,$$

where v_i is a randomly sampled unit vector and $C_A = 4$. Consequently, δ_{env} controls the separation among client sampling distributions. Objective heterogeneity is introduced independently through

$$\theta_i^\star = \theta_{\text{base}} + \delta_{\text{obj}} u_i, \qquad \theta_{\text{base}} \sim \mathcal{N}(0, I_d),$$

where u_i is a randomly sampled unit vector and δ_{obj} controls the magnitude of client-specific objective variation. The first client is used as a reference with $m_1 = 0$ and $\theta_1^\star = \theta_{\text{base}}$. Unless otherwise stated, the main controlled experiments vary δ_{env} and δ_{obj} jointly while holding the remaining construction fixed.

Optimization and evaluation. The main synthetic benchmark uses 10 independent evaluation runs. The constant-step experiment uses $\alpha = 0.01$ for $T = 2400$, while the decaying-step experiment uses $\alpha_t = 0.03/(1 + t/200)$ for $T = 4000$. We report Polyak–Ruppert averages over the second half of training. Pooling begins after 200 rounds and is disabled for $n < 4$. Dimension and communication experiments use the settings specified in their captions, including $\alpha = 0.005$ for $d = 20$. Diagnostic (n, δ) sweeps use three runs and are not used for the numerical claims in the main tables.

Baseline tuning. Personalized-FL baselines are tuned on three runs per setting using seeds disjoint from the ten evaluation runs. Ditto and pFedMe use $\lambda \in 1, 5, 15, 50$, All-for-one uses $\tau \in 0.01, 0.03, 0.1, 1$, and IFCA uses $C = 10$ clusters. Fine-tuning has no additional tuned hyperparameter. The selected configuration is fixed before the reported evaluation runs. IW-EXACT uses the exact client-to-population density ratios required by the importance-weighting method of Zhang & Azizan (2026), whereas IW-NONE applies the same transfer mechanism without density-ratio correction.

H.2 SCARCE-DATA REAL-FEATURE EXPERIMENTS

Features and client construction. The CIFAR-10 and CIFAR-100 experiments use frozen ImageNet ResNet-18 representations reduced by PCA to $d = 64$. FEMNIST uses frozen LeNet features with $d = 84$. The main scarce-data comparisons use 50 clients and five evaluation seeds. Each client receives $n_{\text{tr}} \in 64, 128, 256, 512$ training samples as specified in the corresponding table and $n_{\text{test}} = 1000$ test samples.

For Q2, clients have homogeneous sampling distributions. Q3 introduces structured label skew through the controlled and Dirichlet settings reported in Table 2. Q4 uses the more extreme CIFAR skews in Table 3 and additionally evaluates FEMNIST, including 100 real writers with their natural client distributions. Further details of the label-skew construction are given below.

Optimization and evaluation. Training consists of a single pass without replacement with batch size one, so $T = n_{\text{tr}}$ in the scarce-data experiments. Iterative methods use the common SGD step $\alpha = 0.3$. The collaboration gate is enabled for the transfer methods. Evaluation is performed

throughout training, and the final five evaluations are averaged. For iterative baselines, we use the better of the final and Polyak-averaged iterates; the closed-form local ridge solution is unaffected by iterate averaging. Pooling warm-up and refresh intervals are scaled to $T/10$ as described in Appendix G.

Local and personalized baselines. LOCAL-RIDGE independently selects a ridge coefficient for each client by 5-fold cross-validation over

$$\{10^{-4}, 10^{-3}, 10^{-2}, 0.03, 0.1, 0.3, 1, 3, 10\}.$$

On the well-posed synthetic system, LOCAL-LS instead denotes the unregularized least-squares solution. INDEP-AVG is independent constant-step SGD with Polyak–Ruppert averaging over the second half of training (Bach & Moulines, 2013). Hyperparameters for the personalized-FL baselines are selected using tuning seeds disjoint from the evaluation seeds.

Across the main tables, bold denotes the best deployable method. IW-EXACT is excluded from this designation because it is supplied with oracle density ratios. Additional metrics, partial-participation experiments, communication costs, and nonlinear head results are reported in the following appendices.