FEDBHG: BAYESIAN COALITION STABILITY IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

At a federated learning checkpoint, clients must choose between local and shared continuation before either alternative is trained. We study this decision through posterior gradient-estimation risk under a Gaussian hierarchy. With known scalar variances, clients whose sampling variance is below the between-client variance remain local; the rest pool using inverse-total-variance weights. Under the specified summary disclosure, we prove that this partition almost surely admits no coalition that weakly reduces every member's risk and strictly reduces at least one. For at least three pooling-eligible clients, these are the unique fixed positive linear weights that all members accept over staying local for every disclosed summary configuration. When every covariance direction favors the same local-or-shared choice, the construction extends to general covariances; when directions favor conflicting choices, every partition can be unstable. Under shared quadratic curvature, the posterior-risk comparison also determines the expected-loss comparison between the corresponding Newton updates. FedBHG estimates the variances at a checkpoint and continues with ordinary training. Synthetic experiments isolate failures from route errors, directional conflict and aggregation weights. Across 20 runs in four selected task/clusterer settings with fixed client partitions, FedBHG beats both universal continuations in accuracy in 19 and loss in 17. Four fresh-partition observations improve accuracy in all four and loss in three. These experiments support checkpoint-based selective continuation in the tested settings, with the Bayesian routing rule's incremental benefit over simpler routes varying by task. The theory concerns oracle posterior-risk stability; the neural experiments evaluate predictive transfer under fitted routing and ordinary training.

1 INTRODUCTION

A shared model can improve a federation's average accuracy while some clients would do better training alone. Consider clients that have reached a common training checkpoint. Each client must decide whether to continue locally or share a model with other clients, before the alternatives have been trained. Even if every client prefers one shared model to training alone, a smaller group may prefer to leave and train together. We ask when Bayesian uncertainty can rule out such jointly beneficial departures, and whether the resulting grouping helps choose how to continue federated training.

At the checkpoint, a client's mean gradient describes how its own population loss changes with the model parameters. The client estimates that gradient from finite local data. Sharing can reduce sampling error, but it also combines gradients from clients with different objectives. We model these two sources of variation separately: latent client gradients vary around a population mean, and each observed gradient is a noisy estimate of its client's latent gradient. A coalition uses one weighted average as its common gradient estimate.

The decision criterion is posterior risk: expected squared error in estimating a client's own mean gradient, conditional on the available summaries. When a group is proposed, each member compares its current and proposed estimates using the summaries from both groups. We call a partition *stable* if no proposed group can weakly lower every member's risk and strictly lower at least one. Thus stability concerns incentives to change groups; it does not mean that training has converged or that a fitted assignment resists perturbation.

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.

We call the resulting procedure *FedBHG*, for *Federated Bayesian Hedonic Game*. The name reflects participation decisions based on a client's own coalition and Bayesian beliefs about its gradient. Classical hedonic games assign each client a fixed ranking of coalitions. Here the information changes with the two groups being compared, so those comparisons can be nontransitive. We specify the disclosure rule and deviation test directly: one proposal is chosen independently of the data, and its members compare risks before any votes.

Bayesian FL already separates sampling uncertainty from variation between clients and uses inverse-variance averaging (Chen et al., 2022; Ozkara et al., 2023). Model-sharing games study local and common estimates using utilities averaged over data and client parameters (Donahue & Kleinberg, 2021a;b). We instead ask which groups remain acceptable after particular summaries have been observed, while requiring each group to use one common estimate. Information-dependent coalition objections also have classical predecessors (Wilson, 1978; Dutta & Vohra, 2005); our conclusions concern the specified disclosure and comparison rule.

With known scalar variances, our stable grouping follows a simple rule. Clients stay alone when their sampling variance is below the between-client variance; all other clients pool their estimates. Every group weights its summaries by the inverse of total variance, the sum of these two variances. We prove that, almost surely, no proposed group can lower one member's posterior risk without raising another member's. The proof first excludes proposals containing a client who prefers its local estimate. For a subgroup of the remaining pool, precision weighting makes the sum of the members' risk changes nonnegative. One member can therefore gain only if another loses.

The weights have a separate participation justification. Among positive averaging weights fixed before the summaries are observed, we ask which keep every member willing to pool instead of staying alone. For at least three pooled clients, requiring this for every set of summaries forces the same precision weights. General covariances introduce another issue: a client can favor sharing in one gradient direction and local estimation in another. The stable construction extends when every client's directions agree on one choice. A counterexample shows that conflicting directions can instead make every partition unstable; the positive condition is sufficient, not a complete classification.

To connect gradient estimation to learning, we first consider quadratic losses with shared curvature. Expected loss after a Newton update is determined by posterior gradient-estimation risk in the inverse-curvature metric. The scalar grouping remains stable in this common metric, so it also prevents jointly beneficial departures for those updates. FedBHG estimates checkpoint variances and trains the selected local and shared models with ordinary optimizers. Synthetic experiments isolate failures from routing errors, directional conflict and aggregation; neural experiments test predictive gains, simpler routes and client decisions. The fitted routing rule's incremental value over simpler coverage-matched routes varies across tasks, while client-level gains and harms remain heterogeneous even when aggregate performance improves. Section 7 summarizes the scope, and Appendix I collects the detailed statistical, training, and deployment limitations.

2 Bayesian participation at a common checkpoint

Summaries and common actions. Fix a checkpoint $\boldsymbol{w}^*$ and a finite population of M clients. Client identities, sample sizes and the oracle variance components below are known model inputs, held fixed when conditioning on observed gradients. Write $Y_i = \widehat{\boldsymbol{\theta}}_i \in \mathbb{R}^d$ for client i's sample-mean gradient and $v_i = \varepsilon_i^2/n_i > 0$ for its sampling variance per coordinate. For client population loss L_i, the latent mean gradient is $\boldsymbol{\theta}_i := \nabla L_i(\boldsymbol{w}^*)$. With known $\sigma^2 > 0$ and v_i, assume

$$\boldsymbol{\theta}_i \mid \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d), \qquad Y_i \mid \boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\theta}_i, v_i I_d), \tag{1}$$

independently across clients, with independent observation errors and the flat population-mean density $p(\boldsymbol{\mu}) \propto 1$. This hierarchy is assumed conditional on the checkpoint; training a checkpoint does not establish it. Every posterior conditions on a nonempty group and is proper. Probability-one statements use the proper sampling law for each fixed $\boldsymbol{\mu}$. Define

$$\alpha_i = (\sigma^2 + v_i)^{-1}, \qquad B_i = \frac{v_i}{\sigma^2 + v_i}, \qquad q_C = \boldsymbol{\vartheta}_C = \frac{\sum_{i \in C} \alpha_i Y_i}{\sum_{i \in C} \alpha_i}. \tag{2}$$

Here α_i is the inverse total variance, B_i is the fraction of total variance due to sampling, and q_C is coalition C's common gradient estimate. The index C denotes a nonempty client group. Every

member uses this same estimate; a singleton uses $q_{\{i\}} = Y_i$. A gradient estimate becomes a model update through the learning rule in Section 4. For a disclosed set $J \ni i$, standard Gaussian conditioning gives

$$m_i(J) := \mathbb{E}[\boldsymbol{\theta}_i \mid Y_J] = (1 - B_i)Y_i + B_i q_J. \tag{3}$$

Posterior squared-error risk is $R_i(a \mid J) = \mathbb{E}[\|a - \boldsymbol{\theta}_i\|_2^2 \mid Y_J]$. It equals $\|a - m_i(J)\|_2^2$ plus a variance term independent of a.

Information and consent. For proposed group S and incumbent T, client $i \in S \cap T$ compares

$$S \succeq_i T \quad \Longleftrightarrow \quad R_i(q_S \mid S \cup T) \le R_i(q_T \mid S \cup T). \tag{4}$$

Both alternatives use the same posterior. For a partition π, a proposal S blocks it when $S \succeq_i \pi(i)$ for every $i \in S$, with a strict inequality for at least one member. We call π stable when no proposal satisfies this test. An accepted unilateral departure is a special case: the mover strictly improves and the receiving members weakly accept. The group left behind has no veto. A stable partition need not be unique or welfare-optimal.

A single-round protocol interprets the information rule. The incumbent partition is fixed from the known model inputs. Client i initially receives $Y_{\pi(i)}$. One nonempty proposal, chosen independently of the population mean, latent gradients and summaries, is announced. Its members disclose their own summaries without forwarding other incumbent members' summaries. Before any ballot is observed, member i therefore knows $Y_{S \cup \pi(i)}$. Exogeneity makes the proposal announcement uninformative about the unknowns. Prescribed comparison ballots accept when everyone weakly improves and someone strictly improves.

A client that receives only these summaries has exactly the information set used in the comparison above. Its comparisons can be nontransitive because their conditioning sets change; they are prescribed comparisons rather than equilibrium voting strategies. Appendix B.3 formalizes the protocol; Appendices B.6 and B.7 show nontransitivity and the effect of broader disclosure. Appendix I.2 states the disclosure and information limits of this model.

3 Stable sharing: from scalar uncertainty to general covariance

The scalar variance comparison determines who joins the pool. We first prove that this pool resists subgroup departures, then ask which weights preserve participation and how the construction changes with covariance geometry.

3.1 The scalar switch

Lemma 3.1 (Local versus common estimate)**.** *For every nonempty $C \ni i$ under Equation* (1)*,*

$$R_i(q_C \mid C) - R_i(Y_i \mid C) = (1 - 2B_i)\|q_C - Y_i\|_2^2. \tag{5}$$

Thus, when the actions differ, pooling is preferred if $v_i > \sigma^2$, the local action if $v_i < \sigma^2$, and the client ties at equality.

Proof. The posterior mean lies at $Y_i + B_i(q_C - Y_i)$. Subtracting its squared distances to the two actions gives $[(1 - B_i)^2 - B_i^2]\|q_C - Y_i\|_2^2$; the posterior variance cancels. □

This midpoint identity is standard shrinkage. The no-deviation proof uses the relation $1 - B_i = \sigma^2\alpha_i$ and the common information retained by members leaving the same pool.

3.2 Stable singleton-or-pool sharing

Theorem 3.2 (Stable singleton-or-pool construction)**.** *Under Equations* (1)–(4)*, leave each client with $B_i < 1/2$ alone and pool $U = \{i : B_i \ge 1/2\}$, omitting an empty pool. For every fixed $\boldsymbol{\mu}$, this partition is stable with probability one.*

Proof. For a low-score singleton i, every non-singleton $S \ni i$ has $q_S \neq Y_i$ almost surely. There are finitely many such pairs, so Lemma 3.1 excludes every proposal containing a low-score client outside one null set. Any block must be $S \subseteq U$. Its members all condition on Y_U. Set $\Delta = q_S - q_U$ and let d_i be proposed-minus-incumbent risk. Then

$$d_i = \|\Delta\|_2^2 + 2(1 - B_i)\langle \Delta, q_U - Y_i \rangle,$$
$$\sum_{i \in S} d_i = \left[\sum_{i \in S} (2B_i - 1) \right] \|\Delta\|_2^2 \geq 0. \tag{6}$$

The second equality uses the defining precision-weighted mean of S. An all-weak improvement with one strict improvement would make the sum negative. Zero displacement and threshold clients produce no exception through ties. □

The construction needs one pass over the known variances and one weighted mean, with $O(Md)$ scalar arithmetic; it does not enumerate partitions. Theorem B.10 contains the full null-event argument.

3.3 Why these weights?

Estimating the population mean accurately does not itself ensure participation: each client cares about its own latent target. Fix a group C and consider a common linear action $q_p = \sum_{j \in C} p_j Y_j$, where $p_j > 0$, $\sum_j p_j = 1$, and p may depend on known variances but is fixed before Y_C is observed. Changing the action leaves the posterior in Equation (3) unchanged.

Theorem 3.3 (Participation characterizes fixed weights)**.** *Suppose $|C| \geq 3$ and $B_i \geq 1/2$ for every $i \in C$. Every member weakly prefers q_p to its local action for every realization of Y_C if and only if $p_j = \alpha_j / \sum_{k \in C} \alpha_k$ for all j.*

Proof idea. In one coordinate let $r_j = \alpha_j / \sum_k \alpha_k$ and let e_i be the ith unit vector. The risk difference is

$$((p - e_i)^\top y)^2 - 2B_i((p - e_i)^\top y)((r - e_i)^\top y).$$

If $p - e_i$ and $r - e_i$ are independent, choose y making the first inner product one and the second zero; participation fails. Thus $p - e_i = k_i(r - e_i)$. A third client index forces all k_i to agree, and the diagonal and off-diagonal coefficient equations then force $k_i = 1$. The converse is Lemma 3.1. The full proof treats all dimensions in Appendix B.5. □

For two clients an interval of weights can satisfy participation; with equal variances and $B_1 = B_2 = 3/4$, either member's weight can range from $1/4$ to $3/4$. Data-dependent common actions also escape the characterization. The theorem concerns uniform individual participation against leaving alone, while Theorem 3.2 separately excludes subgroup deviations.

3.4 General covariance

The scalar switch compares sampling and between-client variances along every coordinate in the same way. With general covariances, its sign becomes a semidefinite matrix condition, and the centroid proof extends in matrix form.

Replace $\sigma^2 I_d$ and $v_i I_d$ by known positive-definite matrices Σ and N_i, retaining independence and the flat population-mean density. Write $W_i = \Sigma + N_i$, and define

$$q_C = \left(\sum_{i \in C} W_i^{-1} \right)^{-1} \sum_{i \in C} W_i^{-1} Y_i, \qquad G_i = \tfrac{1}{2}(\Sigma W_i^{-1} + W_i^{-1}\Sigma). \tag{7}$$

For Euclidean posterior risk, Gaussian conditioning gives

$$R_i(q_C \mid C) - R_i(Y_i \mid C) = 2(q_C - Y_i)^\top (G_i - I_d/2)(q_C - Y_i).$$

The symmetric matrix determines whether all directions favor the same route.

Theorem 3.4 (Stable sharing when directions agree)**.** *Suppose every client satisfies either $G_i \succeq I_d/2$ or $G_i \preceq I_d/2$. Assign i to a singleton when $G_i \succeq I_d/2$ and $G_i \neq I_d/2$; pool the rest. Under the union-conditioned risk comparison, this partition is stable almost surely for every fixed $\boldsymbol{\mu}$.*

For a proposed subgroup S of the pool, the matrix centroid identity makes the sum of risk changes $\sum_{i \in S} \Delta^\top (I_d - 2G_i)\Delta \geq 0$. Singletons reject other proposals outside null subspaces, completing the proof (Theorem B.27). Common covariance shape is a special case: $\Sigma = \sigma^2 K$, $N_i = v_i K$ implies $G_i = (1 - B_i) I_d$, even for highly anisotropic K. The condition is sufficient, not necessary.

3.5 The boundary: conflicting directions

Theorem 3.5 (Conflicting directions can defeat every partition)**.** *Take three clients in two dimensions with $\Sigma = I_2$, $N_i = \mathrm{diag}(1/100, 100)$ for all i, and observations $Y_1 = (3, 1)^\top$, $Y_2 = (4, -3)^\top$, $Y_3 = (0, 4)^\top$. Under Equation* (7) *and union-conditioned comparisons, every partition admits an accepted unilateral deviation. The required gains are strict and persist on an open set of positive sampling probability for every fixed $\boldsymbol{\mu}$.*

Coordinate 1 favors the local observation while coordinate 2 favors pooling. Writing 12 for $\{1, 2\}$, accepted moves are: client 1 leaves 123; client 2 joins 1 from $1 \mid 23$; client 3 joins 12 from $12 \mid 3$; client 1 joins 2 from $13 \mid 2$ or $1 \mid 2 \mid 3$. The ten mover/receiver comparisons in Table 1 have minimum gain $203/1212 > 0$; continuity and full Gaussian support establish the open-set assertion (Proposition B.30). Anisotropy alone therefore does not preclude stable sharing: the shared-shape family above can be arbitrarily anisotropic. Disagreement across directions can defeat stability. The sufficient condition and counterexample leave intermediate covariance families unresolved.

Table 1: Exact verification of an accepted deviation from every partition. A vertical bar separates incumbent groups. The first row of each block identifies the departing client; subsequent rows verify the receiving members. Positive gain means strictly smaller posterior risk.

Partition	Role	Client i	$T \to S$	Observed J	Gain g_i
123	Leaves alone	1	$123 \to 1$	123	$33/101$
$1 \mid 23$	Joins 1	2	$23 \to 12$	123	$107/101$
	Accepts 2	1	$1 \to 12$	12	$1485/404$
$12 \mid 3$	Joins 12	3	$3 \to 123$	123	$561/101$
	Accepts 3	1	$12 \to 123$	123	$3163/1212$
	Accepts 3	2	$12 \to 123$	123	$203/1212$
$13 \mid 2$	Joins 2	1	$13 \to 12$	123	$3071/1212$
	Accepts 1	2	$2 \to 12$	12	$1485/404$
$1 \mid 2 \mid 3$	Joins 2	1	$1 \to 12$	12	$1485/404$
	Accepts 1	2	$2 \to 12$	12	$1485/404$

4 FedBHG: estimating and training the construction

The theory chooses a gradient estimate, whereas a federated system must choose how to train a model. FedBHG connects the two by estimating uncertainty once at a common checkpoint and using the scalar switch to choose each client's continuation. We first explain the Newton update that justifies this connection under quadratic losses, then specify the implemented training procedure.

From a gradient estimate to a Newton update. Suppose every client has the same deterministic quadratic curvature $H \succ 0$ at every compared action: $L_i(\boldsymbol{w}^* + u) = L_i(\boldsymbol{w}^*) + \boldsymbol{\theta}_i^\top u + u^\top H u/2$. Replacing the unknown gradient by an estimate a and minimizing this quadratic gives $u = -H^{-1}a$, the Newton step. Thus the local estimate Y_i selects $\boldsymbol{w}(Y_i) = \boldsymbol{w}^* - H^{-1}Y_i$, while the shared estimate q_C selects $\boldsymbol{w}(q_C) = \boldsymbol{w}^* - H^{-1}q_C$. A proposed change of coalition changes which shared estimate a client would use.

Proposition 4.1 (Expected-loss participation)**.** *Under the scalar hierarchy and shared quadratic geometry, the local/common conditional loss difference is*

$$\mathbb{E}[L_i(\boldsymbol{w}(q_C)) - L_i(\boldsymbol{w}(Y_i)) \mid Y_C] = (1/2 - B_i)\|q_C - Y_i\|^2_{H^{-1}}. \tag{8}$$

The singleton-or-pool partition is stable under union-conditioned expected-loss comparisons, almost surely for every fixed $\boldsymbol{\mu}$.

Substitution expresses loss as an action-independent term plus $\|a - \boldsymbol{\theta}_i\|^2_{H^{-1}}/2$. The scalar centroid proof then works in that common metric (Theorem C.6). The proposition uses the scalar hierarchy; the covariance theorem above uses Euclidean risk.

Estimate once, route once. An upstream pooled or clustered method produces the checkpoint. Within each base cluster, FedBHG executes three phases.

1. Each client computes Y_i and the sum of squared per-example gradient deviations, SSE_i, at the identical checkpoint, and uploads them with n_i.
2. The server estimates $\widehat{\varepsilon}_i^2 = \mathsf{SSE}_i/[d(n_i-1)]$, fits $\widehat{\sigma}^2$ by deterministic approximate profile-likelihood maximization, and forms $\widehat{B}_i$ and $\widehat{\alpha}_i$. It routes $\widehat{B}_i < 1/2$ locally and the rest jointly; equality is an implementation tie convention. Membership and weights are frozen.
3. Local clients restart independent optimizers from $\boldsymbol{w}^*$. Joint clients restart there and train one common model using normalized $\widehat{\alpha}_i$. These continuations use SGD or Adam and federated averaging; they do not compute H^{-1} or take the prescribed Newton step.

Appendix D.3 specifies the algorithm and edge cases. Summary communication is $O(d)$ values per client; the current server uses $O(Md + M^2)$ storage and $O(M^2d + GM^2)$ work for G profile evaluations.

Estimation can change participation. High dimension alone does not correct the fitted route. With M fixed, known common sampling variance v and between-client variance t, the profile target converges as $d \to \infty$ to $\widehat{t}_{\mathrm{ML}} \to [(M-1)(t+v)/M - v]_+$. For $M = 6, t = 1.2v$, the true score is $5/11 < 1/2$ but the fitted score tends to $6/11 > 1/2$. Appendix E.2.1 derives the finite-dimensional error probability and a degrees-of-freedom diagnostic. This failure motivates testing the fitted route directly. The experiments use the stated profile algorithm; the degrees-of-freedom correction is a diagnostic comparison.

5 Empirical evaluation

FedBHG departs from the oracle construction by estimating routes and weights and replacing quadratic Newton actions with multi-round SGD or Adam. Synthetic experiments test posterior-risk deviations under estimation and covariance mismatch; neural experiments test prediction and client decisions.

Claim boundary. The neural experiments establish neither strategic stability of trained models nor validity of the Gaussian hierarchy. Repeated-seed intervals describe optimizer variation on selected fixed client populations; fresh partitions are a separate robustness check.

5.1 When does no-deviation survive approximation?

Six-client simulations check all 63 nonempty coalitions to assess route recovery, directional agreement and aggregation weights. **Route estimation error.** Larger threshold margins and more independent coordinates improve recovery (Figure 1). Across 3,750 fitted trials, 93.0% with correct routes admit no block, versus 14.5% after one error and below 1% after two or more. Among the fitted routes and weights tested here, route error produces the larger drop in no-deviation rate.

Directional conflict. All 1,500 direction-consistent trials admit no block, including non-shared covariance shapes. When directions cross the local/joint boundary, rates fall to 21.0%, 2.4% and 1.0% in three conflict settings. Anisotropy alone is not the obstruction in this family.

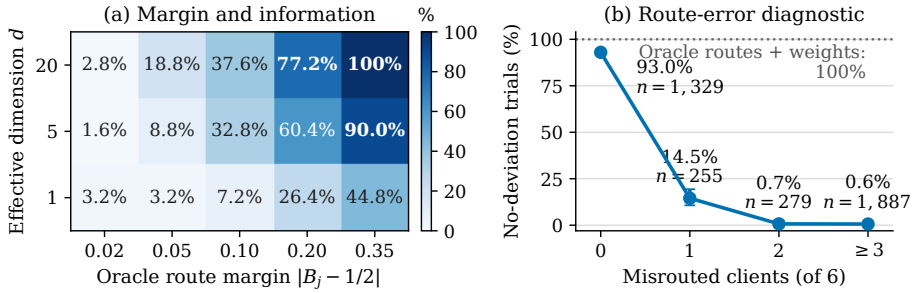

Figure 1: Route recovery and no-deviation under fitted routing. Each six-client trial checks all 63 coalitions. **(a)** No-deviation rates over 250 trials per margin–dimension cell. **(b)** Rates by routing-error count, pooled over the balanced 3,750-trial grid; bars are Wilson 95% intervals and labels are trial counts. The dotted line is the oracle-route/oracle-weight reference. Panel (b) is descriptive, not a randomized causal comparison.

Aggregation-weight error. With oracle routes, sample-count weighting averages 24.61% no-deviation versus 100% for precision weighting. Correct membership alone is insufficient. Appendix F reports all three failure modes, deviation magnitudes and protocols.

5.2 Does selective continuation beat the universal choices?

From one checkpoint, *local-all* trains clients separately, *joint-all* pools each full upstream cluster with precision weights, and *FedBHG* routes low-score clients locally and pools the rest. We use 100 clients on Fashion-MNIST, EMNIST-Letters and natural-client Shakespeare, 200-round CFL or IFCA checkpoints and ten continuation rounds with full participation. Policies share the checkpoint, optimizer and per-client budget; each distinct pool trains its own model. Co-primary outcomes are sample-weighted raw accuracy and client-uniform held-out loss.

The development sweep improves checkpoint accuracy in 10/10 cells and beats the better universal action in accuracy in 8/10, loss in 7/10 and both in 5/10. Both Fashion S4 cells gain accuracy while increasing loss. Four cells were selected afterward; the sweep describes breadth (Appendix G.7).

Optimizer variation. Five training seeds per selected cell keep client populations and partitions fixed. Across 20 runs, FedBHG improves checkpoint accuracy in 20/20, beats both universal actions in accuracy in 19/20 and loss in 17/20, and improves both in 16/20. Fashion S3/CFL and Shakespeare/IFCA have positive paired intervals on both metrics. EMNIST S3/IFCA has unresolved loss improvement; Shakespeare/CFL has unresolved accuracy improvement (Table 2).

New client populations. Two fresh data seeds per selected vision cell change counts, class subsets and corruption, producing new checkpoints and routes. Accuracy beats both universal choices in 4/4 observations and loss in 3/4. The remaining Fashion partition gains 1.45 accuracy points while increasing loss by .013, so accuracy is more consistent than loss across these four fresh-partition observations.

5.3 Does the full score improve on simpler routes?

Coverage-matched controls identify where the full score adds to selective continuation. We match local coverage within each cluster using sample size, per-example gradient variance, or three fixed random routes. FedBHG beats the mean random-route performance on both outcomes in all 20 repeats. Its membership matches sample-size routing in 7/20 and variance-only routing in 5/20.

Fashion S3/CFL differs from both controls in every seed, gaining .925 accuracy points over sample size and 1.118 over variance-only. Loss falls by .0415 against variance-only; the .0020 reduction against sample size is unresolved. EMNIST equals variance-only in 5/5 seeds; both Shakespeare cells are effectively tied with sample size. Fresh Fashion routes differ from both controls; fresh EMNIST routes equal both.

Table 2: Predictive gains over the better universal continuation for each metric; positive values favor FedBHG. Repeated rows: paired means [descriptive 95% t intervals] over five optimizer seeds on a selected fixed partition. Fresh rows: individual new-partition outcomes. F/E/Sh denote Fashion-MNIST/EMNIST-Letters/ Shakespeare; S3 has independently varying sample counts and corruption.

Design	Cell / data seed	Accuracy gain (pp)	Loss reduction
Repeated	F-S3/CFL	$+2.944$ $[2.664, 3.224]$	$+.0089$ $[.0027, .0151]$
	E-S3/IFCA	$+.782$ $[.273, 1.292]$	$+.0071$ $[-.0482, .0624]$
	Sh/CFL	$+.051$ $[-.010, .113]$	$+.0048$ $[.0015, .0081]$
	Sh/IFCA	$+.175$ $[.150, .200]$	$+.0068$ $[.0048, .0088]$
Fresh	F-S3/CFL / 12346	$+1.45$	$-.013$
	F-S3/CFL / 12347	$+1.81$	$+.011$
	E-S3/IFCA / 12346	$+1.03$	$+.010$
	E-S3/IFCA / 12347	$+1.60$	$+.078$

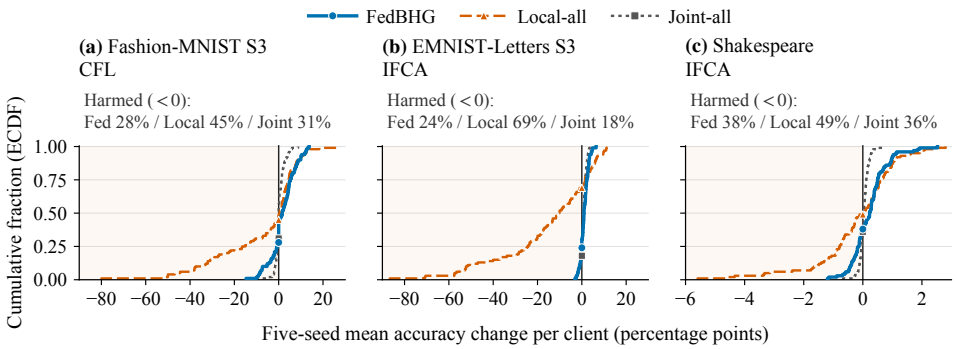

Figure 2: Who is harmed relative to the checkpoint? Accuracy changes after ten rounds are averaged over five optimizer seeds per client; each CDF contains 100 clients from one fixed partition. Negative values indicate harm. Panel (c) uses a narrower horizontal scale. Annotations report strictly harmed fractions, excluding ties, in legend order. These fractions concern seed-averaged clients, whereas the text reports mean run-level harm.

Within a cluster with positive fitted between-client variance, score order equals effective-sampling-variance order. The Bayesian formula sets the cutoff and weights. Matched routes share the predictive gains of selective continuation; additional accuracy gains from the full score are clearest on Fashion S3/CFL.

5.4 Does the score make useful client-level decisions?

Client-level diagnostics relate the score to the benefit of sharing and measure gains and harm relative to the checkpoint. At ten rounds, $\widehat{B}_j$ has mean run-level Spearman association .552 with local-all minus joint-all loss (cell means .301–.758). This measures ranking, not probability calibration, using the full upstream joint-all cluster.

Half-threshold agreement with the lower-loss universal branch increases from 57.5% at round one to 68.9% at round five and 69.3% at round ten. Ten-round cell means range from 57.6% on Shakespeare/CFL to 74.8% on Shakespeare/IFCA. This agreement concerns the universal branches, before selection of the smaller FedBHG pool.

FedBHG reduces checkpoint-relative accuracy harm in every repeated run, with mean harm of 32.2% versus 52.9% for local-all; loss harm improves in 15/20. Figure 2 shows less severe negative tails than local-all. Joint-all has slightly lower strict harm in the displayed EMNIST and Shakespeare cells. Fashion S3/CFL also has higher mean loss harm than local-all despite lower aggregate loss.

5.5 What survives model misspecification?

All five shape checks reject the fitted spherical model in every large cluster from the four fresh cells ($p = 1/201$). Effective rank is 12.2–71.6 versus fitted-null medians of 95.9–99.0. The score nevertheless retains positive association with local-minus-joint loss in all four cells ($\rho = .438$–$.787$). Average route-flip probabilities are 1.39–3.47% under client bootstrap refits and .01–.16% under leave-one-out refits. Sensitivity concentrates near the threshold: the most fragile client in each cell flips in 42.8–51.4% of bootstrap refits. Resampling measures repeatability of the fitted rule, not oracle-route recovery. Table 7 reports the full checks.

5.6 Competitiveness against personalization baselines

These comparisons assess end-to-end competitiveness at a matched communication horizon; the common-checkpoint comparisons isolate the continuation decision. At 210 rounds, FedBHG exceeds the reported Ditto and pFedMe implementations on both outcomes in every aligned repeated seed, and Self-FL and pFedGraph in all 20 aligned vision comparisons (Table 8). Trajectories, optimizer work, objectives and tuning differ; Self-FL and pFedGraph use small validation grids on fixed vision partitions.

5.7 What does the routing decision cost?

Cell-mean summary collection and fitting take 8.16–33.66 seconds in the 100-client RTX 5090 repeats, or 0.6–2.7% of upstream training time. Incremental communication relative to ten-round joint-all falls 41.5% on Fashion S3/CFL, 44.5% on Shakespeare/IFCA and 3.2% on EMNIST S3/IFCA. The summary upload offsets most savings when nearly everyone stays joint. Deployment trains only selected branches; Appendix G.6 separates additional counterfactual evaluation branches.

6 Related work

Normal random-effects inference has the same scalar posterior formulas (Röver, 2020); Self-FL and statistical personalized learning already use hierarchical uncertainty and shrinkage (Chen et al., 2022; Ozkara et al., 2023). For scalar squared error, optimal model averaging characterizes when the shared endpoint outperforms the local one by whether the optimal interpolation weight exceeds one half (Grimberg et al., 2021). The fixed-weight theorem characterizes participation in a common action, not a new posterior estimator.

Donahue–Kleinberg analyze voluntary federation, the corresponding equal-noise threshold and pool/local structures (Donahue & Kleinberg, 2021a;b). Their personalized weights coincide with the scalar posterior coefficients on the overlapping variance model. Their ex-ante utilities and personalized actions do not imply our realized-summary common-action deviation theorem. Information-sensitive objections have longstanding core formulations (Wilson, 1978; Dutta & Vohra, 2005); our exogenous single proposal excludes inference from strategic objections. Core-stable data exchange instead uses fractional bilateral transfers and sharing costs (Song et al., 2025).

7 Discussion and conclusion

With known variances and the specified disclosure, the theory identifies a stable singleton-or-pool partition, characterizes precision weights among fixed positive linear weights for at least three pooling-eligible clients, and extends the construction to general covariances under direction agreement. Conflicting directions can instead make every partition unstable.

The experiments trace this construction through estimated routing and ordinary training. Synthetic studies identify the roles of route recovery, directional agreement and aggregation. Neural studies show frequent predictive gains over universal continuations, with the clearest score-specific accuracy gains on Fashion S3/CFL and informative client-level decisions under a misspecified working model. Appendix I details the statistical, training and deployment limitations. Extending participation guarantees to estimated routes and multi-round training is an open direction.

AI USE STATEMENT

The authors developed FedBHG's research question, Bayesian model, participation claims, proof strategies, and counterexamples. Experimental designs, simulation distributions and parameter choices, and scientific interpretations were also determined by the authors. Generative AI was not used to develop the theoretical model or conceptual frameworks, formulate mathematical claims, contribute critical proof ingredients, propose or refine hypotheses, design or provide feedback on methodology or experiments, or interpret the results.

AI assistance supported the presentation and implementation of that work. It helped write and organize proofs from the authors' existing arguments, improve their exposition, and check for possible gaps or flaws. The authors evaluated those checks and devised and verified any corrections. AI also assisted method implementation, dataset cleaning and reformatting, scientific figure editing, software coding, readability editing, and reference formatting. Synthetic observations were produced by seeded numerical software using distributions specified by the authors; generative AI did not generate these datasets. Translation and qualitative or thematic data analysis were not applicable.

The authors reviewed all AI-assisted material, checked proof exposition against their reasoning, tested code and implementations, and verified numerical outputs and figures against retained experimental records. They take responsibility for the final paper and its text, claims, code, figures, and other accompanying artifacts.

ETHICS STATEMENT

The experiments use established image and natural-client language benchmarks and do not report a deployment involving human subjects. The voluntary-sharing result is a model-based posterior-risk statement, not a fairness, legal, or regulatory guarantee. FedBHG transmits a client mean gradient and dispersion statistic; these summaries can reveal information about local objectives and are not privacy neutral. Privacy noise would alter the variance estimates near the sensitive half-threshold and requires a separate calibration analysis.

REPRODUCIBILITY STATEMENT

The main text states the theorem assumptions, proof mechanism, evaluation selection, and claim boundaries. The appendix contains the complete proofs, estimator derivation, synthetic protocols, empirical protocol, all reported result tables, limitations, and experimental-record analysis. The supplementary material contains the code needed to reproduce the experiments.

REFERENCES

Wenxuan Bao, Haohan Wang, Jun Wu, and Jingrui He. Optimizing the collaboration structure in cross-silo federated learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1718–1736. PMLR, 2023. URL https://proceedings.mlr.press/v202/bao23b.html.

Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1005–1014. PMLR, 2021. URL http://proceedings.mlr.press/v139/blum21a.html.

Anna Bogomolnaia and Matthew O. Jackson. The stability of hedonic coalition structures. *Games Econ. Behav.*, 38(2):201–230, 2002. doi: 10.1006/GAME.2001.0877. URL https://doi.org/10.1006/game.2001.0877.

Jae Ho Chang, Arnab Auddy, and Subhadeep Paul. Personalized federated learning via variance-aware nonparametric empirical bayes, 2026. URL https://arxiv.org/abs/2608.09074.

Bhaskar Ray Chaudhury, Linyi Li, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/25e92e33ac8c35fd49f394c37f21b6da-Abstract-Conference.html.

Huili Chen, Jie Ding, Eric W. Tramel, Shuang Wu, Anit Kumar Sahu, Salman Avestimehr, and Tao Zhang. Self-aware personalized federated learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8265d7bb2db42e86637001db2c46619f-Abstract-Conference.html.

Mengmeng Chen, Xiaohu Wu, Xiaoli Tang, Tiantian He, Yew Soon Ong, Qiqi Liu, Qicheng Lao, and Han Yu. Free-rider and conflict aware collaboration formation for cross-silo federated learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/62ffefbe8dfc8548d22564f3c1d21488-Abstract-Conference.html.

Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J. Su. Minimax estimation for personalized federated learning: An alternative between fedavg and local training? *J. Mach. Learn. Res.*, 24: 262:1–262:59, 2023. URL http://jmlr.org/papers/v24/21-0224.html.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pp. 2921–2926. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966217. URL https://doi.org/10.1109/IJCNN.2017.7966217.

Sen Cui, Jian Liang, Weishen Pan, Kun Chen, Changshui Zhang, and Fei Wang. Collaboration equilibrium in federated learning. In Aidong Zhang and Huzefa Rangwala (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 241–251. ACM, 2022. doi: 10.1145/3534678.3539237. URL https://doi.org/10.1145/3534678.3539237.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986. ISSN 0197-2456. doi: https://doi.org/10.1016/0197-2456(86)90046-2. URL https://www.sciencedirect.com/science/article/pii/0197245686900462.

Canh T. Dinh, Nguyen Hoang Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f4f1f13c8289ac1b1ee0ff176b56fc60-Abstract.html.

Kate Donahue and Jon M. Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event,*

February 2-9, 2021, pp. 5303–5311. AAAI Press, 2021a. doi: 10.1609/AAAI.V35I6.16669. URL https://doi.org/10.1609/aaai.v35i6.16669.

Kate Donahue and Jon M. Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 1287–1298, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/09a5e2a11bea20817477e0b1dfe2cc21-Abstract.html.

Kate Donahue and Jon M. Kleinberg. Fairness in model-sharing games. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (eds.), *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pp. 3775–3783. ACM, 2023. doi: 10.1145/3543507.3583483. URL https://doi.org/10.1145/3543507.3583483.

Bhaskar Dutta and Rajiv Vohra. Incomplete information, credibility and the core. *Math. Soc. Sci.*, 50 (2):148–165, 2005. doi: 10.1016/J.MATHSOCSCI.2005.02.006. URL https://doi.org/10.1016/j.mathsocsci.2005.02.006.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/01cea7793f3c68af2e4989fc66bf8fb0-Abstract-Conference.html.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013. doi: 10.1201/b16018. URL https://doi.org/10.1201/b16018.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *IEEE Trans. Inf. Theory*, 68(12):8076–8091, 2022. doi: 10.1109/TIT.2022.3192506. URL https://doi.org/10.1109/TIT.2022.3192506.

Yan Gong, Bing Luo, Jiawei Jiang, Siping Shi, Chuang Hu, and Dazhao Cheng. Toward personalized federated learning via overlapping coalition formation game. *IEEE Trans. Mob. Comput.*, 25(8): 11790–11803, 2026. doi: 10.1109/TMC.2026.3668791. URL https://doi.org/10.1109/TMC.2026.3668791.

Felix Grimberg, Mary-Anne Hartley, Sai Praneeth Karimireddy, and Martin Jaggi. Optimal model averaging: Towards personalized collaborative learning. *CoRR*, abs/2110.12946, 2021. URL https://arxiv.org/abs/2110.12946.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *CoRR*, abs/1812.04754, 2018. URL http://arxiv.org/abs/1812.04754.

Cengis Hasan. Stable and budget-feasible coalition formation for clustered federated learning: A hedonic potential-game approach. *CoRR*, abs/2607.26788, 2026. doi: 10.48550/ARXIV.2607.26788. URL https://doi.org/10.48550/arXiv.2607.26788.

Salma Kharrat, Marco Canini, and Samuel Horváth. DPFL: decentralized personalized federated learning. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025*, volume 258 of *Proceedings of Machine Learning Research*, pp. 5086–5094. PMLR, 2025. URL https://proceedings.mlr.press/v258/kharrat25a.html.

Minyoung Kim and Timothy M. Hospedales. Fedhb: Hierarchical bayesian federated learning. *J. Mach. Learn. Res.*, 26:272:1–272:50, 2025. URL https://jmlr.org/papers/v26/23-1350.html.

Nikita Kotelevskii, Maxime Vono, Alain Durmus, and Eric Moulines. Fedpop: A bayesian approach for personalised federated learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/39540967927059 1fd2a70abc694cf5a1-Abstract-Conference.html.

Nan M Laird. Computation of variance components using the em algorithm. *Journal of Statistical Computation and Simulation*, 14(3-4):295–303, 1982.

Juan Li, Yanmin Zhu, Jie Wu, Weifan Wu, Tianzi Zang, and Liu Lu. All federated or not: Optimizing personal model performance in cross-silo federated learning. In *30th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2024, Belgrade, Serbia, October 10-14, 2024*, pp. 390–399. IEEE, 2024. doi: 10.1109/ICPADS63350.2024.00058. URL https://doi.org/10.1109/ICPADS63350.2024.00058.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021. URL http://proceedings.mlr.press/v139/li21h.html.

Jianfeng Lu, Yue Chen, Shuqin Cao, Longbiao Chen, Wei Wang, and Yun Xin. LEAP: optimization hierarchical federated learning on non-IID data with coalition formation game. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 4660–4668. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/515.

Disha Makhija, Joydeep Ghosh, and Nhat Ho. A bayesian approach for personalized federated learning in heterogeneous settings. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b9a159b81a0535e3944ffdbd07ae3ba1-Abstract-Conference.html.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 15434–15447, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/82599a4ec94aca066873c99b4c741ed8-Abstract.html.

Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2012. ISBN 9780262018029 0262018020.

Masato Ota, Yuko Sakurai, and Satoshi Oyama. Coalitional game theoretic federated learning. In *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2022, Niagara Falls, ON, Canada, November 17-20, 2022*, pp. 48–55. IEEE, 2022. doi: 10.1109/WI-IAT55865.2022.00017. URL https://doi.org/10.1109/WI-IAT55865.2022.00017.

Kaan Ozkara, Antonious M. Girgis, Deepesh Data, and Suhas N. Diggavi. A statistical framework for personalized federated learning and estimation: Theory, algorithms, and privacy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=FUiDMCr_W4o.

Jonatan Reyes, Lisa Di-Jorio, Cécile Low-Kam, and Marta Kersten-Oertel. Precision-weighted federated learning. *Comput. Intell.*, 41(6), 2025. doi: 10.1111/COIN.70150. URL https://doi.org/10.1111/coin.70150.

Christian Röver. Bayesian random-effects meta-analysis using the bayesmeta r package. *Journal of Statistical Software*, 93(6):1–51, 2020. doi: 10.18637/jss.v093.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v093i06.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJO1_M0Lf.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Networks Learn. Syst.*, 32(8):3710–3722, 2021. doi: 10.1109/TNNLS.2020.3015958. URL https://doi.org/10.1109/TNNLS.2020.3015958.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1992.

Jiaxin Song, Pooja Kulkarni, Parnian Shahkar, and Bhaskar Ray Chaudhury. On the existence and complexity of core-stable data exchanges. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/28870007047360c4cc200bc44fc41de1-Abstract-Conference.html.

Aris Spanos. Revisiting the neyman-scott model: an inconsistent mle or an ill-defined model?, 2013. URL https://arxiv.org/abs/1301.6278.

Yuchang Sun, Marios Kountouris, and Jun Zhang. How to collaborate: Towards maximizing the generalization performance in cross-silo federated learning. *IEEE Trans. Mob. Comput.*, 24(4):3211–3222, 2025. doi: 10.1109/TMC.2024.3509852. URL https://doi.org/10.1109/TMC.2024.3509852.

Nazarii Tupitsa, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Selective collaboration for robust federated learning. In Rebekka Burkholz, Shiwei Liu, Saiprasad Ravishankar, William Redman, Wei Huang, Weijie Su, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pp. 1161–1194. PMLR, 23–26 Mar 2026. URL https://proceedings.mlr.press/v328/tupitsa26a.html.

Robert Wilson. Information, efficiency, and the core of an economy. *Econometrica*, 46(4):807–816, 1978. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1909750.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983. doi: 10.1214/aos/1176346060. URL https://doi.org/10.1214/aos/1176346060.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39801–39817. PMLR, 2023. URL https://proceedings.mlr.press/v202/ye23b.html.

Abbas Zal, Thomas Marchioro, and Leonardo Badia. Evaluating coalition stability in federated learning under voluntary client participation. In *IEEE International Mediterranean Conference on Communications and Networking, MeditCom 2025, Nice, France, July 7-10, 2025*, pp. 1–6. IEEE, 2025. doi: 10.1109/MEDITCOM64437.2025.11104301. URL https://doi.org/10.1109/MeditCom64437.2025.11104301.

Abbas Zal, Alessandro Buratto, Thomas Marchioro, and Leonardo Badia. Costs and incentives for data owners to participate in federated learning seen through game theory. In *IEEE Wireless Communications and Networking Conference, WCNC 2026, Kuala Lumpur, Malaysia, April 13-16, 2026*, pp. 1–6. IEEE, 2026. doi: 10.1109/WCNC65185.2026.11555251. URL https://doi.org/10.1109/WCNC65185.2026.11555251.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and José M. Álvarez. Personalized federated learning with first order model optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ehJqJQk9cw.

Yuqing Zhang, Changli Zhou, Binghuang Huang, and Hui Tian. Fedgain: Toward negative-gain-free client collaboration in federated learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yLn1COP0AN.

Junyi Zhu, Xingchen Ma, and Matthew B. Blaschko. Confidence-aware personalized federated learning via variational expectation maximization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24542–24551. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02351. URL https://doi.org/10.1109/CVPR52729.2023.02351.

A Bayesian Participation at a Common Checkpoint

Clients compare acting on their own estimate with using one common estimate shared by a coalition. Their objectives can differ, and each estimate is noisy. We first study posterior squared-error risk for this decision. Sections B and B.11 establish the participation construction and its covariance boundary. Section C then translates the scalar construction into expected loss under shared quadratic learning geometry. This order separates the structure of the sharing decision from the assumptions needed to interpret an estimate as a trained-model action.

A coalition deploys one common estimate. This describes common-model or clustered deployment, where membership selects the model a client uses. A client allowed its own posterior-optimal estimate faces a different action problem. The information interface is also explicit: a participant evaluates its incumbent and one proposed coalition using their disclosed summaries. It can be interpreted as a participation assessment delegated to a component that receives only those summaries and known model inputs (checkpoint, client identities, sample sizes and oracle variance components). An institution using additional informative private data does not satisfy this observation model without a sufficiency argument. We analyze the stated comparisons rather than strategic reporting.

We use the following notation throughout. Let $\mathcal{M}$ be a finite set of M clients and fix a common checkpoint $\boldsymbol{w}^*$. For client $j \in \mathcal{M}$, let $\boldsymbol{\theta}_j \in \mathbb{R}^d$ denote the latent client mean update gradient, $\hat{\boldsymbol{\theta}}_j \in \mathbb{R}^d$ its observed sample-mean gradient, n_j the effective local sample size, $\varepsilon_j^2 > 0$ the within-client gradient variance, and $\sigma^2 > 0$ the between-client mean-gradient variance. We write

$$\alpha_j := \frac{1}{\sigma^2 + \varepsilon_j^2/n_j} \qquad \text{(precision weight of player } j\text{)},$$

$$B_j := \frac{\varepsilon_j^2/n_j}{\varepsilon_j^2/n_j + \sigma^2} \qquad \text{(shrinkage coefficient of player } j\text{)},$$

$$q_C = \boldsymbol{\vartheta}_C := \frac{\sum_{i \in C} \alpha_i \hat{\boldsymbol{\theta}}_i}{\sum_{i \in C} \alpha_i} \qquad \text{(variance-adaptive coalition mean for coalition } C\text{)}.$$

Here q_C denotes the common gradient estimate used by every member of a nonempty coalition C; $q_{\{j\}} = \hat{\boldsymbol{\theta}}_j$ for a singleton. It is an estimate in gradient space, converted to a model update by a learning rule. We call a partition *stable* if no proposed coalition weakly reduces every proposed member's posterior risk and strictly reduces at least one's, using the disclosed-summary comparison defined below. This concerns incentives to change coalitions, not optimization convergence or robustness to refitting. Note that $B_j \in (0, 1)$ and $1 - B_j = \sigma^2 \alpha_j$.

A.1 Client Mean Gradients and the Reliability Score

The structural theorems use the following model for summaries at one common checkpoint.

Assumption A.1 (Hierarchical Mean-Gradient Model)**.** *Fix a checkpoint $\boldsymbol{w}^* \in \mathbb{R}^d$. We work in $\mathbb{R}^d$ and assume the gradient coordinates are independent under the scalar model. The hierarchy is assumed conditional on the checkpoint, client identities, sample sizes and known variance components. It is not derived from the preceding training trajectory; fixing a learned checkpoint alone does not justify this conditional Gaussian law. The hierarchy has three levels.*

- *(a)* ***Population level.*** *An unknown population mean update $\boldsymbol{\mu} \in \mathbb{R}^d$ is assigned the improper flat prior with Lebesgue density $\pi(\boldsymbol{\mu}) \propto 1$. This is the standard Jeffreys prior for a Gaussian location parameter; it yields a proper posterior as soon as $|C| \geq 1$. The notation does not refer to a proper uniform distribution on $\mathbb{R}^d$.*

- *(b)* ***Client level.*** *A known variance $\sigma^2 > 0$ specifies the dispersion of client mean updates. For each client $j \in \mathcal{M}$, the latent population gradient $\boldsymbol{\theta}_j := \nabla L_j(\boldsymbol{w}^*)$ is drawn independently from $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$.*

- *(c)* ***Gradient-summary level.*** *Let $g_{jk}(\boldsymbol{w}^*)$ be client j's conditionally independent update gradients for local examples k, with conditional mean $\boldsymbol{\theta}_j$ and trace-normalized covariance*

$\varepsilon_j^2 I_d$. *For*

$$\hat{\boldsymbol{\theta}}_j := \frac{1}{n_j} \sum_{k=1}^{n_j} g_{jk}(\boldsymbol{w}^*),$$

the model assumes $\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, (\varepsilon_j^2/n_j) I_d)$. *This is an assumption on the sample-mean gradient. It is exact for independent Gaussian gradients and can be used as a central-limit approximation without requiring every per-example gradient to be Gaussian.*

In summary:

$$\pi(\boldsymbol{\mu}) \propto 1, \quad \boldsymbol{\theta}_j \mid \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d), \quad \hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}\left(\boldsymbol{\theta}_j, \frac{\varepsilon_j^2}{n_j} I_d\right).$$

Observation errors are mutually independent across clients and independent of the latent client gradients and population mean. Conditional on the population mean, the client-gradient draws are independent. The unknown quantities are the population mean update $\boldsymbol{\mu}$ *and the client mean updates* $(\boldsymbol{\theta}_j)_{j \in \mathcal{M}}$*; all other quantities are oracle inputs in the analysis.*

Remark A.2 (Oracle Model and Plug-In Router)**.** *The oracle analysis treats* σ^2 *and* $\{\varepsilon_j^2\}$ *as known, so its partition is determined before the sample means are observed. FedBHG estimates these components from the entire base cluster, then substitutes* $\widehat{B}_j$ *for* B_j*. This changes both the partition and the information available for comparison. Synthetic experiments measure how often the resulting partition remains stable under the known simulation posterior; neural experiments measure predictive outcomes after training. The scope of these three comparisons is collected in Section I.*

Remark A.3 (Meaning of "Almost Surely" Under the Hierarchical Model)**.** *Because Assumption A.1 uses the improper flat prior* $\pi(\boldsymbol{\mu}) \propto 1$ *on* $\mathbb{R}^d$*, statements involving probability are interpreted* conditionally on a fixed value of $\boldsymbol{\mu}$*. More precisely, whenever we say that an event involving the observed quantities* $(\hat{\boldsymbol{\theta}}_j)_{j \in \mathcal{M}}$ *holds "almost surely," we mean that for every fixed* $\boldsymbol{\mu} \in \mathbb{R}^d$*, it holds for almost every realization under the proper conditional law induced by Assumption A.1 given* $\boldsymbol{\mu}$*.*

Equivalently, all null sets are taken with respect to the Gaussian sampling law of $(\hat{\boldsymbol{\theta}}_j)_{j \in \mathcal{M}}$ *conditional on* $\boldsymbol{\mu}$*, rather than with respect to a nonexistent probability law on the full hierarchy. This convention is sufficient for all almost-sure arguments below, since they concern only equalities and inequalities among the observed sample means.*

Given a client group $C \subseteq \mathcal{M}$, we write $\hat{\boldsymbol{\theta}}_C := (\hat{\boldsymbol{\theta}}_j)_{j \in C}$ for the collection of observed data within C.

The routing interpretation of B_j is the local-versus-pooling tradeoff under the model: small B_j corresponds to a precise local summary relative to between-client variation, whereas large B_j corresponds to greater sampling uncertainty. Section C.1 gives the corresponding conditional expected-loss comparison under shared quadratic curvature.

A.2 Supporting Posterior and Risk Derivations

The coalition comparisons require the posterior mean and the cancellation of a common posterior-variance term. We derive both from the hierarchy here; Section C subsequently uses the same mean for learning actions. These are standard hierarchical-Gaussian calculations (Gelman et al., 2013). Because the d coordinates are independent under Assumption A.1, it suffices to treat the scalar case $d = 1$ first.

A.2.1 Posterior Distribution of a Client Mean Update

Proposition A.4 (Conditional Posterior of θ_j Given $(\mu, \hat{\theta}_C)$)**.** *Consider* $d = 1$ *under Assumption A.1. Given the global mean* μ *and the coalition observations* $\hat{\theta}_C$*, the posterior of* θ_j *satisfies*

$$\theta_j \mid \mu, \hat{\theta}_C \;\equiv\; \theta_j \mid \mu, \hat{\theta}_j \;\sim\; \mathcal{N}\big(\theta_j^*, V_j\big),$$

where $\theta_j^* = B_j \mu + (1 - B_j)\hat{\theta}_j$*,* $V_j = B_j \sigma^2$*, and* $B_j = \frac{\varepsilon_j^2/n_j}{\varepsilon_j^2/n_j + \sigma^2}$*. The independence* $\theta_j \perp \hat{\theta}_{C \setminus \{j\}} \mid \mu$ *follows from the conditional independence structure of the model.*

Proposition A.5 (Posterior Distribution of μ Given Coalition Data)**.** *Consider $d = 1$ under Assumption A.1. Given the coalition observations $\hat{\theta}_C$, the posterior of the global mean satisfies*

$$\mu \mid \hat{\theta}_C \;\sim\; \mathcal{N}\left(\vartheta_C,\; \Big(\sum_{j\in C} \alpha_j\Big)^{-1}\right),$$

where $\vartheta_C = \frac{\sum_{j\in C} \alpha_j \hat{\theta}_j}{\sum_{j\in C} \alpha_j}$ and $\alpha_j = (\varepsilon_j^2/n_j + \sigma^2)^{-1}$.

Lemma A.6 (Scalar Posterior of θ_j Given $\hat{\theta}_C$)**.** *Consider $d = 1$ under Assumption A.1. For player j belonging to coalition C,*

$$\theta_j \mid \hat{\theta}_C \;\sim\; \mathcal{N}\left(B_j\,\vartheta_C + (1-B_j)\hat{\theta}_j,\; B_j\sigma^2 + B_j^2\Big(\sum_{i\in C} \alpha_i\Big)^{-1}\right).$$

Proof. The posterior of $\theta_j \mid \hat{\theta}_C$ is obtained by integrating the Gaussian conditional law of θ_j given $(\mu, \hat{\theta}_C)$ against the Gaussian posterior law of $\mu \mid \hat{\theta}_C$. Since the conditional mean is affine in μ and the conditional variance does not depend on μ, the law of total expectation and variance yields the claimed mean and variance directly.

By Proposition A.4,

$$\theta_j \mid \mu, \hat{\theta}_C \;\equiv\; \theta_j \mid \mu, \hat{\theta}_j \;\sim\; \mathcal{N}\big(B_j\mu + (1-B_j)\hat{\theta}_j,\; B_j\sigma^2\big).$$

By Proposition A.5,

$$\mu \mid \hat{\theta}_C \;\sim\; \mathcal{N}\left(\vartheta_C,\; \Big(\sum_{i\in C} \alpha_i\Big)^{-1}\right).$$

Taking conditional expectation over μ gives

$$\mathbb{E}[\theta_j \mid \hat{\theta}_C] = \mathbb{E}\big[\mathbb{E}[\theta_j \mid \mu, \hat{\theta}_C] \mid \hat{\theta}_C\big] = \mathbb{E}\big[B_j\mu + (1-B_j)\hat{\theta}_j \mid \hat{\theta}_C\big] = B_j\vartheta_C + (1-B_j)\hat{\theta}_j.$$

For the variance, apply the law of total variance:

$$\begin{aligned}
\mathrm{Var}(\theta_j \mid \hat{\theta}_C) &= \mathbb{E}\big[\mathrm{Var}(\theta_j \mid \mu, \hat{\theta}_C) \mid \hat{\theta}_C\big] + \mathrm{Var}\big(\mathbb{E}[\theta_j \mid \mu, \hat{\theta}_C] \mid \hat{\theta}_C\big) \\
&= B_j\sigma^2 + \mathrm{Var}\big(B_j\mu + (1-B_j)\hat{\theta}_j \mid \hat{\theta}_C\big) \\
&= B_j\sigma^2 + B_j^2\,\mathrm{Var}(\mu \mid \hat{\theta}_C) \\
&= B_j\sigma^2 + B_j^2\Big(\sum_{i\in C} \alpha_i\Big)^{-1}.
\end{aligned}$$

Finally, $\theta_j \mid \mu, \hat{\theta}_C$ is Gaussian, and its conditional mean is affine in the Gaussian random variable $\mu \mid \hat{\theta}_C$; therefore the marginal law $\theta_j \mid \hat{\theta}_C$ is Gaussian with the mean and variance computed above. □

Applying Lemma A.6 coordinate-wise immediately yields the multivariate posterior.

Theorem A.7 (Hierarchical Gaussian Posterior)**.** *Under Assumption A.1, for player j belonging to coalition C,*

$$\boldsymbol{\theta}_j \mid \hat{\boldsymbol{\theta}}_C \;\sim\; \mathcal{N}\big(\boldsymbol{\theta}_j^*,\; V_j\, I_d\big),$$

where

$$\begin{aligned}
\boldsymbol{\theta}_j^* &:= B_j\,\boldsymbol{\vartheta}_C + (1-B_j)\hat{\boldsymbol{\theta}}_j && \textit{(posterior mean)}, \\
V_j &:= B_j\sigma^2 + B_j^2\Big(\sum_{i\in C} \alpha_i\Big)^{-1} && \textit{(posterior variance)}.
\end{aligned}$$

Proof. Under Assumption A.1 the d coordinates are independent. Applying Lemma A.6 to each coordinate independently yields d independent Gaussian marginals sharing the same variance V_j, giving the stated spherical-covariance Gaussian. □

With the posterior distribution in hand, we can characterize the conditional MSE that drives the local-versus-coalition decision.

A.2.2 Conditional MSE Decomposition

Theorem A.7 immediately implies a bias-variance decomposition of the conditional MSE used in the subsequent no-deviation arguments.

Theorem A.8 (Conditional MSE Decomposition)**.** *Let $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\hat{\boldsymbol{\theta}}_C)$ be any estimator measurable with respect to $\hat{\boldsymbol{\theta}}_C$. Under Assumption A.1, for player $j \in C$,*

$$\mathbb{E}\Big[\|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_C\Big] \;=\; \underbrace{\|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j^*\|_2^2}_{\textit{squared bias}} + \underbrace{d\,V_j}_{\textit{irreducible variance}},$$

where $\boldsymbol{\theta}_j^*$ and V_j are as in Theorem A.7. The second term is independent of $\boldsymbol{\vartheta}$; hence the conditional MSE is uniquely minimized by $\boldsymbol{\vartheta} = \boldsymbol{\theta}_j^*$.

Proof. By Theorem A.7, $\boldsymbol{\theta}_j \mid \hat{\boldsymbol{\theta}}_C \sim \mathcal{N}(\boldsymbol{\theta}_j^*, V_j I_d)$ with $\mathbb{E}[\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^* \mid \hat{\boldsymbol{\theta}}_C] = \mathbf{0}$. The standard bias-variance identity gives

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j\|_2^2 \mid \hat{\boldsymbol{\theta}}_C\Big] &= \|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j^*\|_2^2 + 2(\boldsymbol{\vartheta} - \boldsymbol{\theta}_j^*)^\top \mathbb{E}[\boldsymbol{\theta}_j^* - \boldsymbol{\theta}_j \mid \hat{\boldsymbol{\theta}}_C] + \mathbb{E}\Big[\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\|_2^2 \mid \hat{\boldsymbol{\theta}}_C\Big] \\ &= \|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j^*\|_2^2 + \operatorname{tr}(V_j I_d) = \|\boldsymbol{\vartheta} - \boldsymbol{\theta}_j^*\|_2^2 + d\,V_j,\end{aligned}$$

where the cross term vanishes because $\mathbb{E}[\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^* \mid \hat{\boldsymbol{\theta}}_C] = \mathbf{0}$. Since $d\,V_j$ does not depend on $\boldsymbol{\vartheta}$, the conditional MSE is minimized if and only if $\boldsymbol{\vartheta} = \boldsymbol{\theta}_j^*$, and the minimum is unique since $\|\cdot\|_2^2$ is strictly convex. □

Theorem A.8 also implies the following one-dimensional comparison. For the variance-adaptive coalition estimator $\boldsymbol{\vartheta}_U$ and the local estimator $\hat{\boldsymbol{\theta}}_j$, both measurable with respect to the conditioning set $U \supseteq \{j\}$:

$$\mathbb{E}\Big[\|\boldsymbol{\vartheta}_U - \boldsymbol{\theta}_j\|_2^2 \mid \hat{\boldsymbol{\theta}}_U\Big] - \mathbb{E}\Big[\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_2^2 \mid \hat{\boldsymbol{\theta}}_U\Big] = (1 - 2B_j)\|\boldsymbol{\vartheta}_U - \hat{\boldsymbol{\theta}}_j\|_2^2. \tag{9}$$

To see this, apply Theorem A.8 twice with $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_U$ and $\boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}_j$ respectively. The irreducible variance $d\,V_j$ cancels, leaving $\|\boldsymbol{\vartheta}_U - \boldsymbol{\theta}_j^*\|_2^2 - \|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2^2$. Since $\boldsymbol{\theta}_j^* = B_j\boldsymbol{\vartheta}_U + (1 - B_j)\hat{\boldsymbol{\theta}}_j$, we have $\boldsymbol{\vartheta}_U - \boldsymbol{\theta}_j^* = (1 - B_j)(\boldsymbol{\vartheta}_U - \hat{\boldsymbol{\theta}}_j)$ and $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^* = -B_j(\boldsymbol{\vartheta}_U - \hat{\boldsymbol{\theta}}_j)$, so the difference equals $\big[(1 - B_j)^2 - B_j^2\big]\|\boldsymbol{\vartheta}_U - \hat{\boldsymbol{\theta}}_j\|_2^2 = (1 - 2B_j)\|\boldsymbol{\vartheta}_U - \hat{\boldsymbol{\theta}}_j\|_2^2$. Equation (9) shows that the sign of $\frac{1}{2} - B_j$ determines whether the coalition estimator or the local estimator has lower conditional MSE for player j: unless the two estimators coincide, pooling helps if and only if $B_j > \frac{1}{2}$.

B Stable Sharing: From Scalar Uncertainty to General Covariance

The individual local-versus-common comparison leaves a coalition question: can a proposed group weakly improve every member's posterior risk and strictly improve one member's risk? We define this comparison using the incumbent/proposal summary union, prove a stable construction, and characterize its fixed aggregation weights. We then extend the centroid argument to covariance matrices and identify how conflicting directions can defeat stable sharing. Section C translates the scalar comparisons into expected loss under shared quadratic geometry.

B.1 Coalition Model and Variance-Adaptive Sharing

Voluntary Cooperation for Mean-Update Estimation. Client j seeks to estimate its latent mean update $\boldsymbol{\theta}_j$. Acting alone, it uses $\hat{\boldsymbol{\theta}}_j$; a coalition instead deploys one shared update estimate $\boldsymbol{\vartheta}_C$. Whether sharing helps depends on the within-client uncertainty ε_j^2/n_j relative to the between-client variation σ^2.

Common-Update Constraint. We study the setting in which any cooperating coalition $C \subseteq \mathcal{M}$ must adopt one shared update estimator $\boldsymbol{\vartheta}_C$. This constraint matches the clustered federated learning interface, where members of a cluster receive one server update. The estimator below is the posterior mean of the population mean gradient $\boldsymbol{\mu}$, not the unconstrained client-specific posterior update.

Definition B.1 (Variance-Adaptive Weighted Average)**.** *For a coalition C with clients having quantities $(\sigma^2, \varepsilon_j^2, n_j, \hat{\boldsymbol{\theta}}_j)_{j\in C}$, the **variance-adaptive weighted average** is*

$$\boldsymbol{\vartheta}_C := \frac{\sum_{j\in C} \alpha_j \, \hat{\boldsymbol{\theta}}_j}{\sum_{j\in C} \alpha_j}, \qquad \alpha_j = \frac{1}{\sigma^2 + \varepsilon_j^2/n_j}.$$

As established in Theorem A.7, $\boldsymbol{\vartheta}_C$ equals the posterior mean of $\boldsymbol{\mu}$ given $\hat{\boldsymbol{\theta}}_C$ under Assumption A.1, making it the Bayes-optimal estimator for the population mean. The common-update constraint makes this one coalition estimator the common action evaluated for every member.

Section B.2 formalizes the comparison and deviation rules. The posterior and local-versus-pooled risk identities were established in Sections A.2 and A.2.2.

B.2 Pair-Conditioned Coalition Comparisons and the No-Deviation Test

For every proposed coalition, we first compare each member's posterior update risk with that member's incumbent risk. Under shared quadratic geometry, Proposition C.5 instead expresses expected post-update loss through posterior risk in the common H^{-1} metric. Conventional hedonic games encode such comparisons through a fixed preference ordering over coalitions. Our conditioning set instead depends on the two coalitions being compared, so transitivity is neither assumed nor needed. We borrow the strict-core blocking test—weak improvement for every deviator and strict improvement for at least one—but apply it directly to these pairwise comparisons.

Definition B.2 (Coalition-Comparison Model and Outcome)**.** *Let $\mathcal{M}$ be a finite set of players. For player j and coalitions $S, T \subseteq \mathcal{M}$ containing j, the notation $S \succeq_j T$ denotes the pairwise posterior-risk comparison in Definition B.5. We write $S \succ_j T$ when the corresponding inequality is strict and $S =_j T$ when the risks are equal. No transitive ordering over all coalitions is assumed. An* outcome *is a partition π of $\mathcal{M}$ into disjoint coalitions. Two special outcomes are the* singleton partition $\{\{j\} : j \in \mathcal{M}\}$ *and the* grand coalition $\{\mathcal{M}\}$.

The following criterion borrows the weak-for-all, strict-for-one blocking test used in strict-core analysis, but applies it directly to pair-conditioned posterior-risk comparisons. Because those comparisons need not form a fixed preference ordering, satisfying the criterion is not asserted to be classical strict-core stability.

Definition B.3 (Pair-Conditioned No-Weak-Deviation Criterion)**.** *Let π be a partition of $\mathcal{M}$. A non-empty coalition $S \subseteq \mathcal{M}$* weakly improves upon π with a strict gain *if:*

(i) there exists $i_0 \in S$ such that $S \succ_{i_0} \pi(i_0)$; and

(ii) $S \succeq_i \pi(i)$ for every $i \in S$,

where, for each $i \in S$, the comparison S vs. $\pi(i)$ is evaluated under Definition B.5 with conditioning set $\hat{\boldsymbol{\theta}}_{S\cup\pi(i)}$ (so the conditioning set varies with the deviator's incumbent coalition). The partition π satisfies the pair-conditioned no-weak-deviation criterion *if no such coalition exists. This displayed condition is the entire mathematical claim; we do not identify it with classical strict-core stability unless the comparisons arise from conventional hedonic preferences.*

Remark B.4 (Relation to Standard Stability Notions)**.** *If the pairwise comparisons arise from conventional hedonic preferences, Definition B.3 is standard strict core stability (Bogomolnaia &*

Jackson, 2002). Even without transitivity, its two direct consequences remain valid. First, it excludes a core block because a coalition whose members all strictly improve also satisfies Definition B.3. Second, it excludes an individually accepted deviation. Following Bogomolnaia & Jackson (2002), such a deviation exists if a player j can find $C \in \pi \cup \{\emptyset\}$ such that $C \cup \{j\} \succ_j \pi(j)$ and $C \cup \{j\} \succeq_i C$ for every $i \in C$. Then $C \cup \{j\}$ itself satisfies the displayed deviation test: the deviator strictly improves and every incumbent weakly accepts. Thus Definition B.3 directly implies both no core block and individual stability; we do not use or assert an implication from core stability to individual stability.

Comparison Rule from the Mean-Gradient Model. Suppose a coalition C uses estimator $\boldsymbol{\vartheta}_C$ (Definition B.1). Client $j \in C$ prefers coalitions whose shared update more accurately estimates its latent mean update $\boldsymbol{\theta}_j$. We formalize this using posterior MSE.

Definition B.5 (Data-Conditional Posterior-Risk Comparison)**.** *For player $j \in \mathcal{M}$ and two coalitions $S, T \subseteq \mathcal{M}$ with $j \in S \cap T$, write $S \succeq_j T$ if*

$$\mathbb{E}\Big[\|\boldsymbol{\vartheta}_S - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_{S\cup T}\Big] \;\le\; \mathbb{E}\Big[\|\boldsymbol{\vartheta}_T - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_{S\cup T}\Big],$$

where the expectation is over the posterior randomness in $\boldsymbol{\theta}_j$ given $\hat{\boldsymbol{\theta}}_{S\cup T}$ under Assumption A.1. Write $S \succ_j T$ if the inequality is strict and $S =_j T$ if equality holds.

Remark B.6 (Expected-Loss Interpretation)**.** *Definition B.5 is stated in posterior-risk form because it is defined under Assumption A.1 alone. If Assumption C.2 also holds, Proposition C.5 expresses expected post-update loss as posterior risk in the H^{-1} geometry. This need not give the same pairwise ranking as Euclidean MSE. Theorem C.6 instead proves the loss-based no-deviation conclusion directly by rerunning the argument in that common weighted geometry.*

Conditioning Set Design. When player j compares coalitions S and T, both estimators $\boldsymbol{\vartheta}_S$ and $\boldsymbol{\vartheta}_T$ are functions of observed data. The conditioning set is therefore a substantive modeling choice. We use $\hat{\boldsymbol{\theta}}_{S\cup T}$ so that both options are evaluated with the same information drawn from the two coalitions being compared. This union conditioning also gives a common information set when a proposed coalition is contained in one incumbent pool, which is the case used in Theorem B.10. It is not a canonical hedonic preference construction: because the conditioning set changes with (S, T), the resulting comparisons can be non-transitive (Section B.6). Alternative choices lead to different decisions; Section B.7 shows that conditioning on all players' observations can force singletons even when the union-conditioned rule favors pooling.

B.3 A Single-Round Interpretation of Union Conditioning

The comparison rule has an operational interpretation in a restricted-information, single-round model. Fix the common checkpoint, client identities, sample sizes and oracle variance components. These are the known model inputs for the comparison. We assume the mean-gradient hierarchy conditional on these fixed inputs; the protocol does not derive that hierarchy for a checkpoint learned from these clients' data. The flat prior on $\boldsymbol{\mu}$ specifies generalized-Bayes beliefs. After any nonempty set of summaries is observed, its normalized posterior is proper.

An incumbent partition π is fixed using only the known model inputs. Client i initially receives the labeled mean summaries in $\pi(i)$. The observations of a voting client in this model consist only of the disclosed summaries. Additional raw data, training history, or public statistics are allowed only if they leave the stated posterior unchanged. In particular, knowledge held by a server is not automatically public.

One nonempty proposed coalition S is announced, using a fixed choice or independent public randomization. Its law may depend on the known model inputs but not on the unknown population mean, latent gradients, or observed summaries. Each proposed member automatically discloses its own mean summary to every other proposed member; it does not relay the other summaries from its incumbent group. No additional data-dependent messages or earlier proposals are observed. Members then submit simultaneous comparison ballots: reject if the proposed common action has greater posterior risk, tie if the risks agree, and prefer if the proposed action has lower risk. The proposal is accepted exactly when there is no rejection and at least one preference ballot. These ballots prescribe posterior comparisons, not strategically optimal voting.

Proposition B.7 (Union conditioning before a single-round ballot)**.** *For clients observing only the specified disclosures, conditional on the known model inputs and an announced proposal $S = s$, member $i \in s$'s pre-ballot transcript generates*

$$\mathcal{F}_i(s) = \sigma\Big(\widehat{\boldsymbol{\theta}}_k : k \in s \cup \pi(i)\Big).$$

Its generalized-Bayes posterior is therefore the posterior used by Definition B.5. The same posterior conclusion holds with additional posterior-irrelevant information. In particular, if $\pi = \Pi_c$, the protocol accepts no proposal almost surely for every fixed $\boldsymbol{\mu}$ in the sense of Remark A.3. The same conclusion holds for expected-loss ballots under the assumptions of Theorem C.6.

Proof. Before the ballot, client i knows the summaries in its incumbent group and the individual summaries disclosed by s. Conversely, its transcript contains no other random observation once the known model inputs and $S = s$ are fixed. Exogeneity makes the proposal likelihood constant in the hierarchy's unknowns, so it cancels from the normalized generalized-Bayes posterior. With $U = s \cup \pi(i)$, the resulting posterior mean is

$$\mathbb{E}_{\mathrm{GB}}[\boldsymbol{\theta}_i \mid \mathcal{F}_i(s)] = (1 - B_i)\widehat{\boldsymbol{\theta}}_i + B_i \boldsymbol{\vartheta}_U.$$

Both compared actions are measurable under this information. Their posterior-risk comparison is Definition B.5, and the acceptance test is Definition B.3. Theorem B.10 excludes acceptance for Π_c; Theorem C.6 excludes acceptance under expected-loss ballots. The probability statement concerns fixed-$\boldsymbol{\mu}$ sampling, not a probability distribution induced by the improper prior. □

This interpretation identifies one setting in which union conditioning describes the disclosed summaries. It does not establish an equilibrium of strategic voting or coalition formation. An endogenous proposal or partition may reveal information; previous proposals, votes, or additional disclosures may enlarge a client's information. Full-data clients require a sufficiency argument or a richer model. The fitted FedBHG route, which uses estimated variance components and all-cluster summaries, is not covered by this protocol proposition.

B.4 Singleton-or-Pool Partition and Its No-Deviation Guarantee

We now analyze the data-conditional coalition comparisons. We first identify the classification that determines the routing partition, then prove the section's main structural claim: no coalition can weakly improve every member's posterior risk while strictly improving at least one member's risk. Core blocks and individually accepted deviations are then excluded directly as special cases.

B.4.1 Confident vs. Unconfident Players

By Theorem A.8, the shrinkage weight determines each player's attitude toward pooling:

$$B_j = \frac{\varepsilon_j^2/n_j}{\varepsilon_j^2/n_j + \sigma^2} \in (0, 1).$$

Since the optimal estimator $\boldsymbol{\theta}_j^* = B_j \boldsymbol{\vartheta}_C + (1 - B_j)\hat{\boldsymbol{\theta}}_j$ is a convex combination of the coalition aggregate and the player's local estimate, B_j controls how heavily player j relies on pooling. A player with small B_j (i.e. $\varepsilon_j^2/n_j \ll \sigma^2$) has a precise local estimate and gains little from pooling; they are *confident* in their own data. A player with large B_j (i.e. $\varepsilon_j^2/n_j \gtrsim \sigma^2$) has a noisy local estimate and the posterior comparison favors pooling; they are *unconfident*. Equation (9) makes this precise: whenever the local and pooled estimates differ, pooling strictly lowers conditional MSE for player j if and only if $B_j > \frac{1}{2}$, making $B_j = \frac{1}{2}$ the classification boundary.[1]

[1]This threshold $\varepsilon_j^2/n_j < \sigma^2$ coincides with the "large player" classification of Donahue & Kleinberg (2021a), who call a player large if $n_j > \varepsilon^2/\sigma^2$ (assuming a common local variance $\varepsilon_j = \varepsilon$). Our confident players correspond to their large players; our unconfident players to their small players. We therefore do not claim novelty for the reliability threshold alone. Donahue & Kleinberg analyze an ex-ante model-sharing game; our theorem conditions on common-checkpoint gradient summaries and proves a direct pair-conditioned no-deviation statement for the induced singleton-or-pool action.

Definition B.8 (Confident vs. Unconfident Players)**.** *Under Assumption A.1, player j is* confident *if $B_j < \frac{1}{2}$ (equivalently, $\varepsilon_j^2/n_j < \sigma^2$) and* unconfident *if $B_j \geq \frac{1}{2}$ (equivalently, $\varepsilon_j^2/n_j \geq \sigma^2$). We write $C_{\mathrm{conf}} := \{j \in \mathcal{M} : B_j < \frac{1}{2}\}$ and $C_{\mathrm{unconf}} := \{j \in \mathcal{M} : B_j \geq \frac{1}{2}\}$.*

Given the classification above, the partitioning and estimator-assignment rules follow directly.

Definition B.9 (Confidence Partition)**.** *Under Assumption A.1, the* ***confidence partition*** *is the "singleton-or-pool" partition*

$$\Pi_c := \big\{\{j\} : j \in C_{\mathrm{conf}}\big\} \cup \{C_{\mathrm{unconf}}\},$$

where each confident player forms a singleton coalition and all unconfident players form a single pool. (If $C_{\mathrm{unconf}} = \emptyset$, the pool is omitted.)

For each coalition $C \in \Pi_c$ we assign a ***common estimator*** *$\boldsymbol{\vartheta}_C$ as follows:*

$$\boldsymbol{\vartheta}_C := \begin{cases} \hat{\boldsymbol{\theta}}_j, & \text{if } C = \{j\} \text{ for some } j \in C_{\mathrm{conf}}, \\ \dfrac{\sum_{i \in C_{\mathrm{unconf}}} \alpha_i \hat{\boldsymbol{\theta}}_i}{\sum_{i \in C_{\mathrm{unconf}}} \alpha_i}, & \text{if } C = C_{\mathrm{unconf}}. \end{cases}$$

For the unconfident pool, this estimator is the variance-adaptive weighted average (Definition B.1), which by Proposition A.5 equals the posterior mean of $\boldsymbol{\mu}$ given $\hat{\boldsymbol{\theta}}_{C_{\mathrm{unconf}}}$ and is therefore Bayes-optimal for estimating the population mean. Confident players retain their local estimates; by (9)*, these have lower conditional MSE than pooling whenever $B_j < \frac{1}{2}$ and the two estimates differ.*

The scalar inequality $B_j \gtrless \frac{1}{2}$ lifts the pairwise MSE comparison to the following coalition-level theorem for Π_c.

B.4.2 No Weakly Improving Coalition Deviation

We now establish that the confidence partition Π_c satisfies the precise no-deviation criterion in Definition B.3. The theorem is stated directly so it does not rely on transitivity or on an implication hierarchy among classical hedonic stability notions.

Theorem B.10 (Singleton-or-Pool Pair-Conditioned No-Deviation)**.** *Fix Assumption A.1, the variance-adaptive estimator (Definition B.1), and the data-conditional comparison rule (Definition B.5). Then, almost surely in the sense of Remark A.3, the confidence partition Π_c (Definition B.9) admits no coalition that weakly improves every member's posterior risk and strictly improves at least one member's risk. Equivalently, Π_c satisfies the pair-conditioned no-weak-deviation criterion of Definition B.3.*

Proof. Fix $\boldsymbol{\mu}$. For every confident client i and non-singleton coalition $S \ni i$, the vector $\boldsymbol{\vartheta}_S - \widehat{\boldsymbol{\theta}}_i$ is a nondegenerate Gaussian linear combination of independent summaries. Hence it equals zero with probability zero. There are finitely many such pairs (i, S); work outside their union of null events. By the local–pooled identity, any non-singleton proposal containing a confident client strictly increases that client's risk relative to its incumbent singleton. A singleton proposal cannot strictly improve its unchanged assignment. Consequently, a qualifying deviation must satisfy $\emptyset \neq S \subseteq U := C_{\mathrm{unconf}}$.

All its members compare against U, conditional on the same summaries $\widehat{\boldsymbol{\theta}}_U$. Write $\Delta = \boldsymbol{\vartheta}_S - \boldsymbol{\vartheta}_U$ and let d_i be client i's posterior risk at the proposed action minus its risk at the incumbent action. The posterior covariance cancels, giving

$$d_i = \|\Delta\|_2^2 + 2(1 - B_i)\left\langle \Delta, \boldsymbol{\vartheta}_U - \widehat{\boldsymbol{\theta}}_i \right\rangle.$$

Since $1 - B_i = \sigma^2 \alpha_i$ and $\sum_{i \in S} \alpha_i \widehat{\boldsymbol{\theta}}_i = (\sum_{i \in S} \alpha_i)\boldsymbol{\vartheta}_S$,

$$\sum_{i \in S} d_i = \left(\sum_{i \in S} (2B_i - 1)\right) \|\Delta\|_2^2 \geq 0.$$

Weak improvement by every member and strict improvement by one would instead give $\sum_{i \in S} d_i < 0$, a contradiction. This also covers $\Delta = 0$ and threshold clients: all-weak improvements with zero sum are all ties and cannot satisfy the strict-gain requirement. Thus the conclusion holds outside the stated null set for every fixed $\boldsymbol{\mu}$. □

Corollary B.11 (All-Strict and Accepted-Unilateral-Deviation Consequences)**.** Π_c *has neither an all-strict core block nor an individually accepted deviation under either the posterior-risk comparison or the expected-loss comparison of Theorem C.6, almost surely in the sense of Remark A.3.*

Proof. An all-strict core block is a special case of Definition B.3. An individually accepted deviation by j into C would make $C \cup \{j\}$ weakly better for every incumbent and strictly better for j, and hence would also satisfy that definition. Theorems B.10 and C.6 exclude both cases directly under their respective comparisons; no implication among classical stability notions is used. □

B.5 When Participation Forces Precision Weights

For this subsection write $Y_i = \widehat{\boldsymbol{\theta}}_i$, $q_C = \boldsymbol{\vartheta}_C$, $A_C = \sum_{j \in C} \alpha_j$, and $m_i(C) = (1 - B_i)Y_i + B_i q_C$. For a common action a, let $R_i(a \mid C) = \mathbb{E}[\|a - \boldsymbol{\theta}_i\|_2^2 \mid Y_C]$. These abbreviations retain the hierarchy and the posterior beliefs above.

The preceding construction uses the posterior mean of the population parameter as the common action. Bayes optimality for that parameter does not, by itself, explain why every client should accept the action: clients care about their own latent targets. We now characterize the same weights through that participation requirement.

Fix a group C with $m = |C|$, and retain the hierarchy and posterior beliefs of Section A. Consider an alternative common action

$$q_p = \sum_{j \in C} p_j Y_j, \qquad p_j > 0, \qquad \sum_{j \in C} p_j = 1. \tag{10}$$

The weights may depend on the known variance components but are fixed before Y_C is observed. Changing the deployed aggregate does not change the posterior: $m_i(C)$ is still determined by the Gaussian hierarchy and its precision mean q_C.

We say that p satisfies *uniform posterior individual rationality* when

$$R_i(q_p \mid C) \le R_i(Y_i \mid C) \quad \text{for every } i \in C \text{ and every } y_C \in (\mathbb{R}^d)^m. \tag{11}$$

For each observation, this requirement tests only the local outside option. It does not test subgroup deviations, and it must hold uniformly over all observations.

Theorem B.12 (Fixed positive weights)**.** *Suppose Assumption A.1 holds,* $m \ge 3$*, and* $B_i \ge 1/2$ *for every* $i \in C$*. A fixed positive normalized vector* p *satisfies* (11) *if and only if*

$$p_j = \frac{\alpha_j}{A_C} \quad \textit{for every } j \in C. \tag{12}$$

Equivalently, these are the only such weights for which every member weakly prefers the common action to its local action almost surely under every fixed-μ *sampling law.*

Proof. First let $d = 1$, order the members of C, and write $r_j = \alpha_j / A_C$. For each i, let e_i be the corresponding unit vector, $a_i = p - e_i$ and $c_i = r - e_i$ in $\mathbb{R}^m$. From the posterior mean $m_i(C) = y_i + B_i c_i^\top y$, the risk difference is

$$R_i(q_p \mid C) - R_i(Y_i \mid C) = (a_i^\top y)^2 - 2B_i (a_i^\top y)(c_i^\top y). \tag{13}$$

If a_i and c_i are linearly independent, there exists y with $a_i^\top y = 1$ and $c_i^\top y = 0$. Equation (13) is then positive. Thus (11) forces $a_i = k_i c_i$. The coefficient-vector entry indexed by client $j \ne i$ gives $p_j = k_i r_j$, so positivity gives $k_i > 0$. Substitution into (13) further requires $k_i \le 2B_i$.

For any distinct i, j, choose $k \notin \{i, j\}$, possible because $m \ge 3$. The entry indexed by the third client k gives $p_k = k_i r_k = k_j r_k$, hence $k_i = k_j$. Denote their common value by k_0. For a fixed j, the relation for $i \ne j$ gives $p_j = k_0 r_j$, while its own relation gives $p_j = 1 + k_0(r_j - 1)$. Therefore $k_0 = 1$ and $p = r$.

Conversely, $p = r$ reduces (13) to $(1 - 2B_i)(c_i^\top y)^2 \le 0$. For $d > 1$, necessity follows by considering observations supported in one coordinate; sufficiency follows by adding the coordinate risk differences. Finally, if a candidate p violates the uniform inequality, strict harm occurs at a point and, by continuity, on an open set. The fixed-μ Gaussian sampling density is positive everywhere, so that open set has positive probability for every μ. This proves the almost-sure equivalence as well. □

The theorem characterizes precision aggregation through every client's willingness to participate for every observation. It does not prove uniqueness of a stable partition, nor uniqueness among all possible common actions. The class of fixed scalar coefficients and the requirement over all observations are indispensable parts of the statement.

B.5.1 Two Clients and Data-Dependent Actions

For two clients, the third client index used to equate the coefficient multipliers is unavailable, so participation permits more freedom.

Proposition B.13 (Two-client exception). *For $C = \{1, 2\}$ in the setting above, write $r = \alpha_2/A_C$ and $p_2 = x$. Uniform posterior individual rationality is equivalent to*

$$1 - 2B_2(1 - r) \le x \le 2B_1 r, \qquad 0 < x < 1. \tag{14}$$

Proof. For client 1, the proposed displacement from its observation is $x(Y_2 - Y_1)$, while the precision displacement is $r(Y_2 - Y_1)$. Equation (13) is nonpositive for every observation exactly when $x \le 2B_1 r$, since $x > 0$. For client 2 the corresponding condition is $1 - x \le 2B_2(1 - r)$. These are (14), and summing coordinates extends the argument to $d > 1$. □

The interval can have positive length. For example, equal variances and $B_1 = B_2 = 3/4$ give $r = 1/2$ and $x \in [1/4, 3/4]$. Thus the threshold of three members in Theorem B.12 is substantive.

Data-dependent common actions also escape the characterization. To see this, use three equal-variance scalar clients with $B_i = 3/4$ and observations $y = (-1, 1, 2)$. Their precision mean $2/3$ differs from every local observation. Theorem A.8 implies a strict improvement over each local action. By continuity, a sufficiently small perturbation of the common action preserves all three strict improvements, also for data in a sufficiently small neighborhood of y. An action rule can use a fixed nonzero perturbation of q_C in that neighborhood and use q_C elsewhere. It satisfies the local participation inequalities everywhere and differs from precision aggregation on an open set. This observation does not establish that the altered rule withstands subgroup deviations. It establishes why Theorem B.12 must retain its restriction to weights fixed independently of the observations.

B.6 Pair Dependence and Non-Transitivity

The comparison rule in Definition B.5 is not a classical ex-ante hedonic preference ordering. It is a data-conditional comparison rule: when player j compares two coalitions $S, T \ni j$, the posterior risks are evaluated after conditioning on the observations $\widehat{\boldsymbol{\theta}}_{S \cup T}$. Because this conditioning set depends on the pair (S, T) being compared, the induced comparisons need not assemble into a global transitive ranking over all coalitions.

The following example delimits the model. It is also why Theorem B.10 is stated as a direct no-deviation theorem rather than as a claim that the comparisons define a classical hedonic solution. The proof uses only the posterior-risk comparison between a proposed coalition and each member's incumbent coalition.

Proposition B.14 (Pairwise Comparison Need Not Be Transitive). *The preference relation $\succeq_j$ of Definition B.5 need not be transitive.*

Proof. We exhibit a three-cycle. Consider $d = 1$ with four players i, a, b, c and parameters

$$\sigma^2 = 1, \quad n_j = 1 \text{ for all } j, \quad (\hat{\theta}_i, \hat{\theta}_a, \hat{\theta}_b, \hat{\theta}_c) = (5, 0, 8, 10), \quad (\varepsilon_i^2, \varepsilon_a^2, \varepsilon_b^2, \varepsilon_c^2) = (1, 3, 1, 4).$$

The resulting precision weights and shrinkage coefficients are

$$\alpha_i = \tfrac{1}{2}, \quad \alpha_a = \tfrac{1}{4}, \quad \alpha_b = \tfrac{1}{2}, \quad \alpha_c = \tfrac{1}{5}, \quad B_i = \tfrac{1}{2}.$$

Define three coalitions containing player i: $A := \{i, a\}$, $B := \{i, b\}$, $C := \{i, c\}$, with variance-adaptive means

$$\vartheta_A = \tfrac{10}{3}, \qquad \vartheta_B = \tfrac{13}{2}, \qquad \vartheta_C = \tfrac{45}{7}.$$

By Theorem A.8, when player i compares two coalitions S and T (conditioning on $U = S \cup T$), the preference reduces to comparing $|\vartheta_S - \theta_i^*(U)|$ with $|\vartheta_T - \theta_i^*(U)|$, where $\theta_i^*(U) = B_i \vartheta_U + (1 - B_i)\hat{\theta}_i$.

Comparison 1: $B \succ_i A$. $U = \{i, a, b\}$, so $\vartheta_U = 26/5$ and $\theta_i^*(U) = 51/10$.

$$|\vartheta_B - \theta_i^*(U)| = \tfrac{7}{5} = 1.4, \qquad |\vartheta_A - \theta_i^*(U)| = \tfrac{53}{30} \approx 1.77.$$

Since $1.4 < 1.77$, we have $B \succ_i A$.

Comparison 2: $A \succ_i C$. $U = \{i, a, c\}$, so $\vartheta_U = 90/19$ and $\theta_i^*(U) = 185/38$.

$$|\vartheta_A - \theta_i^*(U)| = \tfrac{175}{114} \approx 1.54, \qquad |\vartheta_C - \theta_i^*(U)| = \tfrac{415}{266} \approx 1.56.$$

Since $1.54 < 1.56$, we have $A \succ_i C$.

Comparison 3: $C \succ_i B$. $U = \{i, b, c\}$, so $\vartheta_U = 85/12$ and $\theta_i^*(U) = 145/24$.

$$|\vartheta_C - \theta_i^*(U)| = \tfrac{65}{168} \approx 0.39, \qquad |\vartheta_B - \theta_i^*(U)| = \tfrac{11}{24} \approx 0.46.$$

Since $0.39 < 0.46$, we have $C \succ_i B$.

These three comparisons yield the cycle $B \succ_i A \succ_i C \succ_i B$, which contradicts transitivity. □

Remark B.15 (Scope of the No-Deviation Theorem)**.** *Proposition B.14 does not invalidate Theorem B.10: the displayed criterion asks only whether a proposed coalition directly improves upon each member's incumbent assignment, and the proof never invokes transitivity. The theorem does not, however, recover a global preference ordering or guarantee path-independent outcomes under iterative coalition negotiation. FedBHG uses the rule only for a one-shot binary routing decision based on B_j.*

B.7 Disclosure Changes the Participation Conclusion

Conditioning every comparison on all client summaries defines a different information arrangement. The following example changes the stable partition under that arrangement. It establishes sensitivity to disclosure, not that less information is generally preferable or that fully informed clients should ignore what they know. Information-dependent coalition conclusions are familiar from informational-core models (Wilson, 1978).

Definition B.16 (All-Data Preference Relation)**.** *For player $j \in \mathcal{M}$ and coalitions S, T with $j \in S \cap T$, write $S \succeq_j^{\mathrm{all}} T$ if*

$$\mathbb{E}\Big[\|\boldsymbol{\vartheta}_S - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_{\mathcal{M}}\Big] \;\le\; \mathbb{E}\Big[\|\boldsymbol{\vartheta}_T - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_{\mathcal{M}}\Big].$$

Proposition B.17 (All-Data Conditioning Forces Singletons)**.** *Consider $d = 1$ under Assumption A.1 with $\sigma^2 = 2$, $\varepsilon_j^2 = 5$ for all j, $(n_a, n_b, n_c) = (4, 2, 2)$, and $(\hat{\theta}_a, \hat{\theta}_b, \hat{\theta}_c) = (-5, -3, 4)$. Under this configuration, player a is confident ($B_a < \frac{1}{2}$), while b and c are unconfident. Under the all-data preference relation $\succeq_j^{\mathrm{all}}$, the singleton partition is the unique partition with no individually accepted deviation*

$$\pi_{\mathrm{sing}} = \big\{\{a\}, \{b\}, \{c\}\big\}.$$

Proof. Write $\mathcal{M} = \{a, b, c\}$. Since $\sigma^2 = 2$ and $\varepsilon_j^2 = 5$ for all j,

$$\frac{\varepsilon_a^2}{n_a} = \frac{5}{4} < 2 = \sigma^2, \qquad \frac{\varepsilon_b^2}{n_b} = \frac{\varepsilon_c^2}{n_c} = \frac{5}{2} \ge 2.$$

Thus a is confident and b, c are unconfident. The corresponding precision and shrinkage weights are

$$\alpha_a = \frac{4}{13}, \qquad \alpha_b = \alpha_c = \frac{2}{9}, \qquad B_a = \frac{5}{13}, \qquad B_b = B_c = \frac{5}{9}.$$

Step 1: Coalition estimators. For each nonempty $S \subseteq \mathcal{M}$, define

$$\vartheta_S := \frac{\sum_{j \in S} \alpha_j \hat{\theta}_j}{\sum_{j \in S} \alpha_j}.$$

A direct calculation gives

$$\vartheta_{\{a\}} = -5, \qquad \vartheta_{\{b\}} = -3, \qquad \vartheta_{\{c\}} = 4,$$
$$\vartheta_{\{a,b\}} = -\frac{129}{31} \approx -4.161, \qquad \vartheta_{\{a,c\}} = -\frac{38}{31} \approx -1.226, \qquad \vartheta_{\{b,c\}} = \frac{1}{2},$$
$$\vartheta_{\mathcal{M}} = \vartheta_{\{a,b,c\}} = -\frac{7}{4}.$$

Step 2: Posterior means under all-data conditioning. Because all comparisons are conditioned on $\hat{\theta}_{\mathcal{M}}$, Theorem A.8 says that player j compares coalitions by their squared distance from the posterior mean

$$\theta_j^*(\mathcal{M}) = B_j \vartheta_{\mathcal{M}} + (1 - B_j)\hat{\theta}_j.$$

Hence

$$\theta_a^*(\mathcal{M}) = \frac{5}{13}\Big(-\frac{7}{4}\Big) + \frac{8}{13}(-5) = -\frac{195}{52} = -3.75,$$
$$\theta_b^*(\mathcal{M}) = \frac{5}{9}\Big(-\frac{7}{4}\Big) + \frac{4}{9}(-3) = -\frac{83}{36} \approx -2.306, \qquad \theta_c^*(\mathcal{M}) = \frac{5}{9}\Big(-\frac{7}{4}\Big) + \frac{4}{9}(4) = \frac{29}{36} \approx 0.806.$$

Step 3: Four strict preferences that eliminate every non-singleton partition. Using the values above,

$$|\vartheta_{\{a\}} - \theta_a^*(\mathcal{M})| = |-5 - (-3.75)| = 1.25 < |\vartheta_{\mathcal{M}} - \theta_a^*(\mathcal{M})| = \Big|-\frac{7}{4} - (-3.75)\Big| = 2,$$

so $\{a\} \succ_a^{\text{all}} \mathcal{M}$.

Likewise,

$$|\vartheta_{\{b\}} - \theta_b^*(\mathcal{M})| = \Big|-3 + \frac{83}{36}\Big| = \frac{25}{36} < \Big|\frac{1}{2} + \frac{83}{36}\Big| = \frac{101}{36} = |\vartheta_{\{b,c\}} - \theta_b^*(\mathcal{M})|,$$

so $\{b\} \succ_b^{\text{all}} \{b, c\}$.

Also,

$$|\vartheta_{\{a\}} - \theta_a^*(\mathcal{M})| = 1.25 < |\vartheta_{\{a,c\}} - \theta_a^*(\mathcal{M})|,$$

so $\{a\} \succ_a^{\text{all}} \{a, c\}$; and

$$|\vartheta_{\{b\}} - \theta_b^*(\mathcal{M})| < |\vartheta_{\{a,b\}} - \theta_b^*(\mathcal{M})|,$$

so $\{b\} \succ_b^{\text{all}} \{a, b\}$.

These strict preferences eliminate every partition containing a non-singleton coalition:

(i) $\{\{a, b, c\}\}$ admits the singleton deviation by a.

(ii) $\{\{a\}, \{b, c\}\}$ admits the singleton deviation by b.

(iii) $\{\{b\}, \{a, c\}\}$ admits the singleton deviation by a.

(iv) $\{\{c\}, \{a, b\}\}$ admits the singleton deviation by b.

Thus the singleton partition is the only remaining candidate.

Step 4: The singleton partition admits no accepted unilateral deviation. It remains to check that no player can profitably join another singleton with the host's consent. Player a would like to join b, since

$$|\vartheta_{\{a,b\}} - \theta_a^*(\mathcal{M})| < |\vartheta_{\{a\}} - \theta_a^*(\mathcal{M})|,$$

but b rejects because $\{b\} \succ_b^{\text{all}} \{a, b\}$. Similarly, player c would like to join b, but b rejects because $\{b\} \succ_b^{\text{all}} \{b, c\}$; and player c would like to join a, but a rejects because $\{a\} \succ_a^{\text{all}} \{a, c\}$. All remaining join attempts fail to benefit the deviator strictly. Hence no accepted unilateral deviation exists from

$$\pi_{\text{sing}} = \big\{\{a\}, \{b\}, \{c\}\big\},$$

so π_{sing} admits no accepted unilateral deviation. Since all other partitions were ruled out above, it is the unique partition with this property. □

The disclosure examples above concern how the information set changes coalition preferences. Separately, adding observations reduces posterior variance; the next proposition records only that variance monotonicity, not a welfare advantage for the selected partition.

B.8 The Unconfident-Client Pool Minimizes Posterior Uncertainty Among Its Subcoalitions

Adding an observation with positive precision decreases the posterior variance. The proposition records this standard consequence for the designated pool. The same monotonicity holds for confident clients and for supersets extending beyond that pool; it is not a property unique to the routing partition. It concerns posterior uncertainty, not the conditional risk of deploying a common action.

Proposition B.18 (The Unconfident Pool Minimizes Individual Posterior Uncertainty Among Its Subcoalitions)**.** *Under Assumption A.1, for every unconfident player $j \in C_{\mathrm{unconf}}$ and every strict sub-coalition $G \subsetneq C_{\mathrm{unconf}}$ with $j \in G$,*

$$V_j(C_{\mathrm{unconf}}) \;<\; V_j(G),$$

where, for any coalition C containing j,

$$V_j(C) := B_j\sigma^2 + B_j^2\Big(\sum_{i\in C}\alpha_i\Big)^{-1}$$

denotes the scalar posterior covariance factor in

$$\boldsymbol{\theta}_j \mid \hat{\boldsymbol{\theta}}_C \sim \mathcal{N}\big(\boldsymbol{\theta}_j^*(C),\, V_j(C)\, I_d\big).$$

Equivalently, among subsets of C_{unconf} containing j, the posterior covariance trace $dV_j(C)$ is strictly minimized at $C = C_{\mathrm{unconf}}$.

Proof. Since $G \subsetneq C_{\mathrm{unconf}}$, there exists $k \in C_{\mathrm{unconf}} \setminus G$. Under Assumption A.1,

$$\alpha_k = (\sigma^2 + \varepsilon_k^2/n_k)^{-1} > 0,$$

so

$$\sum_{i\in C_{\mathrm{unconf}}}\alpha_i > \sum_{i\in G}\alpha_i.$$

Therefore

$$\Big(\sum_{i\in C_{\mathrm{unconf}}}\alpha_i\Big)^{-1} < \Big(\sum_{i\in G}\alpha_i\Big)^{-1}.$$

Since $B_j^2 > 0$ and the term $B_j\sigma^2$ is common to both $V_j(C_{\mathrm{unconf}})$ and $V_j(G)$, the strict inequality

$$V_j(C_{\mathrm{unconf}}) < V_j(G)$$

follows immediately. Multiplying by d yields the equivalent statement for the posterior covariance trace. □

Remark B.19 (Individual, Not Aggregate, Guarantee)**.** *Proposition B.18 holds individually for every unconfident player j, not merely on average. Each unconfident player personally has strictly lower posterior variance in the unconfident-client pool than in any strict sub-coalition, regardless of which sub-coalition is compared and regardless of the data realization $(\hat{\boldsymbol{\theta}}_j)_{j\in\mathcal{M}}$.*

Remark B.20 (Oracle Pre-Data Property)**.** *The scalar posterior covariance factor $V_j(C) = B_j\sigma^2 + B_j^2(\sum_{i\in C}\alpha_i)^{-1}$ depends only on the known parameters σ^2, ε_j^2, n_j, and the coalition structure, not on the unknown $\boldsymbol{\theta}_j$ or the observed $\hat{\boldsymbol{\theta}}_j$. Consequently, every unconfident player can verify, from the oracle variance components before the sample means are observed, that the grand pool minimizes its irreducible estimation uncertainty. The implemented FedBHG router does not have this pre-data property because it estimates the variance components from checkpoint gradients.*

B.9 Further Consequences for the Confidence Partition

Theorem B.10 and Proposition B.18 imply three model-conditional consequences of Π_c.

Corollary B.21 (Consequences of the Confidence Partition)**.** *Under Assumption A.1 and the variance-adaptive estimator, Π_c satisfies the following three model-conditional properties under voluntary participation.*

(1) ***Confident players are individually optimal.*** *Every $j \in C_{\mathrm{conf}}$ strictly prefers $\{j\}$ over any non-singleton coalition, almost surely (in the sense of Remark A.3). The singleton uniquely minimizes j's conditional MSE among the available variance-adaptive coalition estimators. The unconstrained posterior mean remains the minimizer over all measurable estimators.*

(2) ***Unconfident players achieve minimum irreducible uncertainty.*** *By Proposition B.18, every $j \in C_{\mathrm{unconf}}$ achieves strictly lower posterior variance in C_{unconf} than in any strict subcoalition. For each such client j, no strict subcoalition $G \subsetneq C_{\mathrm{unconf}}$ containing j has lower irreducible posterior uncertainty than C_{unconf}.*

(3) ***No weakly improving coalition deviation.*** *By Theorem B.10, no coalition $S \subseteq \mathcal{M}$ weakly improves every member's posterior risk while strictly improving at least one member's risk.*

Consequently, any partition satisfying Definition B.3 must assign every confident player to a singleton. Thus Π_c is the unique partition satisfying the criterion among partitions that pool all unconfident players together; we do not claim uniqueness when C_{unconf} is further subdivided.

Proof. Items (1) and (3) follow from Theorem B.10. Item (2) is Proposition B.18. For the conditional uniqueness claim, consider any partition that assigns a confident player j to a non-singleton coalition. Equation (9) and the same almost-sure argument used in Theorem B.10 show that j strictly prefers the singleton $\{j\}$. That one-player coalition satisfies Definition B.3, a contradiction. Hence every partition satisfying the criterion assigns all confident players to singletons. □

We next show that the *aggregation rule* used within each coalition is essential to the proof: replacing variance-adaptive weighting with standard sample-size weighting can admit an accepted unilateral deviation even when all players are unconfident. Section B.11 treats covariance extensions separately.

B.10 Variance-Adaptive Aggregation is Non-Redundant

The no-deviation guarantee of Theorem B.10 relies on the variance-adaptive estimator of Definition B.1. The following proposition shows that replacing it with the standard sample-size-weighted estimator can admit an accepted unilateral deviation, even when all players are unconfident.

Proposition B.22 (Standard Averaging Admits an Accepted Unilateral Deviation)**.** *Consider $d = 1$ under Assumption A.1 with $\varepsilon^2 = 10$, $\sigma^2 = 1$, $C = \{a, b, c\}$, $(n_a, n_b, n_c) = (5, 5, 2)$, and $(\hat{\theta}_a, \hat{\theta}_b, \hat{\theta}_c) = (-3.57, -1.16, 3.58)$. All three players are unconfident. Under the standard sample-size-weighted estimator $\hat{\vartheta}_C^{\mathrm{std}} = \sum_{j \in C} n_j \hat{\theta}_j / \sum_{j \in C} n_j$, the grand-coalition partition $\{\{a, b, c\}\}$ admits an accepted unilateral deviation: player b strictly prefers the singleton $\{b\}$.*

Proof. The precision weights are $\alpha_a = \alpha_b = 1/3$, $\alpha_c = 1/6$, and $B_b = 2/3 \geq 1/2$ (unconfident). We carry out three steps.

Step 1: Variance-adaptive coalition mean.

$$\vartheta_C = \frac{\frac{1}{3}(-3.57) + \frac{1}{3}(-1.16) + \frac{1}{6}(3.58)}{\frac{1}{3} + \frac{1}{3} + \frac{1}{6}} = \frac{-0.9800}{5/6} \approx -1.176.$$

Player b's posterior mean is $\theta_b^*(C) = B_b\, \vartheta_C + (1 - B_b)\hat{\theta}_b = \frac{2}{3}(-1.176) + \frac{1}{3}(-1.16) \approx -1.171$.

Step 2: Standard coalition mean.

$$\hat{\vartheta}_C^{\mathrm{std}} = \frac{5(-3.57) + 5(-1.16) + 2(3.58)}{5 + 5 + 2} = \frac{-16.49}{12} \approx -1.374.$$

Step 3: Comparison. By Theorem A.8, conditioning on $\hat{\theta}_C$, the conditional MSE of any estimator for player b is minimized at $\theta_b^*(C)$ and increases with squared distance from $\theta_b^*(C)$. Therefore:

$$|\hat{\vartheta}_C^{\mathrm{std}} - \theta_b^*(C)| \approx |-1.374 + 1.171| \approx 0.203, \qquad |\hat{\theta}_b - \theta_b^*(C)| \approx |-1.16 + 1.171| \approx 0.011.$$

Since $0.203 > 0.011$, player b strictly prefers the singleton $\{b\}$ to C under the standard estimator. □

Remark B.23 (Structural Necessity of Variance-Adaptive Weighting)**.** *The variance-adaptive estimator $\boldsymbol{\vartheta}_C$ is precisely the posterior mean of $\boldsymbol{\mu}$ given $\hat{\boldsymbol{\theta}}_C$ under the hierarchical model (Theorem A.7), making it Bayes-optimal under squared loss. The standard estimator weights players by n_j, whereas $\boldsymbol{\vartheta}_C$ weights them by the precision $\alpha_j = (\sigma^2 + \varepsilon_j^2/n_j)^{-1}$. These diverge whenever a player has large n_j but also large ε_j^2: such a player contributes many samples yet each sample is noisy, so they receive* lower *weight under variance-adaptive aggregation than under sample-size weighting. Relative to variance-adaptive aggregation, sample-size weighting can over-weight a client with many but noisy samples, pull the coalition mean away from the Bayes-optimal value, and reduce the value of pooling for other coalition members.*

Within this oracle hierarchy and pair-conditioned comparison rule, the no-deviation theorem and standard-averaging counterexample show that the stated half-threshold guarantee depends on the variance-adaptive common update; replacing it by sample-size weighting can invalidate the guarantee. The deployed FedBHG method is specified in Section D; the next section asks when a single scalar route survives general covariance structure.

B.11 General Covariance

The posterior-risk no-deviation guarantee of Theorem B.10 is proved under Assumption A.1, which requires both the prior covariance and every client's noise covariance to be scalar multiples of the identity. This extension asks what survives once that restriction is removed. We first recall the shared-shape case, then prove a broader theorem: the same no-deviation conclusion holds whenever every direction favors the same route for each client. Finally, a small counterexample shows that conflicting directions can leave every partition vulnerable to an accepted unilateral deviation.

B.11.1 General-Covariance Hierarchical Model

We first state the generalized model that will be used throughout this section. It subsumes Assumption A.1 as the special case $\Sigma = \sigma^2 I_d$ and $E_j = \varepsilon_j^2 I_d$.

Assumption B.24 (Generalized Mean-Gradient Model)**.** *We work in $\mathbb{R}^d$. The model consists of three levels.*

- *(a)* ***Population level.*** *An unknown mean update $\boldsymbol{\mu} \in \mathbb{R}^d$ is assigned the improper flat prior with Lebesgue density $\pi(\boldsymbol{\mu}) \propto 1$; this is not a proper uniform distribution on $\mathbb{R}^d$.*
- *(b)* ***Client level.*** *A known positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ specifies the between-client mean-gradient covariance. For each client $j \in \mathcal{M}$, the latent client mean update $\boldsymbol{\theta}_j$ is drawn independently from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$.*
- *(c)* ***Gradient-summary level.*** *For each client $j \in \mathcal{M}$, the sample size n_j and noise covariance matrix $E_j \in \mathbb{R}^{d\times d}$ (positive definite, known) determine the observable $\hat{\boldsymbol{\theta}}_j \sim \mathcal{N}(\boldsymbol{\theta}_j,\, E_j/n_j)$.*

In summary:

$$\pi(\boldsymbol{\mu}) \propto 1, \quad \boldsymbol{\theta}_j \mid \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu},\, \Sigma), \quad \hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}\left(\boldsymbol{\theta}_j,\, \frac{E_j}{n_j}\right).$$

The observation errors are independent across clients and independent of the latent gradients and population mean. Conditional on the population mean, the client-gradient draws are independent. Assumption A.1 is recovered by setting $\Sigma = \sigma^2 I_d$ and $E_j = \varepsilon_j^2 I_d$ for all j. All probability statements in this section use the conditional-on-fixed- $\boldsymbol{\mu}$ convention of Remark A.3.

Under Assumption B.24 with general positive definite Σ and $\{E_j\}$, the posterior mean of $\boldsymbol{\theta}_j$ given coalition data $\hat{\boldsymbol{\theta}}_C$ is no longer expressible using a single scalar shrinkage coefficient B_j. Different directions can prescribe different amounts of pooling. The next theorems separate the case in which those directions agree from the case in which their recommendations conflict.

B.11.2 SHARED-SHAPE COVARIANCES

The following corollary identifies a broad class of anisotropic models for which the no-deviation analysis of Section B.4 carries over.

Corollary B.25 (Shared-Shape Covariances Reduce to the Spherical Case)**.** *Consider Assumption B.24 under the additional restriction that there exists a fixed positive definite matrix $K \in \mathbb{R}^{d\times d}$ and scalars $\sigma^2 > 0$, $\varepsilon_j^2 > 0$ such that*

$$\Sigma = \sigma^2 K \qquad \text{and} \qquad E_j = \varepsilon_j^2 K \quad \text{for all } j \in \mathcal{M}.$$

Then the scalar routing threshold and the no-deviation theorem of Section B.4 hold with the same scalar quantities σ^2 and $\{\varepsilon_j^2\}_{j\in\mathcal{M}}$. In particular, the scalar shrinkage coefficient $B_j = \varepsilon_j^2/n_j \,/\, (\varepsilon_j^2/n_j + \sigma^2)$ is well-defined and coordinate-independent, and the confidence partition Π_c (Definition B.9) satisfies the pair-conditioned no-weak-deviation criterion under Assumption B.24 with shared shape.

Proof. Let $K = Q\Lambda Q^\top$ be the spectral decomposition, where $Q \in \mathbb{R}^{d\times d}$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_s > 0$ for all s. Define rotated variables $\boldsymbol{\theta}'_j := Q^\top \boldsymbol{\theta}_j$, $\hat{\boldsymbol{\theta}}'_j := Q^\top \hat{\boldsymbol{\theta}}_j$, and $\boldsymbol{\mu}' := Q^\top \boldsymbol{\mu}$. Since Q is orthogonal, the rotation preserves Euclidean distances: $\|\mathbf{x} - \mathbf{y}\|_2 = \|Q^\top \mathbf{x} - Q^\top \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Under the shared-shape assumption, the rotated model satisfies, for each coordinate $s \in [d]$:

$$\theta'_{j,s} \mid \mu'_s \sim \mathcal{N}(\mu'_s,\, \sigma^2\lambda_s), \qquad \hat{\theta}'_{j,s} \mid \theta'_{j,s} \sim \mathcal{N}\left(\theta'_{j,s},\, \frac{\varepsilon_j^2 \lambda_s}{n_j}\right).$$

The coordinates decouple in the rotated system, and each is an independent scalar Gaussian–Gaussian model. The coordinate-s shrinkage coefficient is

$$B_{j,s} = \frac{(\varepsilon_j^2/n_j)\lambda_s}{(\varepsilon_j^2/n_j)\lambda_s + \sigma^2\lambda_s} = \frac{\varepsilon_j^2/n_j}{\varepsilon_j^2/n_j + \sigma^2} =: B_j,$$

which is independent of the coordinate index s. The coordinate-s precision weight is

$$\alpha_{j,s} = \frac{1}{\sigma^2\lambda_s + (\varepsilon_j^2/n_j)\lambda_s} = \frac{1}{\lambda_s} \cdot \frac{1}{\sigma^2 + \varepsilon_j^2/n_j} = \frac{\alpha_j}{\lambda_s},$$

where $\alpha_j := (\sigma^2 + \varepsilon_j^2/n_j)^{-1}$ is the scalar precision weight from Section A. The common factor λ_s^{-1} cancels in the precision-weighted coalition mean:

$$\vartheta_{C,s} = \frac{\sum_{i\in C} \alpha_{i,s}\, \hat{\theta}'_{i,s}}{\sum_{i\in C} \alpha_{i,s}} = \frac{\sum_{i\in C} \alpha_i\, \hat{\theta}'_{i,s}}{\sum_{i\in C} \alpha_i},$$

which is identical to the spherical-case formula applied to coordinate s of the rotated data. Because B_j is coordinate-independent, the confident/unconfident classification of Definition B.8 is unchanged, and hence so is the confidence partition of Definition B.9. All posterior means, MSE comparisons, and blocking-coalition arguments from Sections A and B.4 therefore apply coordinate-wise in the rotated system. Since Euclidean loss is preserved by the rotation, every conclusion carries back verbatim to the original coordinates. □

Remark B.26 (Scope of the Shared-Shape Condition)**.** *The shared-shape condition $\Sigma = \sigma^2 K$, $E_j = \varepsilon_j^2 K$ requires the prior covariance and every client noise covariance to be proportional to the same positive definite matrix K, but imposes no restriction on the anisotropy of K itself. In particular, K may have arbitrary eigenvalues and eigenvectors. The condition is satisfied when $K = I_d$ (Assumption A.1), or more generally whenever all client noise covariances and the prior covariance are scalar multiples of a common matrix. Outside this family, different directions can induce different signal-to-noise ratios, so a single scalar shrinkage coefficient B_j need not represent the posterior trade-off. Failure of shared shape does not by itself invalidate the no-deviation property: Theorem B.27 still applies whenever all directions favor the same route.*

B.11.3 When Every Direction Favors the Same Route

The shared-shape condition makes every direction use the same scalar score. The following theorem keeps only the condition needed by the no-deviation proof: for each client, all directions must favor local estimation or all must favor pooling. In this subsection, the generalized comparison is the posterior-risk inequality of Definition B.5 with Assumption B.24 and the covariance-weighted estimator below substituted for their scalar counterparts. The blocking test retains the same union-conditioned information rule and the same all-weak, one-strict criterion as Definition B.3.

Theorem B.27 (No Weakly Improving Deviations When Directions Agree)**.** *Under Assumption B.24, define*

$$W_j := \Sigma + E_j/n_j, \qquad G_j := \frac{1}{2}\left(\Sigma W_j^{-1} + W_j^{-1}\Sigma\right),$$

and use the covariance-weighted coalition mean

$$\boldsymbol{\vartheta}_C := \left(\sum_{i\in C} W_i^{-1}\right)^{-1} \sum_{i\in C} W_i^{-1}\widehat{\boldsymbol{\theta}}_i.$$

Suppose every client satisfies either $G_j \succeq \frac{1}{2}I_d$ or $G_j \preceq \frac{1}{2}I_d$. Assign a client to a singleton when $G_j \succeq \frac{1}{2}I_d$ and $G_j \neq \frac{1}{2}I_d$; pool all remaining clients. Then, almost surely, this partition admits no coalition that weakly improves every member's posterior risk and strictly improves at least one member's risk under this generalized comparison.

Proof. For any comparison whose conditioning set is C, Gaussian conditioning gives

$$\mathbb{E}[\boldsymbol{\theta}_j \mid \widehat{\boldsymbol{\theta}}_C] = \left(I_d - \Sigma W_j^{-1}\right)\boldsymbol{\vartheta}_C + \Sigma W_j^{-1}\widehat{\boldsymbol{\theta}}_j.$$

The posterior covariance cancels when comparing squared-error risks. Therefore, taking $q := \boldsymbol{\vartheta}_C$ and $y_j := \widehat{\boldsymbol{\theta}}_j$,

$$\begin{aligned}\mathbb{E}[\|q-\boldsymbol{\theta}_j\|_2^2 - \|y_j-\boldsymbol{\theta}_j\|_2^2 \mid \widehat{\boldsymbol{\theta}}_C] \\ = 2(q-y_j)^\top\left(G_j - \tfrac{1}{2}I_d\right)(q-y_j).\end{aligned} \tag{15}$$

Thus a client with $G_j \succeq \frac{1}{2}I_d$ and $G_j \neq \frac{1}{2}I_d$ strictly prefers its singleton to every non-singleton coalition almost surely: for a non-singleton C, the vector $\boldsymbol{\vartheta}_C - y_j$ has a full-dimensional Gaussian density, whereas the null space of the nonzero positive-semidefinite matrix $G_j - \frac{1}{2}I_d$ has measure zero. Such a client cannot belong to a blocking coalition.

It remains to rule out a block formed only by pooled clients. Let C_{pool} be their current coalition and let a nonempty $R \subseteq C_{\mathrm{pool}}$ deviate. Write

$$\Delta := \boldsymbol{\vartheta}_R - \boldsymbol{\vartheta}_{C_{\mathrm{pool}}}, \qquad D_i := \boldsymbol{\vartheta}_{C_{\mathrm{pool}}} - \widehat{\boldsymbol{\theta}}_i.$$

Weak improvement by client $i \in R$, conditioning on the union C_{pool}, requires

$$\|\Delta\|_2^2 + 2\Delta^\top \Sigma W_i^{-1} D_i \le 0. \tag{16}$$

The defining equations for the two covariance-weighted means imply

$$\sum_{i\in R} W_i^{-1} D_i = -\left(\sum_{i\in R} W_i^{-1}\right)\Delta.$$

Summing the left-hand side of (16) over $i \in R$ therefore gives

$$\sum_{i\in R}\left(\|\Delta\|_2^2 - 2\Delta^\top G_i \Delta\right) \ge 0,$$

because every pooled client has $G_i \preceq \frac{1}{2}I_d$. If every member of R weakly improved, each summand in (16) would be nonpositive, so their sum could be nonnegative only if all were zero. No member can then improve strictly. Hence no coalition satisfies Definition B.3 for the stated partition. □

Under shared-shape covariances, $G_j = (1 - B_j)I_d$, so Theorem B.27 reduces to the scalar half-threshold. The theorem is a posterior-MSE extension; the implemented FedBHG score remains scalar and does not estimate full covariance matrices. It does not establish the expected-loss no-deviation theorem for arbitrary loss-curvature matrices.

B.11.4 Posterior Distribution Under Diagonal (Coordinate-Specific) Covariances

To develop the technical machinery for the counterexample of Section B.12, we record the posterior distribution and conditional MSE decomposition for the case where all covariances are diagonal but not necessarily isotropic. This is a direct coordinate-wise application of the scalar identities from Section A.

Corollary B.28 (Posterior Distribution Under Diagonal Covariances)**.** *Consider Assumption B.24 with $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$ and $E_j = \mathrm{diag}(\varepsilon_{j1}^2, \dots, \varepsilon_{jd}^2)$ for all $j \in \mathcal{M}$. Fix a coalition C and a player $j \in C$. Then, conditioned on $\hat{\boldsymbol{\theta}}_C$, the coordinates of $\boldsymbol{\theta}_j$ are independent, and for each coordinate $k \in [d]$:*

$$\theta_{jk} \mid \hat{\boldsymbol{\theta}}_C \;\sim\; \mathcal{N}\big(\theta^*_{jk}(C),\; V_{jk}(C)\big),$$

where

$$\theta^*_{jk}(C) := B_{jk}\,\hat{\mu}_k(C) + (1 - B_{jk})\,\hat{\theta}_{jk},$$
$$V_{jk}(C) := B_{jk}\sigma_k^2 + B_{jk}^2\Big(\textstyle\sum_{i \in C} \alpha_{ik}\Big)^{-1},$$

and the coordinate-specific parameters are

$$B_{jk} := \frac{\varepsilon_{jk}^2/n_j}{\varepsilon_{jk}^2/n_j + \sigma_k^2}, \qquad \alpha_{ik} := \frac{1}{\sigma_k^2 + \varepsilon_{ik}^2/n_i}, \qquad \hat{\mu}_k(C) := \frac{\sum_{i \in C} \alpha_{ik}\,\hat{\theta}_{ik}}{\sum_{i \in C} \alpha_{ik}}.$$

Equivalently, $\boldsymbol{\theta}_j \mid \hat{\boldsymbol{\theta}}_C \sim \mathcal{N}(\boldsymbol{\theta}^*_j(C), \mathrm{diag}(V_{j1}(C), \dots, V_{jd}(C)))$, where $\boldsymbol{\theta}^*_j(C) = (\theta^*_{j1}(C), \dots, \theta^*_{jd}(C))^\top$.

Proof. Diagonal Σ and $\{E_j\}$ decouple the d coordinates in the joint model of Assumption B.24. Conditioned on $\hat{\boldsymbol{\theta}}_C$, each coordinate k constitutes an independent scalar Gaussian–Gaussian model with prior variance σ_k^2 and noise variance ε_{jk}^2/n_j. Applying Lemma A.6 coordinate-wise yields the stated distribution. □

Corollary B.29 (Conditional MSE Under Diagonal Covariances)**.** *Under the diagonal-covariance setting of Corollary B.28, for any estimator $\hat{\boldsymbol{\vartheta}}$ measurable with respect to $\hat{\boldsymbol{\theta}}_C$,*

$$\mathbb{E}\Big[\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_j\|_2^2 \;\Big|\; \hat{\boldsymbol{\theta}}_C\Big] \;=\; \big\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\theta}^*_j(C)\big\|_2^2 \;+\; \sum_{k=1}^{d} V_{jk}(C).$$

The second term is independent of $\hat{\boldsymbol{\vartheta}}$; the conditional MSE is uniquely minimized by $\hat{\boldsymbol{\vartheta}} = \boldsymbol{\theta}^*_j(C)$.

Proof. Apply the bias-variance decomposition from Theorem A.8 coordinate-wise, using the independence of coordinates established in Corollary B.28. The variance term is $\sum_k V_{jk}(C) = \mathrm{tr}(\mathrm{diag}(V_{j1}(C), \dots, V_{jd}(C)))$, which does not depend on $\hat{\boldsymbol{\vartheta}}$. □

In the spherical model, $B_{jk} = B_j$ for all k, so every coordinate delivers the same coalition recommendation: pool if $B_j \geq \frac{1}{2}$ and act alone otherwise. Under coordinate-dependent B_{jk}, a single player may be "confident" along some coordinates and "unconfident" along others. No scalar coalition rule can reconcile these conflicting directional prescriptions in general. The counterexample below shows that, for a simple three-player instance, every partition then admits an accepted unilateral deviation.

B.12 The Boundary: Conflicting Directions

The failure is witnessed by an explicit construction with three players, two dimensions, a spherical prior, and identical diagonal noise covariances. The two noise variances differ by the deliberately extreme factor of $10{,}000$, creating opposing confidence signals across coordinates. The example establishes possibility, not typicality.

Coalition estimator. Under Assumption B.24 with $n_j = 1$ and $E_j = E$ for all j, all players have equal precision in every coordinate, so the coordinate-wise precision weights α_{ik} are equal across i for each fixed k. Any variance-adaptive aggregation rule therefore reduces to the unweighted mean:

$$\boldsymbol{\vartheta}_C := \frac{1}{|C|} \sum_{i \in C} \hat{\boldsymbol{\theta}}_i.$$

This is the coalition estimator used throughout the counterexample.

Proposition B.30 (Every Partition Admits an Accepted Deviation Under Anisotropic Noise)**.** *Consider Assumption B.24 with $M = 3$ players, $d = 2$, $\Sigma = I_2$, and identical anisotropic noise covariances $E_1 = E_2 = E_3 = \mathrm{diag}(1/100, 100)$ with $n_j = 1$ for all j. Let the observed sample means be*

$$\hat{\boldsymbol{\theta}}_1 = (3, 1)^\top, \qquad \hat{\boldsymbol{\theta}}_2 = (4, -3)^\top, \qquad \hat{\boldsymbol{\theta}}_3 = (0, 4)^\top.$$

Under the variance-adaptive estimator and the comparison rule of Definition B.5, every partition of $\{1, 2, 3\}$ admits an individually accepted deviation. All comparisons needed for these moves can be chosen strict. The same five moves defeat every partition throughout an open neighborhood of the displayed observations, which has positive sampling probability for every fixed $\boldsymbol{\mu}$.

Proof. We work with the coordinate-specific parameters induced by $\Sigma = I_2$ and $E_j = \mathrm{diag}(1/100, 100)$:

$$\sigma_1^2 = \sigma_2^2 = 1, \qquad \varepsilon_{j1}^2 = \tfrac{1}{100}, \qquad \varepsilon_{j2}^2 = 100, \qquad n_j = 1.$$

The coordinate-specific shrinkage coefficients are therefore

$$B_{j1} = \frac{1/100}{1/100 + 1} = \frac{1}{101} \approx 0.0099 \qquad \text{(strongly confident along coordinate 1)},$$

$$B_{j2} = \frac{100}{100 + 1} = \frac{100}{101} \approx 0.9901 \qquad \text{(strongly unconfident along coordinate 2)}.$$

These values are identical across all three players and independent of the data. Since all players share the same precision in every coordinate, the coalition mean $\boldsymbol{\vartheta}_C = |C|^{-1} \sum_{i \in C} \hat{\boldsymbol{\theta}}_i$ is the unweighted average. Moreover, because α_{ik} is the same for all i given k, the posterior mean of Corollary B.28 takes the explicit form

$$\boldsymbol{\theta}_j^*(U) = \left(\frac{1}{101} (\boldsymbol{\vartheta}_U)_1 + \frac{100}{101} \hat{\theta}_{j1},\ \frac{100}{101} (\boldsymbol{\vartheta}_U)_2 + \frac{1}{101} \hat{\theta}_{j2} \right)^\top, \tag{17}$$

where $(\boldsymbol{\vartheta}_U)_k$ denotes the k-th coordinate of the coalition mean $\boldsymbol{\vartheta}_U$.

Comparison rule. By Corollary B.29, when player j compares coalitions S and T (conditioning on $\hat{\boldsymbol{\theta}}_{S \cup T}$ with $U = S \cup T$), the variance terms $\sum_k V_{jk}(U)$ cancel and the preference reduces to comparing squared distances:

$$S \succeq_j T \iff \|\boldsymbol{\vartheta}_S - \boldsymbol{\theta}_j^*(U)\|_2^2 \le \|\boldsymbol{\vartheta}_T - \boldsymbol{\theta}_j^*(U)\|_2^2.$$

Precomputed quantities. The coalition averages are:

$$\boldsymbol{\vartheta}_{\{1\}} = (3, 1)^\top, \qquad \boldsymbol{\vartheta}_{\{2\}} = (4, -3)^\top, \qquad \boldsymbol{\vartheta}_{\{3\}} = (0, 4)^\top,$$

$$\boldsymbol{\vartheta}_{\{1,2\}} = \left(\tfrac{7}{2}, -1\right)^\top, \quad \boldsymbol{\vartheta}_{\{1,3\}} = \left(\tfrac{3}{2}, \tfrac{5}{2}\right)^\top, \quad \boldsymbol{\vartheta}_{\{2,3\}} = \left(2, \tfrac{1}{2}\right)^\top, \quad \boldsymbol{\vartheta}_{\{1,2,3\}} = \left(\tfrac{7}{3}, \tfrac{2}{3}\right)^\top.$$

Substituting into (17) with $U = \{1, 2, 3\}$:

$$\boldsymbol{\theta}_1^*(123) = \left(\tfrac{907}{303},\ \tfrac{203}{303} \right)^\top,$$

$$\boldsymbol{\theta}_2^*(123) = \left(\tfrac{1207}{303},\ \tfrac{191}{303} \right)^\top,$$

$$\boldsymbol{\theta}_3^*(123) = \left(\tfrac{7}{303},\ \tfrac{212}{303} \right)^\top.$$

With $U = \{1, 2\}$:

$$\boldsymbol{\theta}_1^*(12) = \left(\tfrac{607}{202}, -\tfrac{99}{101}\right)^\top,$$
$$\boldsymbol{\theta}_2^*(12) = \left(\tfrac{807}{202}, -\tfrac{103}{101}\right)^\top.$$

Exhaustive enumeration of all five partitions. There are five partitions of $\{1, 2, 3\}$. For this auxiliary counterexample only, we use "IS deviation" as shorthand for the classical individually accepted-deviation test described in Remark B.4, and exhibit one in each partition.

(i) $\pi = \{\{1, 2, 3\}\}$. Player 1 deviates to the singleton $\{1\}$, comparing $\{1, 2, 3\}$ vs. $\{1\}$ under $U = \{1, 2, 3\}$:

$$\|\boldsymbol{\vartheta}_{123} - \boldsymbol{\theta}_1^*(123)\|_2^2 = \frac{40001}{91809}, \qquad \|\boldsymbol{\vartheta}_{\{1\}} - \boldsymbol{\theta}_1^*(123)\|_2^2 = \frac{10004}{91809}.$$

Since $10004 < 40001$, we have $\{1\} \succ_1 \{1, 2, 3\}$, so π is not IS.

(ii) $\pi = \{\{1\}, \{2, 3\}\}$. Player 2 proposes to join player 1, forming $\{1, 2\}$. The strict-improvement condition compares $\{1, 2\}$ vs. $\{2, 3\}$ for player 2 under $U = \{1, 2, 3\}$:

$$\|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_2^*(123)\|_2^2 = \frac{1061993}{367236}, \qquad \|\boldsymbol{\vartheta}_{23} - \boldsymbol{\theta}_2^*(123)\|_2^2 = \frac{1451045}{367236}.$$

Since $1061993 < 1451045$, player 2 strictly prefers $\{1, 2\}$. For acceptance, player 1 compares $\{1, 2\}$ vs. $\{1\}$ under $U = \{1, 2\}$:

$$\|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_1^*(12)\|_2^2 = \frac{2504}{10201}, \qquad \|\boldsymbol{\vartheta}_{\{1\}} - \boldsymbol{\theta}_1^*(12)\|_2^2 = \frac{160001}{40804}.$$

Since $2504/10201 < 160001/40804$, player 1 weakly prefers $\{1, 2\}$ and accepts. Hence π is not IS.

(iii) $\pi = \{\{1, 2\}, \{3\}\}$. Player 3 proposes to join $\{1, 2\}$, forming $\{1, 2, 3\}$. Under $U = \{1, 2, 3\}$:

$$\|\boldsymbol{\vartheta}_{123} - \boldsymbol{\theta}_3^*(123)\|_2^2 = \frac{490100}{91809}, \qquad \|\boldsymbol{\vartheta}_{\{3\}} - \boldsymbol{\theta}_3^*(123)\|_2^2 = \frac{1000049}{91809}.$$

Player 3 strictly prefers $\{1, 2, 3\}$. For acceptance, both incumbents must weakly prefer $\{1, 2, 3\}$ over $\{1, 2\}$, also under $U = \{1, 2, 3\}$:

$$\|\boldsymbol{\vartheta}_{123} - \boldsymbol{\theta}_1^*(123)\|_2^2 = \frac{40001}{91809} < \frac{1118393}{367236} = \|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_1^*(123)\|_2^2,$$

$$\|\boldsymbol{\vartheta}_{123} - \boldsymbol{\theta}_2^*(123)\|_2^2 = \frac{250121}{91809} < \frac{1061993}{367236} = \|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_2^*(123)\|_2^2.$$

Both incumbents accept. Hence π is not IS.

(iv) $\pi = \{\{1, 3\}, \{2\}\}$. Player 1 proposes to leave $\{1, 3\}$ and join player 2, forming $\{1, 2\}$. Under $U = \{1, 2, 3\}$:

$$\|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_1^*(123)\|_2^2 = \frac{1118393}{367236} < \frac{2048906}{367236} = \|\boldsymbol{\vartheta}_{13} - \boldsymbol{\theta}_1^*(123)\|_2^2.$$

Player 1 strictly prefers $\{1, 2\}$. Under $U = \{1, 2\}$, player 2 compares $\{1, 2\}$ vs. $\{2\}$ and accepts (similar to case (ii)). Hence π is not IS.

(v) $\pi = \{\{1\}, \{2\}, \{3\}\}$. Player 1 can deviate by joining player 2, forming $\{1, 2\}$. The deviator's strict improvement compares $\{1, 2\}$ vs. $\{1\}$ for player 1 under $U = \{1, 2\}$, which is the acceptance check computed in case (ii):

$$\|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_1^*(12)\|_2^2 = \tfrac{2504}{10201} < \tfrac{160001}{40804} = \|\boldsymbol{\vartheta}_{\{1\}} - \boldsymbol{\theta}_1^*(12)\|_2^2,$$

so $\{1,2\} \succ_1 \{1\}$. For acceptance, player 2 compares $\{1,2\}$ vs. $\{2\}$ under $U = \{1,2\}$. Substituting $\boldsymbol{\theta}_2^*(12) = (807/202,\ -103/101)^\top$, $\boldsymbol{\vartheta}_{12} = (7/2,\ -1)^\top$, and $\boldsymbol{\vartheta}_{\{2\}} = (4,\ -3)^\top$ into the comparison rule yields

$$\|\boldsymbol{\vartheta}_{12} - \boldsymbol{\theta}_2^*(12)\|_2^2 = \tfrac{2504}{10201} < \tfrac{160001}{40804} = \|\boldsymbol{\vartheta}_{\{2\}} - \boldsymbol{\theta}_2^*(12)\|_2^2,$$

i.e. the same fractions as for player 1, by the symmetry between players 1 and 2 in the two-player coalition $\{1,2\}$ under the shared anisotropic noise. Hence $\{1,2\} \succeq_2 \{2\}$ and player 2 accepts the deviation, so π is not IS.

Since every partition of $\{1,2,3\}$ admits an accepted unilateral deviation, no partition is free of such deviations. The ten required proposed-member risk reductions in the displayed checks are positive; their minimum is $203/1212$. Each reduction is a continuous function of the six observed coordinates. A finite intersection of neighborhoods preserves all ten strict inequalities, so the same moves defeat all five partitions on an open set. The joint Gaussian sampling density is positive everywhere for every fixed $\boldsymbol{\mu}$, which proves the positive-probability assertion. □

The five cases cover every partition of the three players, so no partition in this instance is free of an accepted unilateral deviation.

Corollary B.31 (Arbitrary Anisotropy Can Defeat the No-Deviation Guarantee)**.** *The noise isotropy assumption ($E_j = \varepsilon_j^2 I_d$) in Assumption A.1 cannot be replaced by unrestricted anisotropic E_j while retaining the no-deviation guarantee of Theorem B.10, even when the prior is isotropic and the noise covariance is identical across clients. Formally, there exists an instance satisfying Assumption B.24 with $\Sigma = I_d$ and $E_1 = E_2 = \cdots = E_M$ for which every partition admits an accepted unilateral deviation. The instance lies outside the shared-shape regime of Corollary B.25, since its E is not proportional to $\Sigma = I_2$.*

Proof. Proposition B.30 constructs such an instance with $M = 3$, $d = 2$, $\Sigma = I_2$, and $E_j = \mathrm{diag}(1/100,\ 100)$ for all j. The instance is symmetric (spherical prior, identical noise across players), yet every partition admits an accepted unilateral deviation. Any partition satisfying Definition B.3 has no individually accepted deviation by Remark B.4. The counterexample therefore rules out the no-deviation guarantee of Theorem B.10 in this anisotropic setting. A covariance condition cannot be removed altogether. □

The preceding counterexample shows that every partition can admit a deviation. We now identify the algebraic cause and the proposition's scope.

B.12.1 Mechanism and Scope of the Anisotropy Boundary

The following two remarks clarify the structural cause of the failure and the precise scope of Corollary B.31.

Remark B.32 (Conflicting Confidence Classifications Across Coordinates)**.** *The failure in Proposition B.30 has a specific algebraic cause. With $E_j = \mathrm{diag}(1/100,\ 100)$ and $\Sigma = I_2$, every player has $B_{j1} = 1/101 \approx 0.01$ (strongly confident along coordinate 1: pooling* increases *MSE) and $B_{j2} = 100/101 \approx 0.99$ (strongly unconfident along coordinate 2: acting alone* increases *MSE). No single assignment is coordinatewise preferred in both directions: pooling improves coordinate 2 while worsening coordinate 1, whereas acting alone does the reverse. The magnitudes of each player's gain and loss depend on the specific coalition mean, which in turn depends on which other players are present. For the data realization in Proposition B.30, those gains and losses are incompatible across all five candidate partitions.*

More generally, such a failure can arise when the prior-to-noise ratio $\sigma_k^2/(\varepsilon_{jk}^2/n_j)$ varies substantially across coordinates. Each coordinate can then prescribe a different coalition recommendation, and the ℓ_2 loss that enters the comparison rule of Definition B.5 aggregates these coordinate-wise costs without any mechanism for resolving conflicts between them.

Remark B.33 (Scope of the Non-Redundancy Claim)**.** *Corollary B.31 establishes that unrestricted anisotropy can defeat the no-deviation property; it does not assert that every anisotropic instance does so. Theorem B.27 proves a broader positive statement whenever all directions favor the same route, with shared shape as an immediate special case. The matrix switch in the proof describes the*

corresponding client-level picture: the theorem certifies a common scalar route while all directions remain on the same side of $1/2$. Directional conflict lies outside that guarantee and can, but need not, make a scalar route insufficient. The scalar algorithm does not detect a mixed-direction client.

Together, the theorem and counterexample identify a positive region and a failure mode: direction agreement guarantees the no-deviation property, while directional conflict can destroy it. Intermediate covariance families remain unresolved. The implemented algorithm retains its scalar score and cannot detect a mixed-direction client. Section C next connects the scalar construction to learning actions. Section F examines estimation error and covariance conflict using the existing simulations.

C Conditional Utility for Local and Common Actions

C.1 The Exact Local-versus-Joint Loss Switch

The hierarchy implies the posterior identity used in the action comparison.

Lemma C.1 (Common-Checkpoint Posterior Mean)**.** *Under Assumption A.1, for every $j \in C$,*

$$\boldsymbol{m}_j(C) := \mathbb{E}[\boldsymbol{\theta}_j \mid \widehat{\boldsymbol{\theta}}_C] = (1 - B_j)\widehat{\boldsymbol{\theta}}_j + B_j \boldsymbol{\vartheta}_C.$$

Proof. Normal–normal conjugacy gives $\mathbb{E}[\boldsymbol{\theta}_j \mid \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}_C] = (1 - B_j)\widehat{\boldsymbol{\theta}}_j + B_j\boldsymbol{\mu}$ and $\mathbb{E}[\boldsymbol{\mu} \mid \widehat{\boldsymbol{\theta}}_C] = \boldsymbol{\vartheta}_C$. Iterated expectation yields the claim. The full posterior distribution and variance are derived in Theorem A.7. □

FedBHG uses this mean to compare two fixed training actions from one common checkpoint. Theorem C.3 states that, under shared local quadratic geometry, the conditional joint-minus-local loss equals the standard posterior-risk switch in the H^{-1} geometry. Unlike a comparison with an unchanged checkpoint, this action-to-action identity does not require the checkpoint to be stationary.

For a positive-definite matrix H, write $\|\boldsymbol{z}\|^2_{H^{-1}} := \boldsymbol{z}^\top H^{-1}\boldsymbol{z}$.

Assumption C.2 (Shared Local Quadratic Geometry)**.** *Fix a client set C, a common checkpoint $\boldsymbol{w}^*$, and one deterministic matrix $H \succ 0$. Conditional on the hierarchy in Assumption A.1, every realization of every client $j \in C$ has, on every local or coalition action under comparison, the population-loss form*

$$L_j(\boldsymbol{w}^* + \boldsymbol{u}) = L_j(\boldsymbol{w}^*) + \boldsymbol{\theta}_j^\top \boldsymbol{u} + \tfrac{1}{2}\boldsymbol{u}^\top H \boldsymbol{u}.$$

The same curvature H is shared by all clients and all actions in the stated client set. It need not be the global Hessian of a neural loss. For a deviation theorem over the full population, the assumption is imposed with $C = \mathcal{M}$.

For an update estimate $\boldsymbol{a}$, define its curvature-matched action by

$$\boldsymbol{w}(\boldsymbol{a}) := \boldsymbol{w}^* - H^{-1}\boldsymbol{a}.$$

Client j's local action is $\boldsymbol{w}(\widehat{\boldsymbol{\theta}}_j)$, while coalition C uses the common action $\boldsymbol{w}(\boldsymbol{\vartheta}_C)$.

Relation to ordinary mean estimation. At any action covered by the shared quadratic assumption,

$$L_j(\boldsymbol{w}^* - H^{-1}a) = K_j + \tfrac{1}{2}\|a - \boldsymbol{\theta}_j\|^2_{H^{-1}},$$

where K_j is independent of a. Thus, under shared quadratic geometry, comparing model updates is equivalent to comparing gradient estimates under H^{-1}-weighted squared error. The Gaussian posterior and the endpoint comparison below use established shrinkage algebra (Röver, 2020; Ozkara et al., 2023); their role is to specify the learning utilities used by the participation model.

Theorem C.3 (Exact Local-versus-Joint Loss Switch)**.** *Under Assumptions A.1 and C.2, for every $j \in C$,*

$$\begin{aligned}\mathbb{E}\Big[L_j\big(\boldsymbol{w}(\boldsymbol{\vartheta}_C)\big) - L_j\big(\boldsymbol{w}(\widehat{\boldsymbol{\theta}}_j)\big) \,\Big|\, \widehat{\boldsymbol{\theta}}_C\Big] \\ = \big(\tfrac{1}{2} - B_j\big)\,\|\boldsymbol{\vartheta}_C - \widehat{\boldsymbol{\theta}}_j\|^2_{H^{-1}}.\end{aligned} \tag{18}$$

Unless the two update estimates coincide, the local action has lower conditional expected loss if and only if $B_j < 1/2$, and the joint action has lower conditional expected loss if and only if $B_j > 1/2$.

Proof. For any update estimate $\boldsymbol{a}$ measurable with respect to $\widehat{\boldsymbol{\theta}}_C$, Assumption C.2 gives

$$\begin{aligned}&\mathbb{E}\Big[L_j\big(\boldsymbol{w}(\boldsymbol{a})\big)\,\Big|\,\widehat{\boldsymbol{\theta}}_C\Big]\\&\quad= K_j(C) - \boldsymbol{m}_j(C)^\top H^{-1}\boldsymbol{a} + \tfrac{1}{2}\boldsymbol{a}^\top H^{-1}\boldsymbol{a},\end{aligned}$$

where $K_j(C)$ does not depend on $\boldsymbol{a}$ and, by Lemma C.1,

$$\boldsymbol{m}_j(C) := \mathbb{E}[\boldsymbol{\theta}_j \mid \widehat{\boldsymbol{\theta}}_C] = (1-B_j)\widehat{\boldsymbol{\theta}}_j + B_j\boldsymbol{\vartheta}_C.$$

Set $\boldsymbol{y} = \widehat{\boldsymbol{\theta}}_j$, $\boldsymbol{q} = \boldsymbol{\vartheta}_C$, and $\boldsymbol{d} = \boldsymbol{q} - \boldsymbol{y}$. Then $\boldsymbol{m}_j(C) = \boldsymbol{y} + B_j\boldsymbol{d}$. Subtracting the expression for $\boldsymbol{a} = \boldsymbol{y}$ from that for $\boldsymbol{a} = \boldsymbol{q}$ yields

$$\begin{aligned}&-(\boldsymbol{y}+B_j\boldsymbol{d})^\top H^{-1}\boldsymbol{d} + \tfrac{1}{2}\big((\boldsymbol{y}+\boldsymbol{d})^\top H^{-1}(\boldsymbol{y}+\boldsymbol{d}) - \boldsymbol{y}^\top H^{-1}\boldsymbol{y}\big)\\&\quad= (\tfrac{1}{2}-B_j)\boldsymbol{d}^\top H^{-1}\boldsymbol{d},\end{aligned}$$

which is (18). □

Remark C.4 (The common-action restriction)**.** *For $y = \widehat{\boldsymbol{\theta}}_j$, $q = \boldsymbol{\vartheta}_C$, and $m = (1-B_j)y + B_jq$, the quadratic surrogate risk is minimized at m. The smaller endpoint risk exceeds that minimum by*

$$\tfrac{1}{2}\min\{B_j^2, (1-B_j)^2\}\|q-y\|_{H^{-1}}^2.$$

For distinct endpoints and positive variance components this is strictly positive. The same statement holds for actual loss only if the quadratic assumption also covers $\boldsymbol{w}(m)$. A coalition in our model commits to one common action, whereas these posterior-optimal actions generally differ across clients. The participation theorem therefore concerns that deployment constraint, not optimality among personalized estimators. The midpoint structure also appears in ex-ante optimal model averaging (Grimberg et al., 2021).

Interpretation. The score has a direct decision meaning among the two fixed curvature-matched actions under the theorem assumptions. When the sample-mean gradient is reliable relative to between-client variation, the local direction has lower conditional expected loss; when sampling uncertainty dominates, the precision-weighted coalition direction does. The identity does not say that either action must improve on the unchanged checkpoint.

C.2 Expected Loss and Coalition Comparisons

The loss identity extends beyond the local-versus-joint pair. Let $S, T \ni j$ be two coalitions, let $U = S \cup T$, and condition on $\widehat{\boldsymbol{\theta}}_U$. Completing the square expresses expected loss as a posterior risk in the curvature-weighted geometry.

Proposition C.5 (Expected Loss as Curvature-Weighted Posterior Risk)**.** *Under Assumptions A.1 and C.2, let $\boldsymbol{a}_S$ and $\boldsymbol{a}_T$ be two common update estimates measurable with respect to $\widehat{\boldsymbol{\theta}}_U$. Then*

$$\begin{aligned}&\mathbb{E}\Big[L_j\big(\boldsymbol{w}(\boldsymbol{a}_S)\big) - L_j\big(\boldsymbol{w}(\boldsymbol{a}_T)\big)\,\Big|\,\widehat{\boldsymbol{\theta}}_U\Big]\\&\quad= \tfrac{1}{2}\left(\|\boldsymbol{a}_S - \boldsymbol{m}_j(U)\|_{H^{-1}}^2 - \|\boldsymbol{a}_T - \boldsymbol{m}_j(U)\|_{H^{-1}}^2\right),\end{aligned} \tag{19}$$

where $\boldsymbol{m}_j(U) = (1-B_j)\widehat{\boldsymbol{\theta}}_j + B_j\boldsymbol{\vartheta}_U$.

Proof. For either action $\boldsymbol{a}$,

$$-\boldsymbol{m}_j(U)^\top H^{-1}\boldsymbol{a} + \tfrac{1}{2}\boldsymbol{a}^\top H^{-1}\boldsymbol{a} = \tfrac{1}{2}\|\boldsymbol{a} - \boldsymbol{m}_j(U)\|_{H^{-1}}^2 - \tfrac{1}{2}\|\boldsymbol{m}_j(U)\|_{H^{-1}}^2.$$

All action-independent terms cancel in the comparison. □

Proposition C.5 connects the action comparison to the voluntary-sharing theorems in Section B. Its H^{-1} norm is generally not the Euclidean posterior MSE in Definition B.5; the two rankings can differ when H is anisotropic. The loss no-deviation theorem therefore reruns the coalition argument in one common H^{-1} inner product rather than claiming that the Euclidean preferences are unchanged.

Theorem C.6 (Expected-Loss No-Deviation)**.** *Retain the setting of Theorem B.10 and additionally impose Assumption C.2 with $C = \mathcal{M}$: one deterministic $H \succ 0$ is shared by every client and every candidate coalition action. Suppose every coalition C deploys the curvature-matched common action $\boldsymbol{w}(\boldsymbol{\vartheta}_C)$, while a singleton deploys $\boldsymbol{w}(\widehat{\boldsymbol{\theta}}_j)$. Compare two coalitions by conditional expected post-update client loss, conditioning on the union of their observations as in Definition B.5. Then, almost surely in the sense of Remark A.3, the confidence partition Π_c admits no coalition that weakly lowers every member's expected post-update loss and strictly lowers at least one member's loss.*

Proof. By Proposition C.5, each expected-loss comparison equals one half of the difference between squared distances to the same posterior mean in the H^{-1} inner product. The proof of Theorem B.10 uses only inner-product expansions, the scalar identity $1 - B_i = \sigma^2\alpha_i$, and linearity of the precision-weighted coalition mean. It therefore applies verbatim after replacing the Euclidean inner product by $\langle \boldsymbol{x}, \boldsymbol{z}\rangle_{H^{-1}} = \boldsymbol{x}^\top H^{-1}\boldsymbol{z}$. Equivalently, apply the invertible change of coordinates $\boldsymbol{a} \mapsto H^{-1/2}\boldsymbol{a}$ and invoke the same argument. Thus no proposed coalition satisfies the stated expected-loss deviation test. □

C.3 Beyond the Exact Quadratic Boundary

The quadratic local-versus-joint decision is preserved under locally nonquadratic loss when its margin dominates the Taylor remainder.

Theorem C.7 (Nonquadratic Remainder Bound)**.** *Retain Assumption A.1. Suppose $\nabla^2 L_j(\boldsymbol{w}^*) = H \succ 0$ for every posterior realization and that one deterministic constant ρ_j uniformly bounds the Hessian Lipschitz constant in operator norm over every posterior realization and over both the local and coalition update segments. Then*

$$\begin{aligned}&\mathbb{E}\Big[L_j\big(\boldsymbol{w}(\boldsymbol{\vartheta}_C)\big) - L_j\big(\boldsymbol{w}(\widehat{\boldsymbol{\theta}}_j)\big)\,\Big|\,\widehat{\boldsymbol{\theta}}_C\Big]\\ &\quad = (\tfrac{1}{2} - B_j)\|\boldsymbol{\vartheta}_C - \widehat{\boldsymbol{\theta}}_j\|^2_{H^{-1}} + R_j(C),\end{aligned} \tag{20}$$

where

$$|R_j(C)| \le \frac{\rho_j}{6}\left(\|H^{-1}\boldsymbol{\vartheta}_C\|_2^3 + \|H^{-1}\widehat{\boldsymbol{\theta}}_j\|_2^3\right). \tag{21}$$

Consequently, the half-threshold action remains lower-loss whenever the absolute quadratic margin in (20) *exceeds the right-hand side of* (21)*.*

Proof. Third-order Taylor expansion around $\boldsymbol{w}^*$ gives, for either update estimate $\boldsymbol{a}$,

$$L_j\big(\boldsymbol{w}(\boldsymbol{a})\big) = L_j(\boldsymbol{w}^*) - \boldsymbol{\theta}_j^\top H^{-1}\boldsymbol{a} + \tfrac{1}{2}\boldsymbol{a}^\top H^{-1}\boldsymbol{a} + r_j(\boldsymbol{a}),$$

with $|r_j(\boldsymbol{a})| \le \rho_j\|H^{-1}\boldsymbol{a}\|_2^3/6$. The quadratic terms reduce as in Theorem C.3; the triangle inequality bounds the difference of the two remainders. □

C.4 Plug-In Margins and Action Error

No uniform plug-in guarantee is possible at the threshold: an arbitrarily small score error can change the action of a client whose oracle score is arbitrarily close to $1/2$. Any plug-in guarantee must therefore depend on the route margin.

Proposition C.8 (Deterministic Plug-In Route and Action Sensitivity)**.** *If*

$$\max_j |\widehat{B}_j - B_j| < \min_j |B_j - \tfrac{1}{2}|, \tag{22}$$

then every plug-in route agrees with its oracle route.

Under Assumption C.2, fix the observed $\widehat{\boldsymbol{\theta}}_C$ and define the oracle conditional-loss functional

$$\mathcal{G}_{j,C}(\boldsymbol{a}) := \mathbb{E}[L_j(\boldsymbol{w}(\boldsymbol{a})) - L_j(\boldsymbol{w}(\widehat{\boldsymbol{\theta}}_j)) \mid \widehat{\boldsymbol{\theta}}_C].$$

For any realized perturbation $\boldsymbol{e}_C$, let the implemented coalition vector be $\widehat{\boldsymbol{\vartheta}}_C = \boldsymbol{\vartheta}_C + \boldsymbol{e}_C$. Define

$$D_{jC} := \|\boldsymbol{\vartheta}_C - \widehat{\boldsymbol{\theta}}_j\|_{H^{-1}}, \qquad E_C := \|\boldsymbol{e}_C\|_{H^{-1}}.$$

The value of the oracle conditional-loss functional at the perturbed vector differs from its value at $\boldsymbol{\vartheta}_C$ by at most

$$(1 - B_j)D_{jC}E_C + \tfrac{1}{2}E_C^2. \tag{23}$$

Thus this pointwise perturbation of the oracle action preserves the oracle preference when its absolute margin exceeds (23)*. This is a sensitivity statement, not a posterior theorem after conditioning on the extra summaries used to estimate $\boldsymbol{e}_C$.*

Proof. Condition (22) prevents any score from crossing $1/2$. For the action statement, write $\boldsymbol{d} = \boldsymbol{\vartheta}_C - \widehat{\boldsymbol{\theta}}_j$. Direct expansion gives

$$\begin{aligned}&\mathcal{G}_{j,C}(\widehat{\boldsymbol{\vartheta}}_C)\\ &\quad = (\tfrac{1}{2} - B_j)\|\boldsymbol{d}\|_{H^{-1}}^2 + (1 - B_j)\langle \boldsymbol{d}, \boldsymbol{e}_C\rangle_{H^{-1}} + \tfrac{1}{2}\|\boldsymbol{e}_C\|_{H^{-1}}^2.\end{aligned}$$

Cauchy–Schwarz yields (23). □

Remark C.9 (Near-threshold counterexample)**.** *Take $d = 1$, $H = 1$, $\widehat{\theta}_j = 0$, $\vartheta_C = 1$, and $B_j = 1/2 + \delta$. The joint action is better by δ, but $\widehat{B}_j = 1/2 - \delta$ selects local. The required score error 2δ tends to zero with δ.*

Remark C.10 (Continuation-horizon counterexample)**.** *Take $d = 1$, $H = 1$, local estimate 0, coalition estimate 1, and posterior mean $B = 0.2$. If both optimization branches have moved only a fraction c toward their empirical quadratic optima, the joint-minus-local expected loss gap is $c(c/2 - B)$. It is negative at $c = 0.1$ but positive at $c = 1$. Thus step scale and continuation horizon can reverse a decision outside the curvature-matched action in Theorem C.3. This example does not assert that reversal must occur in neural training; it shows why multi-round behavior requires empirical evaluation.*

Claim boundary. Theorem C.3 is exact for one curvature-matched comparison at a common checkpoint and a fixed coalition under its stated information set. A coalition selected from plug-in summaries outside that information set is not covered. Theorem C.7 is local and margin-dependent, and Proposition C.8 is deterministic rather than a concentration theorem for the profile estimator. Arbitrary multi-round stochastic accuracy remains an empirical question for Section G. Section F instead tests finite-sample and covariance boundaries of the no-deviation guarantee.

D Operationalizing the Checkpoint Decision

Theorem C.3 derives a checkpoint-conditioned preference between one local action and one coalition-common action. FedBHG (Federated Bayesian Hedonic Game) turns that model into an estimate-once, route-once policy: estimate the reliability score from summaries of the clients' realized data, choose which continuation each client will enter, and train only the selected branches. Low-score clients continue independently; high-score clients share one continuation. The route and checkpoint-derived aggregation weights remain fixed after training resumes. Section E contains the full variance-estimation derivation.

D.1 Gradient Summaries at a Trained Model

Let $\mathcal{M}$ contain M clients. Client j holds a local dataset D_j of size n_j, sampled from a client distribution $\mathcal{P}_j$. For per-example loss $\ell_x(\boldsymbol{w})$, client j's population risk is

$$L_j(\boldsymbol{w}) := \mathbb{E}_{x\sim\mathcal{P}_j}[\ell_x(\boldsymbol{w})].$$

An upstream pooled or clustered federated method produces a checkpoint $\boldsymbol{w}^*$. Every client in one fitted population evaluates its gradients at the same parameter vector:

$$\widehat{\boldsymbol{\theta}}_j = \frac{1}{n_j}\sum_{x\in D_j}\nabla\ell_x(\boldsymbol{w}^*), \qquad \mathsf{SSE}_j = \sum_{x\in D_j}\|\nabla\ell_x(\boldsymbol{w}^*) - \widehat{\boldsymbol{\theta}}_j\|_2^2.$$

The server receives $(\widehat{\boldsymbol{\theta}}_j, \mathsf{SSE}_j, n_j)$, estimates the variance components, and computes

$$\widehat{B}_j = \frac{\widehat{\varepsilon}_j^2/n_j}{\widehat{\varepsilon}_j^2/n_j + \widehat{\sigma}^2}, \qquad \widehat{\alpha}_j = \frac{1}{\widehat{\sigma}^2 + \widehat{\varepsilon}_j^2/n_j}.$$

The one-time routing pass communicates $O(d)$ values per client. Raw examples and per-example gradients remain local.

D.2 The Two Continuation Actions

The plug-in route forms

$$C_{\text{local}} := \{j : \widehat{B}_j < 1/2\}, \qquad C_{\text{joint}} := \{j : \widehat{B}_j \geq 1/2\}.$$

Each client in C_{local} starts from $\boldsymbol{w}^*$ and follows an independent local optimizer on D_j. Clients in C_{joint} also restart from $\boldsymbol{w}^*$, but train one common model. Their fixed normalized aggregation weights are

$$\widehat{p}_j = \frac{\widehat{\alpha}_j}{\sum_{i \in C_{\text{joint}}} \widehat{\alpha}_i}, \qquad j \in C_{\text{joint}}. \tag{24}$$

In each joint round, members begin from the same current group model, perform the same number of local epochs over their own training iterators, and the server averages their updates using (24). The weights are not recomputed from updates evaluated at different client parameters.

For the shared quadratic loss, minimizing its local expansion with gradient estimate a gives the Newton step $-H^{-1}a$. If the two continuation routines apply these curvature-matched actions $-H^{-1}\widehat{\boldsymbol{\theta}}_j$ and $-H^{-1}\boldsymbol{\vartheta}_C$, Theorem C.3 compares them. The default implementation instead uses ordinary local optimizers and federated aggregation; neither its first step nor its later steps are claimed to equal that action. Therefore the theorem does not certify the SGD/Adam continuation itself; whether the routing rule remains useful after ordinary multi-round training must be tested empirically. We consequently evaluate multiple training horizons in the empirical study; the synthetic study instead isolates the effects of finite estimation and covariance structure on coalition stability.

D.3 The FedBHG Algorithm

If C_{joint} has one member, the joint branch degenerates to that client's single-client continuation; the one-step local and coalition actions coincide. This edge case is retained rather than silently converted into checkpoint retention. A singleton base cluster has no estimable between-client dispersion and no routing choice, so Algorithm 1 sends it directly to local continuation.

D.4 Post-Clustering Use

Given a partition $\{C_1, \ldots, C_K\}$ produced by an upstream method such as CFL or IFCA, FedBHG operates independently inside each base cluster. It collects summaries at that cluster's checkpoint, estimates one $\widehat{\sigma}^2_{C_k}$, partitions the clients, and trains at most one joint continuation for the high-score group. Different base clusters never share variance estimates or post-routing models.

This plug-in procedure observes more than the fixed-coalition theorem: both $\widehat{\sigma}^2_{C_k}$ and membership use summaries from the full base cluster, including clients later routed locally. The all-data example in Section B.7 shows that information-set choices can matter. Accordingly, the fitted group and its optimized continuation are empirical objects, not an oracle posterior guarantee in disguise.

D.5 Why the Route and Weights Are Fixed

The common checkpoint places every gradient summary at the same reference, which is required to interpret their dispersion as the between-client variance σ^2. After clients move to different parameters, update dispersion mixes client heterogeneity with curvature and path-dependent optimization noise. FedBHG therefore does not re-estimate σ^2 from later client updates. It carries the initial precision weights into the joint branch as a fixed design choice.

Algorithm 1: FedBHG Local-versus-Joint Continuation

Input: Common checkpoint $\boldsymbol{w}^*$; client set C with datasets $(D_j : j \in C)$; continuation budget T
Output: Partition $\Pi = \{\{j\} : j \in C_{\text{local}}\} \cup (\{C_{\text{joint}}\} \text{ if } C_{\text{joint}} \neq \emptyset)$; per-client models $(\boldsymbol{w}_j : j \in C)$

if $|C| = 1$ **then**
  $\boldsymbol{w}_j \leftarrow \textsc{LocalContinue}(\boldsymbol{w}^*, D_j, T)$ for the unique $j \in C$
  return $\{\{j\}\}$ and $\boldsymbol{w}_j$
end

`// Phase 1: common-checkpoint summaries`
for *each client* $j \in C$ **do**
  Compute $\widehat{\boldsymbol{\theta}}_j$ and SSE_j at the identical $\boldsymbol{w}^*$
  Transmit $(\widehat{\boldsymbol{\theta}}_j, \mathsf{SSE}_j, n_j)$
end

`// Phase 2: fit once and route once`
$\widehat{\varepsilon}_j^2 \leftarrow \mathsf{SSE}_j/[d(n_j - 1)]$ for every j with $n_j \geq 2$
For $n_j = 1$, use the pooled or shrunk within-client variance convention of Remark E.3
$\widehat{\sigma}^2 \leftarrow \textsc{ProfileMLE}((\widehat{\boldsymbol{\theta}}_j, \widehat{\varepsilon}_j^2/n_j)_{j \in C})$
$\widehat{\alpha}_j \leftarrow (\widehat{\sigma}^2 + \widehat{\varepsilon}_j^2/n_j)^{-1}$ for every j
$\widehat{B}_j \leftarrow (\widehat{\varepsilon}_j^2/n_j)/(\widehat{\varepsilon}_j^2/n_j + \widehat{\sigma}^2)$ for every j
$C_{\text{local}} \leftarrow \{j : \widehat{B}_j < 1/2\}$
$C_{\text{joint}} \leftarrow C \setminus C_{\text{local}}$

`// Phase 3a: independent continuation`
for *each* $j \in C_{\text{local}}$ **do**
  $\boldsymbol{w}_j \leftarrow \textsc{LocalContinue}(\boldsymbol{w}^*, D_j, T)$
end

`// Phase 3b: one separately trained joint continuation`
if $C_{\text{joint}} \neq \emptyset$ **then**
  $\widehat{p}_j \leftarrow \widehat{\alpha}_j / \sum_{i \in C_{\text{joint}}} \widehat{\alpha}_i$ for $j \in C_{\text{joint}}$
  $\boldsymbol{w}_{\text{joint}} \leftarrow \textsc{JointContinue}(\boldsymbol{w}^*, (D_j, \widehat{p}_j)_{j \in C_{\text{joint}}}, T)$
  $\boldsymbol{w}_j \leftarrow \boldsymbol{w}_{\text{joint}}$ for every $j \in C_{\text{joint}}$
end

return Π and $(\boldsymbol{w}_j : j \in C)$

Algorithm 1 is the deployed FedBHG procedure: local clients train separately, and joint clients train together with the fixed precision weights in (24). Each distinct routed coalition in an experiment therefore receives its own branch from the same checkpoint. The next section derives the one-time estimator used to produce the route and fixed weights. The coalition theorems concern the corresponding oracle actions and do not certify the fitted plug-in procedure.

E Estimating the Gradient-Reliability Score

Section D stated the profile-likelihood estimator used by the router. This section derives it and specifies its statistical scope. The theoretical score depends on scalar variance parameters σ^2 and $\{\varepsilon_j^2\}_{j \in \mathcal{M}}$, which in practice must be estimated from client-level gradient data. The implemented path has three steps: estimate each within-client trace variance, profile out the population mean and deterministically maximize the resulting one-dimensional likelihood for σ^2, then form $\widehat{B}_j$ and the precision weights. We first justify the scalar covariance working model (Section E.1), derive this procedure (Section E.2), and record exact moments for its untruncated moment components (Section E.3). The later EM derivation is an equivalent likelihood characterization, not a second default router. We finish by situating the estimator within empirical Bayes (Section E.4) and stating its misspecification boundaries (Section E.5).

E.1 Why the Implemented Score Uses Scalar Covariances

The scalar hierarchical mean-gradient model (Assumption A.1) uses isotropic covariances $E_j = \varepsilon_j^2 I_d$ and $\Sigma = \sigma^2 I_d$. Before deriving the estimation procedure, we explain why this is a tractable working approximation in the federated gradient setting.

Theoretical simplification. Under shared-shape covariances, the same scalar score B_j applies in every direction, so the common shape matrix cancels from the routing decision. The general theorem in Section B.11.3 is broader, but it requires matrix-valued covariance information. The implemented scalar score therefore targets the shared-shape approximation; it is a computationally light sufficient model, not a necessary condition for stable partitions to exist.

Computational motivation. A full $d \times d$ covariance matrix is infeasible at the neural scale considered here. For a model with $d = 10^6$ parameters, the covariance matrix has 10^{12} entries; at 32-bit precision this requires approximately 4 TB of memory per client (about 2 TB if only one triangle is stored). Even a low-rank approximation of rank r costs $O(dr)$ storage per client and $O(dr^2)$ computation per round, which is prohibitive for $r \gtrsim 10^3$. The scalar model reduces covariance storage to one real number per client; the client mean gradient is still a d-vector, as in Section D.

Choice of normalization. Under the shared-shape model, the trace of the sample covariance matrix satisfies $\mathrm{trace}(\widehat{K}) = d \cdot s^2$ where s^2 is the scalar sample variance (see Lemma E.8 below). Under the normalization $\mathrm{trace}(K) = d$ (which is consistent with the routing rule depending only on the ratio ε_j^2/σ^2), s^2 is an unbiased estimator of $\sigma^2 + \frac{1}{M}\sum_j \varepsilon_j^2/n_j$, which is the quantity needed for the MoM diagnostic. The oracle score is also invariant to a common shape. These facts do not make the implemented isotropic profile likelihood the true likelihood for arbitrary K: unless $K = I_d$, that likelihood would use K^{-1}-weighted residuals. The profile estimator below is therefore part of the scalar working model.

With the scalar covariance model motivated on both theoretical and computational grounds, we now derive the concrete estimators of σ^2 and $\{\varepsilon_j^2\}$ used by Algorithm 1.

E.2 Trace Variance Estimation and Profile Likelihood

We estimate the variance components used by FedBHG from client-level gradient summaries. The within-client variance estimates $\widehat{\varepsilon}_j^2$ are obtained by an unbiased trace estimator. FedBHG then estimates the between-client variance σ^2 by a deterministic approximate maximization of the marginal likelihood of the observed client means after profiling out $\boldsymbol{\mu}$. The resulting procedure is a one-dimensional optimization with an observable finite search interval. We retain the method-of-moments expression as a diagnostic and derive the equivalent empirical-Bayes EM characterization to clarify which posterior variance belongs to this likelihood.

Let $g_{jk} \in \mathbb{R}^d$ denote the per-sample gradient for client j, sample k, so that

$$\widehat{\boldsymbol{\theta}}_j = n_j^{-1} \sum_{k=1}^{n_j} g_{jk}.$$

Within-client variance. The standard unbiased estimator of ε_j^2 under independent isotropic per-example gradients is

$$\widehat{\varepsilon}_j^2 := \frac{\mathrm{SSE}_j}{d(n_j - 1)}, \qquad \mathrm{SSE}_j := \sum_{k=1}^{n_j} \|g_{jk} - \widehat{\boldsymbol{\theta}}_j\|_2^2.$$

Unbiasedness requires only the covariance and independence conditions. Under the stronger condition that the per-example gradient residuals are conditionally Gaussian, $\mathrm{SSE}_j/\varepsilon_j^2 \sim \chi^2_{d(n_j-1)}$; see Lemma E.7. In the default implementation, these within-client estimates are fixed during the subsequent between-client variance estimation.

Method-of-moments diagnostic. The between-client variance admits the method-of-moments expression, equivalently the one-way ANOVA estimator. Let

$$\bar{\boldsymbol{\theta}} := |C|^{-1} \sum_{j \in C} \widehat{\boldsymbol{\theta}}_j.$$

Define

$$\widehat{\sigma}^2_{\mathrm{ANOVA}} := \left(\frac{1}{d} \cdot \frac{\sum_{j \in C} \|\widehat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}\|_2^2}{|C| - 1} - \frac{1}{|C|} \sum_{j \in C} \frac{\widehat{\varepsilon}_j^2}{n_j} \right)_+,$$

where $(\cdot)_+$ denotes the positive part. Lemma E.6 establishes that the untruncated moment expression is unbiased for σ^2 when $\widehat{\varepsilon}_j^2 = \varepsilon_j^2$. The truncation prevents negative variance estimates, but it can collapse the estimate to zero in nearly homogeneous populations.

FedBHG profile marginal likelihood. FedBHG defines its ideal profile target by maximizing the marginal likelihood of the observed client means after profiling out μ. With $\widehat{\varepsilon}_j^2$ fixed, define

$$\nu_j^2(\sigma^2) := \sigma^2 + \widehat{\varepsilon}_j^2 / n_j.$$

For fixed $\sigma^2 \geq 0$, the maximum-likelihood estimate of the population mean is the precision-weighted average

$$\widehat{\boldsymbol{\mu}}(\sigma^2) := \frac{\sum_{j \in C} \widehat{\boldsymbol{\theta}}_j / \nu_j^2(\sigma^2)}{\sum_{j \in C} 1/\nu_j^2(\sigma^2)}.$$

The profiled marginal log-likelihood, up to constants independent of σ^2, is

$$\ell_{\mathrm{prof}}(\sigma^2) := -\frac{d}{2} \sum_{j \in C} \log \nu_j^2(\sigma^2) - \frac{1}{2} \sum_{j \in C} \frac{\|\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\mu}}(\sigma^2)\|_2^2}{\nu_j^2(\sigma^2)}.$$

The ideal profile estimator is any global maximizer of ℓ_{prof} over $\sigma^2 \geq 0$. This objective accounts for heterogeneous client precisions through $\nu_j^2(\sigma^2)$. Its maximizer may still occur at the boundary $\sigma^2 = 0$ when the data support a nearly homogeneous population, but this boundary solution follows from the profile objective. It is not evidence that the underlying population is homogeneous: the calibration example below shows how profiling can underestimate its variance.

The search interval is bounded. Let

$$D := \max_{i,j \in C} \|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j\|_2.$$

At least one global maximizer of ℓ_{prof} lies in $[0, D^2/d]$. Indeed, $\widehat{\boldsymbol{\mu}}(\sigma^2)$ lies in the convex hull of the observed means, so every residual norm is at most D. Differentiating the profile likelihood then shows that its derivative is strictly negative for $\sigma^2 > D^2/d$. The implementation computes D^2 from the client-summary Gram matrix and performs a deterministic grid search followed by bounded local refinement on $[0, D^2/d]$. Without an additional unimodality certificate, this numerical routine is an approximation to the global profile maximum rather than an exact global optimizer.

E.2.1 Finite-Dimensional Calibration of the Profile Target

The following calculation isolates the effect of profiling the population mean. It concerns the exact maximum of the stated likelihood, so numerical optimization error is not needed for the failure. This is the incidental-mean mechanism of the Neyman–Scott example (Spanos, 2013). The following formulas diagnose the profile target; the evaluated estimator remains unchanged.

Proposition E.1 (Finite-dimensional profile-ML routing law)**.** *Fix $M \geq 2$, $d \geq 1$, a known effective sampling variance $v > 0$, and a true between-client variance $t \geq 0$. For every fixed population mean vector, suppose the coordinates are independent and $y_{jr} \sim \mathcal{N}(\mu_r, t + v)$ independently over clients $j = 1, \ldots, M$. Let*

$$S_d := \sum_{r=1}^{d} \sum_{j=1}^{M} (y_{jr} - \bar{y}_r)^2, \qquad \bar{y}_r := \frac{1}{M} \sum_{j=1}^{M} y_{jr}, \qquad k := d(M-1).$$

Writing F_k for the χ^2_k distribution function, the exact profile target and its common fitted score satisfy

$$\widehat{t}_d = \left(\frac{S_d}{Md} - v\right)_+, \qquad \frac{S_d}{t+v} \sim \chi^2_k, \qquad \widehat{B}_d = \frac{v}{v+\widehat{t}_d}. \tag{25}$$

Under the convention that equality at $1/2$ is routed jointly, all clients take the same route, with

$$p_d := \Pr(\textit{joint route}) = F_k\left(\frac{2Mdv}{t+v}\right), \qquad \Pr(\widehat{t}_d = 0) = F_k\left(\frac{Mdv}{t+v}\right). \tag{26}$$

The probability of disagreeing with the oracle assignment is

$$e_d = \begin{cases} 1 - p_d, & 0 \le t \le v, \\ p_d, & t > v. \end{cases} \tag{27}$$

At $t = v$, the two ideal actions have equal conditional loss, so assignment disagreement is not a loss error.

For fixed M and v, as independent coordinates are added,

$$\widehat{t}_d \longrightarrow t_* := \left(\frac{M-1}{M}(t+v) - v\right)_+ \quad \textit{almost surely}, \tag{28}$$

and

$$\lim_{d\to\infty} e_d = \begin{cases} 0, & 0 \le t \le v, \\ 1, & v < t < \dfrac{M+1}{M-1}v, \\ \frac{1}{2}, & t = \dfrac{M+1}{M-1}v, \\ 0, & t > \dfrac{M+1}{M-1}v. \end{cases} \tag{29}$$

Proof. After profiling the d coordinate means, the log likelihood at a candidate between-client variance $s \ge 0$ is, up to an additive constant,

$$\ell_d(s) = -\frac{Md}{2}\log(s+v) - \frac{S_d}{2(s+v)}.$$

Its derivative is $[S_d - Md(s+v)]/[2(s+v)^2]$, which gives the unique maximum in (25). In each coordinate, centering M independent Gaussian observations removes one degree of freedom; the independent residual sums therefore give $S_d/(t+v) \sim \chi^2_{d(M-1)}$. Since $v > 0$, the joint-route condition $\widehat{B}_d \ge 1/2$ is equivalent to $\widehat{t}_d \le v$, hence to $S_d \le 2Mdv$. The zero-variance condition is $S_d \le Mdv$. These establish (26). The oracle score $v/(t+v)$ is at least $1/2$ exactly when $t \le v$, proving (27).

The coordinate residual sums are independent with common mean $(M-1)(t+v)$, so the strong law proves (28). Thus the fitted variance is eventually below or above v when $t_* < v$ or $t_* > v$, respectively. The equality $t_* = v$ holds at $t = (M+1)v/(M-1)$; there the argument of F_k in the joint-route probability equals k. The central limit theorem for χ^2_k gives $F_k(k) \to 1/2$. Combining these cases proves (29). □

For example, $M = 6$ and $t = 6v/5$ give oracle score $5/11$, but $\widehat{t}_d \to 5v/6$ and $\widehat{B}_d \to 6/11$. Every oracle client is local, while the fitted rule pools all clients with probability tending to one. The increasing number of fitted coordinate means causes the persistent bias. The variance boundary itself is included: $t_* = 0$ whenever $t \le v/(M-1)$, and the second formula in (26) gives the exact finite-dimensional mass at zero.

The existing homoscedastic ANOVA diagnostic instead has target

$$\widetilde{t}_d = \left(\frac{S_d}{d(M-1)} - v\right)_+, \qquad \Pr(\text{joint route under } \widetilde{t}_d) = F_k\left(\frac{2kv}{t+v}\right). \tag{30}$$

The same estimator follows by maximizing the likelihood of the $d(M-1)$ orthonormal error contrasts that remove the coordinate means. It converges almost surely to t and recovers the oracle route in

probability for $t \neq v$. At $t = v$, its assignment-disagreement probability tends to $1/2$, but both ideal actions tie. This comparison identifies a limit of the chosen profile target; it does not replace the estimator used in the reported experiments or establish that one target dominates under heterogeneous fitted variances.

Equivalent empirical-Bayes EM characterization. The same marginal likelihood can alternatively be optimized with the standard EM algorithm for the empirical-Bayes hierarchical model. This characterization is not a second FedBHG estimator; it identifies the likelihood fixed points and the posterior variance term that is valid for that objective. The observed data are the client summaries

$$X := \left(\widehat{\boldsymbol{\theta}}_j, \widehat{\varepsilon}_j^2, n_j\right)_{j \in C},$$

the latent variables are the unobserved client mean updates

$$Z := (\boldsymbol{\theta}_j)_{j \in C},$$

and the parameters estimated by EM are

$$\phi := (\boldsymbol{\mu}, \sigma^2).$$

Thus $\boldsymbol{\mu}$ is treated as a fixed but unknown empirical-Bayes parameter, not as a latent variable. With $\widehat{\varepsilon}_j^2$ fixed, the model is

$$\boldsymbol{\theta}_j \mid \boldsymbol{\mu}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d), \qquad \widehat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}\left(\boldsymbol{\theta}_j, \frac{\widehat{\varepsilon}_j^2}{n_j} I_d\right).$$

The observed-data likelihood of this model is the likelihood whose profile in $\boldsymbol{\mu}$ was displayed above; EM optimizes the same objective over $(\boldsymbol{\mu}, \sigma^2)$ by introducing the latent $(\boldsymbol{\theta}_j)_{j \in C}$.

Given current iterates $(\boldsymbol{\mu}^{(t)}, \sigma_t^2)$, define

$$B_j^{(t)} := \frac{\widehat{\varepsilon}_j^2 / n_j}{\widehat{\varepsilon}_j^2 / n_j + \sigma_t^2}.$$

The conditional posterior of $\boldsymbol{\theta}_j$ given $\widehat{\boldsymbol{\theta}}_j$ and the current parameters is

$$\boldsymbol{\theta}_j \mid \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\mu}^{(t)}, \sigma_t^2 \sim \mathcal{N}\left(\boldsymbol{m}_j^{(t)}, V_j^{(t)} I_d\right),$$

where

$$\boldsymbol{m}_j^{(t)} := B_j^{(t)} \boldsymbol{\mu}^{(t)} + \left(1 - B_j^{(t)}\right) \widehat{\boldsymbol{\theta}}_j, \qquad V_j^{(t)} := B_j^{(t)} \sigma_t^2.$$

The variance $V_j^{(t)}$ is the conditional posterior variance given the current empirical-Bayes mean parameter. It intentionally excludes the additional uncertainty term

$$\left(B_j^{(t)}\right)^2 \left(\sum_{i \in C} \alpha_i^{(t)}\right)^{-1}, \qquad \alpha_i^{(t)} := \left(\sigma_t^2 + \widehat{\varepsilon}_i^2 / n_i\right)^{-1},$$

which appears in the fully hierarchical posterior after marginalizing μ. That fully hierarchical variance is appropriate for posterior-risk comparisons in the theory, but not for the complete-data likelihood maximized by this fixed-parameter EM procedure.

E-step. Let $\mathbb{E}^{(t)}[\cdot]$ denote expectation with respect to the conditional posterior of $\boldsymbol{\theta}_j$ given $(\widehat{\boldsymbol{\theta}}_j, \boldsymbol{\mu}^{(t)}, \sigma_t^2)$. For each $j \in C$, compute

$$\mathbb{E}^{(t)}[\boldsymbol{\theta}_j] = \boldsymbol{m}_j^{(t)}.$$

Moreover, for any candidate value $\boldsymbol{u} \in \mathbb{R}^d$,

$$\mathbb{E}^{(t)}\left[\|\boldsymbol{\theta}_j - \boldsymbol{u}\|_2^2\right] = \|\boldsymbol{m}_j^{(t)} - \boldsymbol{u}\|_2^2 + d V_j^{(t)}.$$

This identity is the quantity used in the M-step, where $\boldsymbol{u}$ is optimized over and then set to $\boldsymbol{\mu}^{(t+1)}$.

M-step. Maximizing the expected complete-data log-likelihood gives

$$\boldsymbol{\mu}^{(t+1)} = \frac{1}{|C|} \sum_{j \in C} \boldsymbol{m}_j^{(t)}$$

and

$$\sigma_{t+1}^2 = \frac{1}{|C|d} \sum_{j \in C} \left(\|\boldsymbol{m}_j^{(t)} - \boldsymbol{\mu}^{(t+1)}\|_2^2 + dV_j^{(t)} \right).$$

The final routing score is then computed as

$$\widehat{B}_j = \frac{\widehat{\varepsilon}_j^2 / n_j}{\widehat{\varepsilon}_j^2 / n_j + \widehat{\sigma}^2},$$

where $\widehat{\sigma}^2$ is the final EM iterate. For this alternative implementation, one may initialize σ_0^2 at $\widehat{\sigma}^2_{\mathrm{ANOVA}}$ when positive and initialize $\boldsymbol{\mu}^{(0)}$ by the associated precision-weighted mean. By standard EM convergence theory (Dempster et al., 1977; Wu, 1983; Laird, 1982), these updates weakly increase the empirical-Bayes observed-data likelihood.

Proposition E.2 (Variance Term in the Empirical-Bayes EM Update)**.** *The term* $\left(B_j^{(t)}\right)^2 (\sum_i \alpha_i^{(t)})^{-1}$ *must not be added to* $V_j^{(t)}$ *in the empirical-Bayes EM update above. With that addition, the iteration need not maximize the stated observed-data likelihood and can converge to a nonstationary point of it.*

Proof. Consider one coordinate, $M \geq 2$ clients with the same effective local variance $r > 0$, and identical observed means. After profiling out μ, the observed-data log-likelihood is, up to an additive constant, $-(M/2)\log(\sigma^2 + r)$. It is strictly decreasing on $\sigma^2 \geq 0$ and therefore has its unique maximum at $\sigma^2 = 0$.

For identical means, the posterior means also coincide. If the extra term is inserted into the variance used by the M-step, the variance iteration becomes

$$\sigma_{t+1}^2 = \frac{r\left(\sigma_t^2 + r/M\right)}{\sigma_t^2 + r}.$$

Its unique positive fixed point is $r/\sqrt{M}$. Moreover, writing $x_t = \sigma_t^2/r$ and $a = M^{-1/2}$ shows that the update $f(x) = (x + 1/M)/(x + 1)$ satisfies

$$f(x) - a = \frac{(1-a)(x-a)}{x+1}, \qquad f(x) - x = \frac{a^2 - x^2}{x+1}.$$

Hence every positive initialization converges monotonically to a, even though the likelihood is still decreasing there. The modified iteration is therefore not EM for the stated empirical-Bayes model. The additional term belongs to a different posterior-uncertainty calculation, not to this complete-data likelihood update. □

Optional joint update of ε_j^2. A full joint EM could also include the within-client noise variances in the parameter vector ϕ. Keeping the same latent variables $Z = (\boldsymbol{\theta}_j)_{j \in C}$, the M-step for these additional parameters is

$$\varepsilon_j^2 \leftarrow \frac{1}{dn_j} \left[\mathrm{SSE}_j + n_j \left(\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{m}_j^{(t)}\|_2^2 + dV_j^{(t)} \right) \right], \qquad j \in C.$$

This update applies when the initial within-client variance estimates are unstable, for example when n_j is very small.

The default FedBHG implementation does not use this joint update: it fixes $\widehat{\varepsilon}_j^2$ at the unbiased within-client estimates and maximizes the profile likelihood over the between-client variance σ^2.

This keeps the routing statistic simple and avoids coupling the between-client estimate to poorly identified client-specific noise parameters.

Remark E.3 (Clients with $n_j = 1$)**.** *When $n_j = 1$, the within-client variance ε_j^2 is not identifiable from client j alone. Two practical remedies are available.* (i) Pooled estimation*: assume a common noise variance ε^2 and estimate it by pooling all clients with $n_j \geq 2$ via*

$$\widehat{\varepsilon}^2 = \frac{\sum_{j:n_j \geq 2} \mathrm{SSE}_j}{d \sum_{j:n_j \geq 2} (n_j - 1)}.$$

(ii) Shrinkage*: shrink the formal sample variance toward a pooled target s_0^2 via*

$$\widetilde{\varepsilon}_j^2 = \frac{(n_j - 1)s_j^2 + \lambda s_0^2}{(n_j - 1) + \lambda},$$

where $\lambda > 0$ controls the shrinkage strength. In either case, the resulting $\widehat{\varepsilon}_j^2$ is substituted into the ANOVA, profile-likelihood, or EM formulae above. The reported synthetic experiments have $n_j \geq 8$, so this edge case does not arise there; a deployment containing one-example clients must choose one of the remedies explicitly.

E.3 Finite-Sample Properties of the Moment Estimators

The following lemmas record the exact first and second moments of the untruncated components used by the moment initialization. They do not establish unbiasedness of the positive-part estimator or of the subsequent EM iterate.

Lemma E.4 (Conditional Unbiasedness of the Client Mean)**.** *Under Assumption A.1,*

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j] = \boldsymbol{\theta}_j.$$

Proof. By definition,

$$\hat{\boldsymbol{\theta}}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} g_{jk},$$

and under the data-level model each sample gradient satisfies

$$\mathbb{E}[g_{jk} \mid \boldsymbol{\theta}_j] = \boldsymbol{\theta}_j.$$

Taking conditional expectation and averaging over k gives the claim. □

Lemma E.5 (Unbiasedness of SSE)**.** *Under Assumption A.1, for every fixed $\boldsymbol{\mu}$, $\mathbb{E}[\mathrm{SSE}_j \mid \boldsymbol{\mu}] = d(n_j - 1)\varepsilon_j^2$. Consequently, $\hat{\varepsilon}_j^2 = \mathrm{SSE}_j/(d(n_j - 1))$ is unbiased for ε_j^2.*

Proof. For conditionally independent vectors with covariance $\varepsilon_j^2 I_d$, the standard sample-variance identity gives

$$\mathbb{E}\left[\sum_{k=1}^{n_j} \|g_{jk} - \widehat{\boldsymbol{\theta}}_j\|_2^2 \,\middle|\, \boldsymbol{\theta}_j\right] = (n_j - 1)\,\mathrm{tr}(\varepsilon_j^2 I_d) = d(n_j - 1)\varepsilon_j^2.$$

Taking expectation over $\boldsymbol{\theta}_j$ conditional on the fixed $\boldsymbol{\mu}$ proves the claim; Gaussian per-example gradients are not required. □

Lemma E.6 (Expectation of the Untruncated MoM Component)**.** *Define the unweighted client-mean and trace-normalized sample variance across clients:*

$$\bar{\boldsymbol{\theta}} := \frac{1}{M} \sum_{j=1}^{M} \hat{\boldsymbol{\theta}}_j, \qquad s^2 := \frac{1}{d(M-1)} \sum_{j=1}^{M} \|\hat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}\|_2^2.$$

Under Assumption A.1, for every fixed $\boldsymbol{\mu}$,

$$\mathbb{E}[s^2 \mid \boldsymbol{\mu}] = \sigma^2 + \frac{1}{M} \sum_{j=1}^{M} \frac{\varepsilon_j^2}{n_j}.$$

Consequently, if each $\hat{\varepsilon}_j^2$ is unbiased, the untruncated *difference $s^2 - \frac{1}{M}\sum_j \hat{\varepsilon}_j^2/n_j$ is unbiased for σ^2. The positive-part estimator $\hat{\sigma}^2_{\mathrm{ANOVA}} := (s^2 - \frac{1}{M}\sum_j \hat{\varepsilon}_j^2/n_j)_+$ need not be unbiased.*

Proof. Fix coordinate $\ell \in [d]$. The coordinate-ℓ sample variance is $s_\ell^2 := (M-1)^{-1}\sum_j(\hat{\theta}_{j,\ell} - \bar{\theta}_\ell)^2$. Conditional on fixed $\boldsymbol{\mu}$, the $\hat{\boldsymbol{\theta}}_j$ are independent, $\mathrm{Var}(\hat{\theta}_{j,\ell} \mid \boldsymbol{\mu}) = \sigma^2 + \varepsilon_j^2/n_j$, and $\mathrm{Var}(\bar{\theta}_\ell \mid \boldsymbol{\mu}) = \frac{1}{M}\sigma^2 + \frac{1}{M^2}\sum_j \varepsilon_j^2/n_j$. By the standard identity for sample variance:

$$\begin{aligned}\mathbb{E}[s_\ell^2 \mid \boldsymbol{\mu}] &= \frac{1}{M-1}\left(\sum_j \mathrm{Var}(\hat{\theta}_{j,\ell} \mid \boldsymbol{\mu}) - M\,\mathrm{Var}(\bar{\theta}_\ell \mid \boldsymbol{\mu})\right) \\ &= \frac{1}{M-1}\left(M\sigma^2 + \textstyle\sum_j \varepsilon_j^2/n_j - \sigma^2 - \frac{1}{M}\sum_j \varepsilon_j^2/n_j\right) = \sigma^2 + \frac{1}{M}\sum_j \frac{\varepsilon_j^2}{n_j}.\end{aligned}$$

Since this holds for every coordinate ℓ, averaging gives the stated conditional expectation. □

Lemma E.7 (Variance of the Trace Estimator Under Gaussian Residuals)**.** *Under Assumption A.1, additionally suppose the per-example gradient residuals are conditionally Gaussian. Then, conditional on $\boldsymbol{\theta}_j$,*

$$\mathrm{Var}(\hat{\varepsilon}_j^2 \mid \boldsymbol{\theta}_j) = \frac{2\varepsilon_j^4}{d(n_j-1)}.$$

Proof. Conditional on $\boldsymbol{\theta}_j$, $\mathrm{SSE}_j/\varepsilon_j^2 \sim \chi^2_{d(n_j-1)}$. Since $\mathrm{Var}(\chi_k^2) = 2k$, we have $\mathrm{Var}(\mathrm{SSE}_j \mid \boldsymbol{\theta}_j) = 2d(n_j-1)\varepsilon_j^4$. Dividing by $(d(n_j-1))^2$ proves the stated variance formula. Under this independent-coordinate Gaussian model, the variance is $O((d(n_j-1))^{-1})$, so increasing d stabilizes the trace estimator. □

Lemma E.8 (Trace of a Sample Covariance Matrix)**.** *Let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be the sample covariance matrix of samples $\{\mathbf{X}_i\}_{i=1}^N$ with sample mean $\bar{\mathbf{X}}$. Then*

$$\mathrm{trace}(\boldsymbol{\Sigma}) = \frac{1}{N-1}\sum_{i=1}^N \|\mathbf{X}_i - \bar{\mathbf{X}}\|_2^2.$$

Proof. The ℓ-th diagonal entry of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}_{\ell\ell} = \frac{1}{N-1}\sum_{i=1}^N (X_{i,\ell} - \bar{X}_\ell)^2$. Summing over $\ell \in [d]$ and exchanging the order of summation gives

$$\mathrm{trace}(\boldsymbol{\Sigma}) = \sum_{\ell=1}^d \boldsymbol{\Sigma}_{\ell\ell} = \frac{1}{N-1}\sum_{i=1}^N\sum_{\ell=1}^d (X_{i,\ell} - \bar{X}_\ell)^2 = \frac{1}{N-1}\sum_{i=1}^N \|\mathbf{X}_i - \bar{\mathbf{X}}\|_2^2.$$

□

Lemma E.9 (Matrix-Form Unbiasedness of the Sample Covariance)**.** *Under the shared-shape model with $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\mu}) = (\sigma^2 + \varepsilon_j^2/n_j)K$, let $\mathbf{S}$ denote the full sample covariance matrix computed from $\{\hat{\boldsymbol{\theta}}_j\}_{j=1}^M$. Then*

$$\mathbb{E}[\mathbf{S} \mid \boldsymbol{\mu}] = \sigma^2 K + \frac{1}{M}\sum_{j=1}^M \frac{\varepsilon_j^2}{n_j}K = \left(\sigma^2 + \frac{1}{M}\sum_{j=1}^M \frac{\varepsilon_j^2}{n_j}\right)K.$$

Consequently, the scalar estimator s^2 is an unbiased estimator of $\sigma^2 + \frac{1}{M}\sum_j \varepsilon_j^2/n_j$.

Proof. Write $Y_j = \widehat{\boldsymbol{\theta}}_j$ and $\overline{Y} = \frac{1}{M}\sum_j Y_j$. Conditional on fixed $\boldsymbol{\mu}$, the vectors are independent with common mean $\boldsymbol{\mu}$ and possibly different covariances. The standard matrix identity gives

$$\mathbb{E}\left[\frac{1}{M-1}\sum_{j=1}^M (Y_j - \overline{Y})(Y_j - \overline{Y})^\top \,\middle|\, \boldsymbol{\mu}\right] = \frac{1}{M}\sum_{j=1}^M \mathrm{Cov}(Y_j \mid \boldsymbol{\mu}).$$

Substituting $\mathrm{Cov}(Y_j \mid \boldsymbol{\mu}) = (\sigma^2 + \varepsilon_j^2/n_j)K$ yields the displayed expression. Taking traces and using $\mathrm{trace}(K) = d$ gives the scalar conclusion. □

Lemmas E.4–E.8 justify the untruncated moment components used to initialize the procedure. They do not imply that the positive-part ANOVA estimator or the EM iterate is unbiased. We next relate the construction to classical empirical Bayes before examining model misspecification.

E.4 Relation to Classical Empirical Bayes

Estimating σ^2 by profile likelihood and then applying Bayesian posterior inference to $\boldsymbol{\theta}_j$ is an instance of *empirical Bayes* (also called Type-II maximum likelihood). Here we briefly situate this estimator relative to earlier formulations.

Murphy (2012, Section 5.6.2) discusses the corresponding one-dimensional homoscedastic marginal model. Let r denote the common observation variance of a client mean—in our notation, $r = \varepsilon^2/n$ when both the per-example variance and sample size are common. With $\hat{\mu}' = \frac{1}{M}\sum_j \hat{\theta}_j$, the constrained MLE is

$$s_M^2 := \frac{1}{M}\sum_{j=1}^{M}(\hat{\theta}_j - \hat{\mu}')^2, \qquad \hat{\sigma}^2_{\mathrm{MLE}} = \max\{0,\ s_M^2 - r\}.$$

The positive part enforces the parameter constraint $\sigma^2 \geq 0$. This estimator is closely related to, but does not coincide at finite M with, the untruncated ANOVA/MoM expression, which uses denominator $M-1$. Equal per-example variances do not imply equal client-mean variances when the n_j differ. The profile likelihood in Section E.2 directly handles the heterogeneous effective variances $r_j = \varepsilon_j^2/n_j$ and is the default FedBHG estimator.

This construction is also structurally the Gaussian random-effects estimator used in meta-analysis: coalition observations receive inverse total-variance weights $(\sigma^2 + \varepsilon_j^2/n_j)^{-1}$ (DerSimonian & Laird, 1986). Hierarchical and empirical-Bayes personalization has already been developed in Self-FL, statistical frameworks for personalized estimation, pFedVEM, Makhija et al., and FedHB (Chen et al., 2022; Ozkara et al., 2023; Zhu et al., 2023; Makhija et al., 2024; Kim & Hospedales, 2025); concurrent VANEB studies heteroskedastic local estimators through variance-aware nonparametric empirical Bayes (Chang et al., 2026). FedBHG does not claim these estimation ingredients. It places the hierarchy on common-checkpoint sample-mean gradients and uses the resulting noise-to-total score in the action-level identity and pair-conditioned coalition analysis.

Murphy (2012) also discusses the fully Bayesian approach of placing priors on the variance parameters $(\sigma^2, \varepsilon_j^2)$ using, e.g., inverse-Gamma priors. Standard conjugate variants yield tractable full conditional distributions; joint or marginal inference generally requires numerical integration or Markov chain Monte Carlo (MCMC), at substantially higher computational cost. We leave the fully Bayesian treatment of the multivariate model (with anisotropic covariances $\varepsilon_j^2 K$) as an open problem.

Returning from the literature context to the practical deployment of FedBHG, we now examine the robustness of the three structural assumptions underlying the posterior and no-deviation calculations: sample-mean Gaussianity, client-level Gaussianity, and covariance structure.

E.5 Model Misspecification and Practical Scope

We discuss the practical status of each of the three assumptions underlying Theorem B.10.

Data-level Gaussianity. Assumption A.1 requires $\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, (\varepsilon_j^2/n_j)\, I_d)$. This concerns the distribution of the *sample mean*, not the individual gradient samples. A central-limit approximation may be reasonable when within-client gradients are weakly dependent and have controlled moments, but neither condition follows merely from evaluating gradients near convergence. Moderate sample size alone does not establish Gaussianity, which remains a modeling approximation rather than an asymptotic guarantee. Study S1 retains this sampling model to isolate the effect of estimating variance components and routes. Whether the approximation remains informative for trained neural models is a separate empirical question.

Client-level Gaussianity. The client-level Gaussian prior directly governs how client mean updates are shrunk toward the population mean and underlies the posterior-optimality statement for $\boldsymbol{\theta}_j^*(C)$. When the true distribution of $(\boldsymbol{\theta}_j)$ is multimodal or skewed, the scalar posterior mean and the

half-threshold lose their oracle interpretation. The coalition synthetics do not test this separate misspecification; the theory does not guarantee that the score remains predictive for an unresolved mixture. Section G therefore treats the saved real-gradient summaries as model criticism rather than evidence that the prior is correct.

Covariance structure. Real gradient covariance matrices are anisotropic: eigenvalue spectra can span several orders of magnitude. Our scalar estimator computes the trace-average across all d coordinates, collapsing the direction-specific information into a single number. Empirical studies of gradients and Hessians in overparameterized networks report anisotropic spectra and structured low-dimensional subspaces (Sagun et al., 2018; Gur-Ari et al., 2018). These observations motivate concern about trace averaging rather than validating it. A scalar approximation can preserve the decision when the route-relevant directions agree, but it can misclassify a client when different directions favor different routes. For the generalized covariance-weighted construction, Theorem B.27 guarantees no deviation when every direction favors the same route, while Corollary B.31 shows that conflicting directions can defeat that guarantee. The implemented scalar score does not test this matrix condition. The prospective real-gradient audit indeed strongly contradicts the fitted spherical model on directional and radial diagnostics. Its separate route-resampling analysis asks the narrower operational question of whether this misspecified plug-in decision is stable.

Soft-threshold robustness and misclassification cost. Equation (9) shows that the *excess conditional MSE* incurred by using a coalition estimator instead of the local estimator for player j scales with $(1 - 2B_j)$ and with the squared gap between the two estimators. For a fixed gap between the two actions, the magnitude of this excess risk decreases as B_j approaches $1/2$. Near-threshold routing is inexpensive only when that action gap is also controlled. More concretely, Proposition C.8 shows that a client with $|B_j - 1/2| \geq \delta$ can be rerouted only if its score error reaches at least δ. The paper does not derive a finite-sample concentration rate for the profile estimator. It therefore reports margins and realized client outcomes rather than presenting oracle-reference cells as evidence for plug-in correctness. Study S1 uses the known simulation posterior to measure how score and action errors affect the resulting partition across a complete margin–information grid. On real checkpoints, Section G complements that controlled study with leave-one-client-out and client-bootstrap refits; the most unstable observed routes are, as the proposition predicts, concentrated near the half-threshold.

Threshold scope. The threshold $B_j = 1/2$ emerges directly from the posterior-risk difference in Equation (9); it is not fitted to downstream outcomes. Outside the model, however, a score can remain discriminative without making the half-threshold policy optimal for a particular optimizer and horizon. The coalition synthetics evaluate posterior-risk routes rather than trained predictive models. The threshold is a theorem-derived decision under its assumptions, not a universal neural boundary; actual local and joint continuation must be evaluated on real tasks.

Summary. The three conditions underlying the no-deviation guarantee have the following practical status:

(1) *Data-level Gaussianity*: A sample-mean approximation whose adequacy depends on within-client dependence and tail behavior; sample size alone does not establish it.

(2) *Group-level Gaussianity*: A modeling assumption for the between-client distribution. Under misspecification, whether the routing score remains predictive is an empirical question.

(3) *Covariance structure*: Exact stability extends beyond isotropy when every direction favors the same route. The scalar score does not certify this condition, so near-boundary clients should be treated with caution under strong anisotropy.

The scalar and covariance theorems establish participation under known components and prescribed common actions. The synthetic experiments next test how the no-deviation property changes under fitted routes and directional conflict; the neural experiments then evaluate the trained continuations.

F SYNTHETIC EVALUATION: WHEN DOES NO-DEVIATION SURVIVE APPROXIMATION?

The oracle construction uses known variance components and prescribed common actions. FedBHG estimates the routes and weights, then continues with SGD or Adam. We evaluate these two transfers separately. Synthetic experiments ask whether the no-deviation property survives route estimation, covariance mismatch and changes to the coalition action. Section G asks whether the fitted rule improves prediction after ordinary training.

Claim boundary. The synthetic experiments measure posterior-risk deviations under controlled models in which every candidate coalition can be enumerated. The neural experiments measure predictive outcomes; they establish neither strategic stability of trained models nor validity of the Gaussian hierarchy. Their repeated-seed intervals describe optimizer variation on selected fixed client populations, while fresh partitions form a separate robustness check.

Both studies use six clients, allowing exhaustive evaluation of all $2^6 - 1 = 63$ nonempty candidate coalitions. For a proposed coalition S and incumbent partition π, let

$$G_j(S;\pi) := R_j\big(\pi(j) \mid \widehat{\boldsymbol{\theta}}_{S\cup\pi(j)}\big) - R_j\big(S \mid \widehat{\boldsymbol{\theta}}_{S\cup\pi(j)}\big)$$

be client j's reduction in true conditional posterior risk. The risk is normalized per coordinate, which changes gain magnitudes but not any preference. Coalition S blocks π when $G_j(S;\pi) \geq 0$ for every $j \in S$ and the inequality is strict for at least one member. Besides the no-deviation indicator, we report the number of blocks, the largest individual gain in a block, and

$$\Gamma(\pi) := \max_{S \text{ blocks } \pi} \min_{j\in S} G_j(S;\pi),$$

with $\Gamma(\pi) = 0$ if no block exists. Thus Γ records the largest common member gain and defines a continuous measure of deviation severity. Numerical risk differences within 10^{-12} are treated as ties.

F.1 ROUTE ESTIMATION ERROR

If the oracle route admits no blocking coalition, how often does its fitted replacement preserve that property? The first study varies the margin from the half-threshold and the independent information available to estimate the route.

Experiment. We set the between-client variance to one and draw, conditional on a fixed population mean,

$$\boldsymbol{\theta}_j \sim \mathcal{N}(\mathbf{0}, I_d), \qquad \widehat{\boldsymbol{\theta}}_j \mid \boldsymbol{\theta}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, r_j I_d), \qquad r_j = \varepsilon_j^2/n_j.$$

Every client has $n_j = 20$. Three clients have oracle score $B_j = 1/2 - \delta$ and three have $B_j = 1/2 + \delta$, where $\delta \in \{.02, .05, .10, .20, .35\}$. Because $\sigma^2 = 1$, we set $r_j = B_j/(1 - B_j)$. We cross this margin sweep with effective dimension $d \in \{1, 5, 20\}$. Increasing d adds independent observations for estimating both the within-client and between-client components.

The implementation observes finite-sample within-client variance estimates rather than r_j. Their sampling distribution is

$$\widehat{\varepsilon}_j^2 = \varepsilon_j^2 \frac{\chi^2_{d(n_j-1)}}{d(n_j-1)}.$$

It then applies the profile estimator of Section E, forms $\widehat{B}_j$, routes at one half, and uses the corresponding fitted precision for every coalition action. We compare four variants: oracle routes and weights, fitted routes with oracle weights, oracle routes with fitted weights, and the complete plug-in procedure. Every variant is evaluated under the true posterior risk, so route and action error are separated without using downstream outcomes to choose a policy.

We ran 250 fresh confirmation replicates in each of the 15 cells. The complete plug-in transition and its route-error diagnostic appear in Figure 1.

Table 3: No-deviation rate of the full plug-in partition. Every entry is based on 250 fresh confirmation trials and all 63 candidate coalitions.

Effective d	$\delta = .02$	.05	.10	.20	.35
1	3.2%	3.2%	7.2%	26.4%	44.8%
5	1.6%	8.8%	32.8%	60.4%	90.0%
20	2.8%	18.8%	37.6%	77.2%	100.0%

Margin and information improve no-deviation recovery. Table 3 shows the complete grid. No-deviation is rare near the half threshold and becomes common only when the route margin and effective information are both large. The broad increase across the margin–information grid includes fluctuations between adjacent low-information cells. The difficult endpoint $(d, \delta) = (1, .02)$ has a 3.2% no-deviation rate and 2.872 route errors per trial on average. At $(20, .35)$, all routes are recovered and all 250 trials admit no block.

Among the two fitted components tested here, route error produces the larger drop in no-deviation rate. Across the 3,750 full plug-in trials, all routes are correct in 1,329 trials; 93.00% of those partitions admit no block. After one route error the rate falls to 14.5%, and after two or more it is below 1%; pooled over the 2,421 trials with any error, it is 2.11%. Equally averaging the 15 cells, fitted weights with oracle routes retain a 64.64% no-deviation rate, fitted routes with oracle weights retain 36.03%, and the complete procedure retains 34.32%. In this comparison, route estimation has the larger observed effect, while estimated coalition actions produce a smaller but nonzero gap.

The continuous gain endpoint distinguishes recovery probability from deviation size. At $d = 1$, an unstable trial's mean Γ grows from .145 at margin .02 to .475 at margin .35: a wrong route becomes rarer but can be more consequential farther from the threshold. Larger margins improve route recovery while the remaining mistakes can generate sizable deviations, consistent with the route-margin conditions of Proposition C.8.

F.2 Directional Conflict

Does covariance anisotropy itself break the scalar partition, or does failure begin when different directions recommend different routes?

Experiment. Theorem B.27 requires every direction to recommend the same route. We test this boundary under a known two-dimensional Gaussian posterior. Six clients have base scalar scores $b_j \in \{.35, .65\}$, with three clients of each type. For a conflict ratio q and an independent orientation $s_j \in \{-1, 1\}$, the two directional scores are

$$B_{j1} = b_j + s_j(.15q), \qquad B_{j2} = b_j - s_j(.15q),$$

where $q \in \{0, .5, .9, 1.1, 1.5, 2.0\}$. We set the between-client covariance to I_2 and choose each diagonal mean-noise covariance to realize these scores. Every coalition uses the exact coordinate-wise precision action, while the partition retains the one scalar route given by b_j.

At $q = 0$, both directions have the scalar score. At $q = .5$ and .9, the covariance shapes differ but both directions remain on the scalar side of one half. Once $q > 1$, every client has one locally confident and one pool-reliant direction. We ran 500 fresh replicates per ratio and again enumerated all 63 candidate coalitions.

Directional conflict in this simulation family. Figure 3 shows that the no-deviation rate falls sharply only after the directional scores cross to opposite sides of the half-threshold. All 1,500 direction-consistent trials admit no block, including the non-shared-shape conditions $q = .5$ and .9. The no-deviation rate falls to 21.0% immediately after the threshold crossing and then to 2.4% and 1.0%. The number and severity of blocks increase with the conflict ratio: mean Γ rises from .019 at $q = 1.1$ to .368 at $q = 2$. Thus covariance anisotropy by itself is not the empirical obstruction in this construction. The scalar partition fails when different directions prescribe incompatible actions, matching the distinction between Theorem B.27 and the counterexample of Proposition B.30.

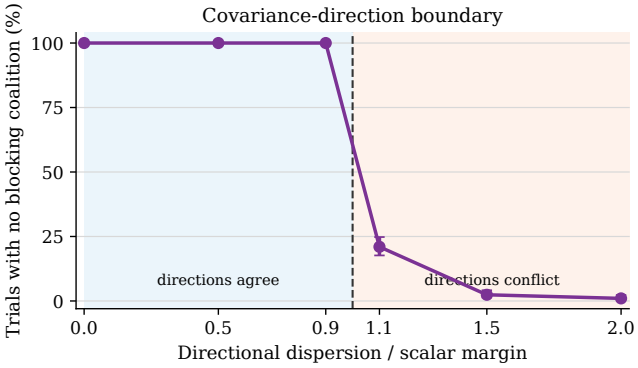

Figure 3: Covariance-direction boundary. The curve is the fraction of 500 trials per conflict ratio with no blocking coalition; bars are Wilson 95% intervals. The dashed line marks the first possible directional threshold crossing. Non-shared covariance shapes have no blocking coalition in these trials while every direction recommends the same route; the no-deviation rate in this simulation family drops sharply once the directions conflict.

Table 4: Covariance-direction boundary. "Common gain" is mean Γ; "strict gain" is the mean largest individual gain among blocking coalitions.

q	Directions agree	No deviation	Blocks	Common / strict gain
0.0	100%	100.0%	0.000	.000 / .000
0.5	100%	100.0%	0.000	.000 / .000
0.9	100%	100.0%	0.000	.000 / .000
1.1	0%	21.0%	2.182	.019 / .346
1.5	0%	2.4%	5.914	.137 / .765
2.0	0%	1.0%	9.578	.368 / 1.661

F.3 Aggregation-Weight Error

The scalar no-deviation theorem also depends on the coalition action. We therefore retain one supporting ablation with oracle routes and scores but replace posterior precision by sample count. Counts are drawn independently of precision from a log-normal distribution with median 80 and log standard deviation .8, then clipped to $[5, 1000]$; the implied per-example variance is adjusted so that the specified mean-gradient variance remains coherent. The realized rank correlation between count and precision averages .009.

Variance-adaptive aggregation admits no block in any of the 3,750 matched reference trials. Sample-count aggregation has a 2.8% no-deviation rate at $(d, \delta) = (1, .02)$, 74.4% at $(20, .35)$, and 24.61% when the 15 cells are equally averaged. Correct membership alone is insufficient: replacing precision aggregation by sample counts can reintroduce blocking coalitions even with oracle routes.

Route recovery, directional agreement and aggregation. Together, the experiments identify how estimated membership and coalition actions affect no-deviation. Larger margins and more independent information improve route recovery; direction agreement permits anisotropic sharing, and precision aggregation yields no blocking coalition in all 3,750 matched oracle-route reference trials. Route errors, directional conflict and sample-count aggregation can each reintroduce blocking coalitions. The next section evaluates predictive outcomes of ordinarily trained continuations. Reproduction details appear in Section G.7.

G Neural Evaluation: When Does Selective Continuation Help?

Given one checkpoint, should all clients continue locally, all continue jointly, or should checkpoint statistics choose for each client? *Local-all* trains every client separately. *Joint-all* trains one model

Table 5: Five-seed repeated evaluation with fixed client partitions. The four cells were selected after seed 1 and frozen before seeds 2–5. Ckpt, Local, Joint, and FedBHG are absolute outcomes after ten continuation rounds (mean ± standard deviation). The paired Δ best column reports the mean improvement over the better universal action and its descriptive 95% t interval.

Cell	Metric	Ckpt	Local	Joint	FedBHG	Δ best [95% CI]
F-S3/CFL	Acc. (%)	84.98±.32	85.30±.37	85.21±.31	88.25±.21	+2.944 [2.664, 3.224]
	Loss	.770±.006	.868±.011	.712±.003	.703±.006	+.0089 [.0027, .0151]
E-S3/IFCA	Acc. (%)	69.25±.72	65.72±.52	70.11±.48	70.89±.77	+0.782 [.273, 1.292]
	Loss	1.011±.073	2.414±.119	.973±.080	.966±.038	+.0071 [−.0482, .0624]
Sh/CFL	Acc. (%)	41.95±.75	41.89±.74	42.00±.74	42.05±.72	+.051 [−.010, .113]
	Loss	2.091±.031	2.094±.030	2.091±.032	2.086±.031	+.0048 [.0015, .0081]
Sh/IFCA	Acc. (%)	43.16±.92	43.22±.94	43.20±.92	43.42±.93	+.175 [.150, .200]
	Loss	1.968±.042	1.973±.043	1.968±.042	1.961±.044	+.0068 [.0048, .0088]

per full upstream cluster using fitted precision weights. *FedBHG* estimates $\widehat{\sigma}^2$ and $\widehat{B}_j$ within each cluster, sends clients with $\widehat{B}_j < 1/2$ to local training, and trains the remaining clients jointly with fixed fitted precision weights. Each distinct joint membership trains its own model from the same checkpoint. Routes and weights are fixed before any continuation outcome is observed.

G.1 Does Selective Continuation Beat the Universal Choices?

We use 100 clients on Fashion-MNIST, EMNIST-Letters and natural-client Shakespeare, with CFL or IFCA checkpoints trained for 200 rounds and ten further rounds of continuation. All clients participate; policies within a cell share the checkpoint, optimizer and per-client training budget. The co-primary outcomes are sample-weighted raw accuracy and client-uniform held-out loss. Architecture, heterogeneity and optimizer details appear in Section G.7.

Development breadth. The ten-cell seed-1 sweep improves checkpoint accuracy in 10/10 cells, exceeds the more accurate universal continuation in 8/10 and the lower-loss continuation in 7/10, and improves both in 5/10. Median improvements over the better universal action are 1.05 accuracy points and .0042 loss reduction. Both Fashion S4 cells gain accuracy while increasing loss. Because four repeated cells were selected after this sweep, it describes breadth rather than uncertainty. The complete grid remains in Table 10.

G.1.1 Do the Gains Survive Optimizer Randomness?

Five training seeds per selected cell hold the client population and data partition fixed while varying optimization, including upstream training. The paired intervals therefore describe optimizer variation, not generalization to new client populations. Table 5 reports every absolute universal reference and FedBHG outcome. Absolute columns are mean ± standard deviation across training seeds. The last column is the paired mean FedBHG improvement over the better universal action, with its 95% t interval. For accuracy, improvement is a percentage-point gain; for loss, it is a reduction.

Across the 20 runs in the four repeated cells, FedBHG improves checkpoint raw accuracy in 20/20, beats the more accurate universal action in 19/20, lowers loss relative to the lower-loss universal action in 17/20, and improves both in 16/20. In two of the four cells, the accuracy and loss intervals both exclude zero. Fashion S3/CFL improves both outcomes in 5/5 seeds; its mean accuracy gain is

2.944 points. Shakespeare/IFCA also improves both outcomes in 5/5 seeds, with a smaller but tightly repeated .175-point accuracy gain and .0068 loss reduction.

The other two cells delimit the claim. EMNIST S3/IFCA improves raw accuracy in all five seeds, by .782 points on average, but its loss difference changes sign and its interval includes zero. Shakespeare/CFL lowers loss in every seed, while its .051-point raw-accuracy advantage is unresolved. The gains repeat most clearly on both outcomes for Fashion S3/CFL and Shakespeare/IFCA; the other two cells retain uncertainty in one outcome.

G.1.2 Do the Gains Transfer to New Client Populations?

To test dependence on the original client population, we generate two new partitions per selected vision task. Table 6 reports all four observations. Data seed, rather than optimizer seed, changes the client sample counts, class subsets, and corruption draws. Training seed 1 and every other protocol choice remain fixed. Each row is therefore a new end-to-end upstream checkpoint, route, and set of continuation branches, not a replay from the original partition.

FedBHG improves checkpoint accuracy and beats the more accurate universal action in 4/4 cells, with gains from 1.03 to 1.81 percentage points. It lowers loss relative to the lower-loss universal action in 3/4 and improves both outcomes in those same three cells. The remaining Fashion partition gains 1.45 accuracy points while its loss is .013 higher than joint-all, so accuracy is more consistent than loss across these four fresh-partition observations. These observations are reported separately from the optimizer-seed intervals.

Table 6: Prospectively frozen check on two new client partitions per selected vision task. Accuracy is sample weighted and loss is client uniform after ten continuation rounds. Δ compares FedBHG with the better universal local or joint action under the corresponding metric.

Cell / data seed	L/J	Ckpt acc.	Local acc.	Joint acc.	FedBHG acc.	Δ acc.	FedBHG loss	Δ loss
F-S3/CFL / 12346	61/39	84.72	84.19	84.57	86.02	+1.45	.738	$-.013$
F-S3/CFL / 12347	45/55	83.72	82.92	84.32	86.13	+1.81	.727	+.011
E-S3/IFCA / 12346	3/97	75.75	70.51	76.15	77.18	+1.03	1.048	+.010
E-S3/IFCA / 12347	3/97	68.53	66.58	68.93	70.53	+1.60	1.454	+.078

G.2 Where Does the Full Score Add Value?

Coverage-matched comparisons identify where the full score adds to selective continuation. We compare FedBHG with routes having the same number of local clients within each base cluster, selected by sample size, per-example gradient variance, or three fixed random masks. The controls retain the fitted client precisions and renormalize them within each new pool. Every distinct joint group is trained separately under the same continuation protocol.

FedBHG beats the mean of three frozen random routes on both co-primary outcomes in all 20 repeated runs and beats the within-client-variance route on both in 14/20. Comparisons with deterministic controls vary by setting: its membership matches the sample-size route in 7/20 runs and the variance-only route in 5/20; it beats sample-size routing on both outcomes in 6/20.

Fashion S3/CFL is the only repeated cell whose FedBHG route differs from both simple routes in every seed. Its mean pairwise local-set Jaccard similarity across seeds is .983; it gains .925 raw-accuracy points over sample-size routing and 1.118 points over variance-only routing. Its loss reduction over variance-only is .0415; its smaller .0020 mean reduction over sample-size is unresolved (95% interval $[-.000016, .003949]$). EMNIST S3/IFCA matches the variance-only route in all five seeds, while both Shakespeare cells are effectively tied with sample-size routing. Fashion S3/CFL is therefore the clearest repeated case of incremental accuracy gains over the simple routes. Where memberships coincide, the experiments identify the value of selective continuation without isolating an additional benefit from the full score.

The fresh-partition check preserves this task-specific membership pattern: FedBHG differs from both controls on both Fashion partitions and matches both on both EMNIST partitions. The fresh partitions retain the membership pattern of the repeated cells; the repeated performance contrasts above assess the score-specific predictive gains.

What the score adds to a within-cluster ranking. For a base cluster with $\widehat{\sigma}^2 > 0$, $\widehat{B}_j = \widehat{v}_j/(\widehat{\sigma}^2 + \widehat{v}_j)$ is strictly increasing in $\widehat{v}_j = \widehat{\varepsilon}_j^2/n_j$. Consequently, at a fixed number of local clients, ranking by effective sampling variance reproduces the score ranking, up to the same tie convention. The recorded sample-size and variance-only controls isolate n_j and $\widehat{\varepsilon}_j^2$, respectively. Effective sampling variance $\widehat{\varepsilon}_j^2/n_j$ already determines the within-cluster ranking. The Bayesian model sets a variance-calibrated cutoff and aggregation weights. Its empirical distinction must be assessed through those choices and the resulting trained branches.

G.3 Does the Score Make Useful Client-Level Decisions?

Client-level diagnostics relate the score to the benefit of sharing and describe how gains vary across clients. We ask whether the score ranks that benefit, whether its threshold chooses the lower-loss universal branch, and whether clients are harmed relative to the checkpoint.

Ranking the benefit of sharing. Define the observed benefit of joint-all over local-all after h rounds as

$$D_j^{(h)} := L_j\big(\boldsymbol{w}_{j,\mathrm{local}}^{(h)}\big) - L_j\big(\boldsymbol{w}_{C,\mathrm{joint}}^{(h)}\big).$$

Here C is the client's full base cluster: the two trajectories are local-all and joint-all, before selecting the smaller FedBHG pool. Positive $D_j^{(h)}$ means that joint-all has lower held-out loss. We report the Spearman association between $\widehat{B}_j$ and $D_j^{(h)}$, together with agreement between the fixed half-threshold and the lower-loss universal branch, assigning loss ties to joint-all. This diagnostic does not measure preference for the post-routing FedBHG pool.

At ten rounds, the mean of the 20 run-level Spearman correlations between $\widehat{B}_j$ and $D_j^{(10)}$ is .552. Cell means are .567 for Fashion S3/CFL, .758 for EMNIST S3/IFCA, .301 for Shakespeare/CFL and .581 for Shakespeare/IFCA. This is ranking information, not probability calibration. On the four fresh cells the correlations range from .438 to .787 (Table 7).

Choosing the lower-loss universal branch. Mean half-threshold agreement rises from 57.5% after one round to 68.9% after five and 69.3% after ten. The ten-round cell means are 70.4% for Fashion S3/CFL, 74.4% for EMNIST S3/IFCA, 57.6% for Shakespeare/CFL and 74.8% for Shakespeare/IFCA. The score thus retains agreement with the observed universal branch comparison after multi-round training, with variation across tasks and clusterers.

Harm relative to the checkpoint. Harm compares a selected model with the checkpoint, while branch agreement compares two continuations. We measure checkpoint-relative harm in both accuracy and loss. The client-level ECDFs in Figure 2 show the distribution of client-level accuracy changes. The figure uses one repeated cell per dataset and first averages each client's outcome over the five optimizer seeds, so each client appears once. Relative to local-all, FedBHG lowers strict harm and removes the most extreme left tail in all three panels. Against joint-all, the curves cross: FedBHG has the larger equal-client mean gain in every panel, whereas joint-all has slightly lower strict harm in the displayed EMNIST and Shakespeare cells. These seed-averaged client fractions differ from the mean run-level harm rates reported below; averaging and thresholding need not commute. The tables retain all four repeated cells and all 20 runs.

Across the 20 repeated-cell runs, FedBHG's mean accuracy-harm rate relative to the common checkpoint is 32.2%, compared with 52.9% for local-all. FedBHG has the lower accuracy-harm rate in every run. FedBHG lowers the loss-harm rate in 15/20 runs, and Fashion S3/CFL has higher mean loss-harm than local-all despite improving aggregate accuracy and loss. Aggregate fit and the fraction of clients harmed are therefore distinct empirical criteria.

Aggregate gains remain compatible with heterogeneous client benefits and a substantial minority harmed relative to the checkpoint.

G.4 What Survives Model Misspecification?

The fitted score retains positive association with the benefit of sharing even when checkpoint gradients depart from its spherical working model. We compare gradient-shape statistics with fitted-model draws, then measure the proportion of refits that change a client's local/joint assignment. Section G.7 specifies the checks and resampling. Table 7 reports the four cellwise refitting summaries and the large-cluster effective-rank diagnostic. In every large cluster, all five predeclared statistics fall in the most adverse tail attainable with 200 parametric bootstrap draws (Monte Carlo $p = 1/201$), rejecting the fitted model on all five checks. Relative to the fitted null, residual spectra are more concentrated, directional cosines are larger and radial dispersion is more heterogeneous. Effective rank is 12.2–71.6 versus fitted-null medians of 95.9–99.0. These checkpoints depart substantially from the fitted spherical Gaussian hierarchy. EMNIST's additional three-client clusters are retained in the route-sensitivity aggregation but are too small for headline shape criticism.

Despite the covariance misspecification, average route-flip probability is 1.39–3.47% under client resampling and .01–.16% under deletion. The score also retains positive association with trained local-minus-joint loss in all four cells ($\rho = .438$–$.787$). Sensitivity concentrates near the half-threshold: the most fragile client in each cell changes route in 42.8–51.4% of bootstrap refits. The positive predictive association in these four cells is an empirical property of the fitted score, not evidence that the rejected spherical hierarchy describes their gradient shape. Resampling measures repeatability of the fitted rule, not recovery of the oracle route.

Table 7: Real-gradient model criticism and route sensitivity on the four prospective cells. Boot. mean/max is the clientwise route-flip probability under 500 client bootstrap refits. LOO is the mean flip probability across all one-client deletions. Agree and ρ compare $\widehat{B}_j < 1/2$ with the realized lower-loss action and rank $\widehat{B}_j$ against local-minus-joint loss. Effective rank is shown for the large cluster; parentheses report its fitted-null median.

Cell / data seed	Boot. mean	Boot. max	LOO mean	Agree	ρ	Eff. rank (null)
F-S3/CFL / 12346	3.29%	51.4%	.16%	62%	.438	12.2 (99.0)
F-S3/CFL / 12347	3.47%	46.4%	.11%	66%	.561	16.4 (99.0)
E-S3/IFCA / 12346	1.48%	43.8%	.03%	72%	.787	66.7 (95.9)
E-S3/IFCA / 12347	1.39%	42.8%	.01%	62%	.774	71.6 (95.9)

Table 8: End-to-end comparison after 210 communication rounds. Values are mean $\pm$ standard deviation over the same five training seeds and one fixed client partition. Accuracy is sample weighted and loss is client uniform. FedBHG rows use the repeated task/clusterer cells; the baselines train from initialization and therefore do not share FedBHG's round-200 checkpoint. Self-FL and pFedGraph were run on the two vision tasks.

Dataset	Method	Accuracy (%)	Loss
Fashion-MNIST S3	FedBHG/CFL	$88.25 \pm .21$	$.703 \pm .006$
	Ditto	$84.73 \pm .45$	$.738 \pm .010$
	pFedMe	71.20 ± 1.56	$.917 \pm .023$
	Self-FL	$73.51 \pm .28$	$1.133 \pm .039$
	pFedGraph	$78.09 \pm .42$	$1.317 \pm .023$
EMNIST-Letters S3	FedBHG/IFCA	$70.89 \pm .77$	$.966 \pm .038$
	Ditto	$67.67 \pm .54$	$1.091 \pm .026$
	pFedMe	66.00 ± 1.12	$1.182 \pm .028$
	Self-FL	67.24 ± 1.23	$1.616 \pm .063$
	pFedGraph	58.85 ± 1.26	$2.570 \pm .232$
Shakespeare	FedBHG/CFL	$42.05 \pm .72$	$2.086 \pm .031$
	FedBHG/IFCA	$43.42 \pm .93$	$1.961 \pm .044$
	Ditto	38.23 ± 1.78	$2.179 \pm .073$
	pFedMe	$18.17 \pm .00$	$3.211 \pm .002$

G.5 Competitiveness Against End-to-End Personalization Baselines

These end-to-end comparisons assess competitiveness at a matched communication horizon; the common-checkpoint experiments assess the continuation decision. We compare Ditto (Li et al., 2021) and pFedMe (Dinh et al., 2020) using the implementations and tuning protocol specified in Section G.7 for 210 communication rounds on the same architectures, clients, S3 vision partitions, natural Shakespeare partition, data seed, and training seeds. This matches FedBHG's total horizon of 200 base rounds plus ten continuation rounds. The baselines train from initialization with their own optimization trajectories and objectives.

Table 8 compares the repeated FedBHG cells with four end-to-end baselines. The Ditto implementation maintains independent global FedAvg and client-personalized trajectories. The pFedMe implementation maintains separate personalized inner iterates and client centers and applies the server mixing update. Self-FL controls local work and aggregation from estimated inter- and intra-client uncertainty, while pFedGraph learns client-specific collaboration weights from model similarity and data size.

Ditto has the highest accuracy and lowest client-uniform loss among the non-FedBHG comparators in every dataset. FedBHG nevertheless exceeds Ditto on both outcomes in every aligned seed. The paired mean FedBHG-minus-Ditto accuracy gaps are 3.517 points on Fashion-MNIST, 3.222 on EMNIST-Letters, 3.822 on Shakespeare/CFL, and 5.185 on Shakespeare/IFCA; their descriptive 95% intervals are $[3.005, 4.030]$, $[2.371, 4.072]$, $[2.252, 5.391]$, and $[3.671, 6.699]$. The corresponding

loss reductions are .0350, .1257, .0928, and .2180, and all four intervals exclude zero. FedBHG also exceeds pFedMe on both outcomes in every aligned seed. pFedMe's identical Shakespeare accuracies across the five seeds reflect convergence to the same low-accuracy solution on the fixed writer partition; the losses retain small seed variation. The checkpoint-identity check is documented in Section G.7.

Against the two closer vision baselines, FedBHG likewise has higher accuracy and lower client-uniform loss in all 20 aligned seedwise comparisons. On Fashion-MNIST, its paired accuracy advantage is 14.739 points over Self-FL (95% interval $[14.285, 15.192]$) and 10.157 over pFedGraph ($[9.831, 10.482]$); the corresponding loss reductions are .4301 and .6141. On EMNIST-Letters, the accuracy advantages are 3.645 points over Self-FL ($[2.212, 5.078]$) and 12.039 over pFedGraph ($[9.889, 14.189]$), with loss reductions of .6509 and 1.6042. All four loss intervals also exclude zero. pFedGraph illustrates why both endpoints are needed: on Fashion-MNIST it attains higher sample-weighted accuracy than Self-FL, but its client-uniform accuracy is 64.20% versus 72.04%, and its bottom-decile accuracy is 11.63% versus 40.46%.

Optimizer work and tuning are documented in Section G.7; Table 5 contains the same-checkpoint comparison.

G.6 What Does the Routing Decision Cost?

Routing takes a small fraction of upstream training time at the tested scale, but its communication savings depend on how many clients continue locally. Table 9 reports measurements for the four repeated cells, averaged over seeds 2–5. Each client uploads one float32 d-vector and two float64 scalars, or $4d + 16$ bytes, during routing. The "FedBHG GB" column then adds the ten-round upload and download traffic of only the clients assigned to joint continuation; locally routed clients communicate no model after routing. "Saving" compares this total with ten rounds of joint-all continuation, which requires no routing summary.

The client-summary pass and server fit together take 8.16–33.66 seconds across the four cell means, or 0.6–2.7% of the corresponding recorded 200-round base training time. The fit itself takes less than five seconds on average in every cell. The implementation stores the $M \times d$ summary array and cluster Gram matrices; their analytic lower-bound workspace is 0.33–0.88 GB in these runs. Its storage scales as $O(Md + M^2)$, and forming the Gram matrices plus G profile evaluations costs $O(M^2 d + GM^2)$. The reported 109–286 MiB CUDA peaks cover the selected continuation sources.

Communication savings vary with the fraction routed locally. Fashion S3/CFL and Shakespeare/IFCA route roughly half the clients locally and reduce traffic relative to joint-all by 41.5% and 44.5%. Shakespeare/CFL reduces it by 26.5%. EMNIST S3/IFCA routes about 92% of clients jointly, so the one-time summary upload nearly offsets the traffic avoided by its locally routed clients and the reduction is 3.2%. The savings therefore depend on local coverage and the one-time summary upload.

The four prospective partition cells fall within the same order of routing time, communication, and continuation-memory use observed in the repeated cells. Summary collection takes 3.80–7.41 seconds and server fitting .59–4.00 seconds. Operational incremental communication ranges from 7.64 to 17.99 GB, and selected-continuation peak PyTorch allocation is 109–123 MiB. These observations are reported separately from the repeated-cell means in Table 9. Operational continuation trains only the selected local and joint branches. The additional counterfactual branches needed to compare policies belong to evaluation cost, detailed below.

G.7 Reproducibility and Execution Details

Implementation and code availability. Our federated-learning implementation is based on the FedEM repository of Marfoq et al. (2021), available at `https://github.com/omarfoq/fedem`. The code used in our experiments is included in the supplementary material.

Tasks and heterogeneity. We use 100 clients on Fashion-MNIST (Xiao et al., 2017), EMNIST-Letters (Cohen et al., 2017), and next-character prediction on the natural-client Shakespeare benchmark. Fashion-MNIST and EMNIST-Letters use two 5×5 convolutional layers with 32 and 64 channels, max pooling, and a 2048-unit hidden layer. Shakespeare uses an eight-dimensional character embedding, a two-layer 256-unit LSTM, and an output decoder over the character vocabulary. The

Table 9: Operational cost on the uniform RTX 5090 repeated runs (means over seeds 2–5). Sum. is sequential in-process collection of 100 client summaries; fit is server profile-likelihood time. MB/client and FedBHG GB use decimal units. Saving is communication reduction relative to ten-round joint-all. CUDA is peak PyTorch allocation during selected continuation, not total VRAM. Op. $L + J$ counts operational independent models and distinct joint branches; Eval. J counts evaluation-only joint branches in addition to all 100 local candidates. Measurements use one host; payloads exclude serialization, protocol headers, transport framing and encryption. Routing-phase process RSS and distributed critical-path latency were not recorded.

Cell	Sum. (s)	Fit (s)	MB/client	FedBHG GB	Saving	CUDA MiB	Op. $L + J$	Eval. J
F-S3/CFL	3.80	4.36	8.69	10.16	41.5%	109	$46.5 + 1.0$	8.0
E-S3/IFCA	6.69	2.32	8.82	17.06	3.2%	119	$8.2 + 2.0$	11.8
Sh/CFL	31.22	.85	3.30	4.85	26.5%	286	$31.5 + 10.0$	55.5
Sh/IFCA	32.84	.82	3.30	3.66	44.5%	258	$49.5 + 1.8$	9.5

Table 10: Complete training-seed-1 development sweep after ten continuation rounds. Accuracy is sample weighted; loss is client uniform. L/J is the number of clients routed locally and jointly. Positive Δ values favor FedBHG over the better universal action for the corresponding metric. F and E denote Fashion-MNIST and EMNIST-Letters; Sh denotes Shakespeare.

Cell	Clusters	L/J	Acc. (%)	Δ best (pp)	Loss	Δ best	ρ	Agree
F-S3/CFL	1	47/53	87.91	+3.16	.713	+.002	.566	69%
F-S3/IFCA	2	33/67	84.41	+1.57	.754	+.012	.552	71%
F-S4/CFL	1	38/62	88.82	+2.84	.920	−.260	.874	90%
F-S4/IFCA	2	31/69	87.38	+2.36	.939	−.147	.841	91%
E-S3/CFL	16	4/96	65.77	−.44	1.565	+.004	−.018	34%
E-S3/IFCA	2	4/96	71.62	+1.27	.994	+.056	.783	69%
E-S4/CFL	16	10/90	70.68	−.01	2.224	+.005	.156	57%
E-S4/IFCA	2	6/94	82.25	+.83	1.219	−.024	.839	91%
Sh/CFL	12	33/67	42.36	+.03	2.100	+.009	.252	55%
Sh/IFCA	2	58/42	43.69	+.15	1.952	+.008	.596	79%

vision experiments contain two quantity–quality settings. In S3, client sample counts are log-normal and label-flip rates increase independently from zero to .7. In S4, counts are again log-normal, but corruption is inversely coupled to sample count and capped at .7. Label flips affect training data; held-out test labels remain uncorrupted. Each vision client receives a random half of the task's classes. Shakespeare retains its natural writer partition and uses length-80 sequences, stride-eight training windows, and nonoverlapping routing summaries.

Common-checkpoint training. We evaluate clustered federated learning (CFL) (Sattler et al., 2021) and IFCA (Ghosh et al., 2022) as upstream clusterers. Every base model is trained for 200 rounds with full client participation and one local epoch per round; every continuation lasts ten rounds, also with one local epoch per participating client. An epoch traverses that client's training iterator and can contain multiple minibatch optimizer steps. Vision uses SGD with learning rate .01 and batch size 128. Shakespeare uses Adam with learning rate 10^{-3} and batch size 128. The main development and optimization-seed study uses data seed 12345; the prospective partition check below uses data seeds 12346 and 12347. Within each cell and seed, every policy receives the identical base checkpoint.

Complete development sweep. Table 10 retains all ten cells, including adverse accuracy–loss tradeoffs. L/J counts locally and jointly routed clients; ρ and Agree are the universal-branch diagnostics defined above. For each metric, positive Δ best favors FedBHG over the better universal action, so the comparator can differ between accuracy and loss.

Synthetic development and confirmation runs. The S1 confirmation grid was chosen independently of the earlier development run and reproduced its no-deviation pattern. Confirmation runs also reproduced the directional-conflict and aggregation-weight patterns; the complete confirmation outcomes are reported above.

Evaluation data and trajectory reuse. All reported outcomes are computed on held-out test data.

Local trajectories are reused only when the client, checkpoint, optimizer and trajectory are identical.

Selection and random seeds. The performance evaluation has three stages. First, training seed 1 covers ten cells: Fashion-MNIST and EMNIST-Letters under S3/S4 and CFL/IFCA, plus Shakespeare under both clusterers. This sweep establishes breadth but is not used for uncertainty claims. After inspecting the seed-1 outcomes, we selected four repeated-evaluation cells and froze their design before running seeds 2–5: Fashion-MNIST S3/CFL, EMNIST-Letters S3/IFCA, and both Shakespeare clusterers. Combining seed 1 with the four untouched additional seeds gives five optimization seeds per repeated cell. We report paired 95% t intervals across those five seeds. Because the cells were selected after seed 1 and the data seed is fixed, the intervals are descriptive measures of optimization variation, not prospective confirmation or client-partition variation.

Third, before generating new data, we froze Fashion-MNIST S3/CFL and EMNIST-Letters S3/IFCA at training seed 1 and ran the complete common-checkpoint policy set on data seeds 12346 and 12347. These four partition–cell observations are reported in full and are not pooled with the optimization-seed intervals. Two partitions per task are a robustness check, not a basis for a partition-level confidence interval.

Gradient-model checks and refitting protocol. For the four prospective cells, the routing pass saves each base cluster's centered client-mean Gram matrix and estimated mean variances. We first require an independent refit to reproduce the recorded route. We then refit after every one-client deletion and after 500 nonparametric client bootstrap samples. Separately, 200 parametric Gram draws from the fitted spherical model estimate reference distributions for effective rank, leading-eigenvalue share, mean absolute pairwise cosine, radial coefficient of variation, and maximum radial deviation. The five statistics were declared before these checks. Bootstrap refits measure changes in the fitted local/joint assignments using only saved gradient summaries; continuation outcomes are not refit inputs.

Baseline tuning and implementations. For Fashion-MNIST and EMNIST-Letters, hyperparameters are selected once using training seed 0 and validation data only. Ten percent of each client's original training portion is held out during selection, giving a 72/8/20 train/validation/test split; the chosen configuration is then retrained on the full original 80% training portion and evaluated on the untouched 20% test portion for seeds 1–5. Ditto searches $\lambda \in \{.1, 1, 2\}$ and selects 2 on Fashion-MNIST and .1 on EMNIST-Letters. pFedMe compares two prespecified configurations and selects the CNN configuration ($\lambda = 15$, server mixing $\beta = 1$, five inner steps, and personalized learning rate .01) on both vision tasks. Shakespeare uses the author-standard settings ($\lambda = 1$ for Ditto and personalized learning rate .09 for pFedMe) without a dataset-specific search. Candidate selection never accesses the reported test data. In a preliminary implementation check, two seed-1 pFedMe vision runs using the released nonconvex learning rate .09 produced non-finite personalized losses. Neither contributes to reported results; the vision runs reported here use the released CNN learning rate .01 selected by the validation protocol.

We additionally compare the two repeated vision cells with Self-FL's uncertainty- controlled local training and aggregation (Chen et al., 2022) and pFedGraph's learned collaboration graph (Ye et al., 2023). These are closer to the paper's reliability and selective-collaboration mechanism than generic personalized objectives. They use the same clients, architectures, fixed data seed, 210-round horizon, training seeds, and validation protocol. With training seed 0 only, Self-FL selects its learning rate from $\{.001, .01\}$ and pFedGraph selects its data-size regularization from $\{.4, .8, 1.6\}$; all other settings are fixed before selection. The selected Self-FL learning rate is .001 on both tasks. The selected pFedGraph regularization is 1.6 on Fashion-MNIST and .4 on EMNIST-Letters. The two larger EMNIST candidates produce non-finite validation losses and are excluded by the predeclared finite-metric gate, not by their test accuracy. The selected configurations are then retrained on the full

original training split for seeds 1–5. Like Ditto and pFedMe, these methods train end to end and do not share FedBHG's round-200 checkpoint.

Cost-accounting convention. The main cost table uses the 16 repeated FedBHG executions from seeds 2–5 because all ran on the same RTX 5090 host; it does not pool their timings with the seed-1 P100 runs. The four prospective partition cells ran on the same RTX 5090 and form a separate confirmation of the recorded ranges. We separate four components. The 200-round CFL or IFCA checkpoint is upstream system training shared by every common-checkpoint policy. Routing then requires one streaming gradient-statistic pass and one summary upload per client. Operational continuation counts only the local trajectories and distinct joint branches selected by FedBHG. Evaluation cost separately counts the counterfactual local trajectories and joint branches needed to score FedBHG and all controls. The cost audit reads the frozen route, branch, timing, communication, and memory records; it launches no additional training.

Optimizer work and counterfactual evaluation. Operational FedBHG trains a single ten-round continuation trajectory per client: 1,820 optimizer steps in Fashion S3/CFL, 3,180 in EMNIST S3/IFCA, and 6,090 in each Shakespeare cell. The evaluation harness is larger by design. To reconstruct all matched policies, it trains all 100 local candidates and 7–63 distinct joint branches, using 4.09–6.67 times the operational continuation optimizer steps. Those counterfactual trajectories explain the experimental runtime and storage but are not part of a deployed FedBHG route.

Counting the upstream checkpoint and selected continuation together gives 38,220 optimizer steps for Fashion S3/CFL, 66,780 for EMNIST S3/IFCA, and 127,890 for either Shakespeare cell. Mean recorded base-training times are 455, 1,484, 1,187, and 2,112 seconds for Fashion S3/CFL, EMNIST S3/IFCA, Shakespeare/CFL, and Shakespeare/IFCA. The evaluator serializes counterfactual trajectories, so its wall-clock execution time cannot be relabeled as operational FedBHG latency.

The end-to-end baselines also use unequal amounts of optimizer work. Across five seeds, mean Ditto runtimes are 908, 1,147, and 1,449 seconds on Fashion-MNIST, EMNIST-Letters, and Shakespeare, with 80,220, 108,780, and 169,890 optimizer steps. Mean pFedMe runtimes are 5,121, 5,504, and 7,551 seconds; each run performs 2.1 million inner optimizer steps and 420,000 client-center updates. We report these counts to make the round-matched comparison auditable. On Fashion-MNIST and EMNIST-Letters, Self-FL averages 526 and 578 seconds with 21,820 and 23,180 optimizer steps; its post-warm-up controller selects one local step in every final run. pFedGraph averages 689 and 902 seconds with 38,220 and 66,780 steps, and allocates 3,339 and 3,390 MiB of CUDA memory. These measurements describe the audited implementations; they do not support a speed ranking across different objectives, update rules, or amounts of local work.

Evaluation protocol. FedBHG routes are fixed before continuation outcomes. All continuation policies within a FedBHG cell start from a common checkpoint, and every member of a joint branch receives the same trained model. Reported test results use the selected implementations and seeds; validation candidates with non-finite loss are excluded by the finite-metric gate described above. Fresh-partition diagnostics reuse saved gradient summaries. Counterfactual branches contribute to evaluation cost rather than operational FedBHG cost (Table 9). Scientific scope and untested settings are collected in Section I.

H Related Work

FedBHG studies summary-based voluntary common-model participation after a checkpoint and before continuation outcomes are observed. The contribution is the stated disclosure protocol applied to Gaussian shrinkage and conditional utility. Table 11 compares the information and action assumptions; the guarantees are not ordered by general strength.

H.1 Statistical Personalization and Uncertainty

The statistical foundation of FedBHG is a normal–normal empirical-Bayes model applied to common-checkpoint gradient summaries. Its interpolation between a local estimate and a population estimate is classical Bayesian shrinkage (Gelman et al., 2013; Murphy, 2012). Its common-action weight

Table 11: Closest precedents along the paper's claimed distinction. The final column identifies differences in scope without ranking the guarantees.

Work	Information and timing	Decision or guarantee	Boundary relative to FedBHG
Self-FL (Chen et al., 2022)	Inter- and intra-client uncertainty during iterative training	Uncertainty-controlled local steps and aggregation	Continuous personalization, not a fixed common-checkpoint action identity
Chen et al. (2023)	Population heterogeneity class before optimization	Minimax-rate transition between FedAvg and local training	Regime-level, not client-specific conditional comparison of observed actions
Even et al. (2022)	Distribution distances and streamed stochastic gradients	Gradient filtering with optimal sample complexity	Iterative filtering, not one binary action pair at a fixed checkpoint
FedGain (Zhang et al., 2026)	Scaling-law estimates and candidate client clusters	Effective-capacity clustering aimed at mitigating negative gain	No mean-gradient loss identity or pair-conditioned deviation theorem
Donahue–Kleinberg (Donahue & Kleinberg, 2021a;b)	Data sizes and ex-ante expected model errors	Voluntary model-sharing coalitions and stability	The half-threshold overlaps; utilities are ex ante rather than observed-gradient conditional
Coalition games (Cui et al., 2022; Ota et al., 2022; Chen et al., 2024; Zal et al., 2025)	Learned benefit graphs, gradient similarity, or trained local/coalition accuracy	Data-driven coalition structures, equilibria, and benchmark evaluations	Utilities are learned or measured outcomes, not a pre-continuation conditional loss identity
CoreFed (Chaudhury et al., 2022)	Client utilities of a learned shared predictor	Core-stable fairness for convex and smooth nonconvex losses	Stabilizes a predictor rather than choosing between two checkpoint actions
Hasan (2026)	Coalition benefit, costs, transfers, and pairwise validation gains	Potential-game stability, budget feasibility, and welfare bounds	Designs transfer-induced utility from coalition values; no conditional update-loss identity
FedBHG	Sample-mean update gradients at one common checkpoint	Binary action identity and pair-conditioned no-deviation theorem	Narrow oracle comparison; multi-round continuation is empirical

$(\sigma^2 + \varepsilon_j^2/n_j)^{-1}$ is likewise the inverse-total-variance weight of Gaussian random-effects models and meta-analysis (DerSimonian & Laird, 1986; Röver, 2020). In FL, precision-weighted aggregation has also used stochastic-gradient variance to weight client models (Reyes et al., 2025). The profile likelihood, ANOVA diagnostic, and EM characterization follow standard variance-component principles (Dempster et al., 1977; Laird, 1982; Wu, 1983; Searle et al., 1992). In particular, Röver (2020, Equations (17), (28), and (29)) gives the heteroscedastic mean and covariance used here after identifying study sampling variance with ε_j^2/n_j and between-study variance with σ^2. This correspondence includes uncertainty in the population mean.

Hierarchical Bayesian and empirical-Bayes personalization is already well developed in federated learning. Ozkara et al. connect personalized estimation to empirical Bayes and extend the same generative hierarchy to personalized learning (Ozkara et al., 2023). Self-FL is especially close conceptually: it uses inter-client and intra-client uncertainty to control local training and aggrega-

tion (Chen et al., 2022). pFedVEM derives confidence-aware aggregation and regularization through variational EM (Zhu et al., 2023); FedHB develops a broader hierarchical Bayesian treatment (Kim & Hospedales, 2025); FedPop models client parameters through common population parameters and random effects and quantifies posterior uncertainty (Kotelevskii et al., 2022); and Makhija et al. use functional-space Bayesian priors for heterogeneous client models (Makhija et al., 2024). A concurrent arXiv preprint, VANEB, studies heteroskedastic local estimators through variance-aware nonparametric empirical Bayes (Chang et al., 2026). Accordingly, FedBHG does not claim novelty for uncertainty-driven personalization, heteroskedastic shrinkage, or confidence-weighted aggregation.

Compared with the Bayesian methods above, FedBHG differs in modeled object, timing, and guarantee. Its random effects are sample-mean update gradients measured at one common checkpoint, and the hierarchy compares one fixed client action with one fixed common action rather than maintaining a continuous or iteratively updated personalized posterior. The participation characterization uses these posterior utilities under the single-round disclosure model. Neither conditioning on client information nor using gradient variance is itself the novelty.

End-to-end methods such as Ditto and pFedMe optimize personalized objectives throughout training (Li et al., 2021; Dinh et al., 2020). They and Self-FL are empirical comparators, whereas the common-checkpoint local and joint branches isolate FedBHG's binary routing decision. The comparison therefore does not make FedBHG a general replacement for personalized FL.

H.2 Local, Joint, and Selective Collaboration

The local-versus-collaborative tradeoff has an established theoretical literature. Chen et al. identify a sharp, up-to-constants transition between regimes in which FedAvg and pure local training are minimax-rate optimal, and show that a dichotomous choice can be rate-optimal (Chen et al., 2023). Even et al. filter and aggregate other agents' stochastic gradients according to distributional proximity, attaining an optimal bias–variance tradeoff in sample complexity (Even et al., 2022). These theorems show that collaboration should depend on heterogeneity and estimation error. Their guarantees concern population regimes, known proximity information, or iterative stochastic optimization; they do not derive FedBHG's client-specific conditional identity for two actions formed from one observed checkpoint.

Selective-collaboration methods address who should train together. FedCollab partitions clients into non-overlapping coalitions using distribution distances and data quantities (Bao et al., 2023); pFedGraph learns client-specific collaboration strengths from model similarity and data size (Ye et al., 2023); and DPFL constructs a collaboration graph under communication and resource constraints (Kharrat et al., 2025). MeritFed assigns dynamic target-specific weights to beneficial collaborators using stochastic gradients (Tupitsa et al., 2026). Most directly, FedGain targets negative collaboration gain through an effective-capacity criterion derived from a scaling-law model (Zhang et al., 2026). These works already study beneficial collaborator selection and negative transfer.

Other close methods make the local-or-collaborative choice explicit. FedFomo uses a client validation set to select and weight other clients' models (Zhang et al., 2021). HCCT compares collaborative and independent training through client-level generalization bounds and learns non-overlapping groups from gradient similarity and data quantity (Sun et al., 2025). All Federated or Not predicts personal-model improvement before a new FL run using a surrogate trained on prior FL outcomes, then uses it as utility in a cross-silo coalition game (Li et al., 2024). These are direct precedents for the decision problem. They use validation performance, generalization bounds, or predicted model performance rather than FedBHG's fixed common-checkpoint conditional loss identity.

Clustered methods such as CFL and IFCA learn several shared models during federated training (Sattler et al., 2021; Ghosh et al., 2022). FedBHG can operate inside an upstream cluster, but it neither learns an arbitrary partition or graph nor adapts its route during continuation. It estimates one score per client at a fixed checkpoint and routes that client either to a singleton continuation or to one common continuation in the existing cluster. Theorem C.3 concerns a single comparison between curvature-matched actions. The implemented branches use ordinary multi-round optimizers, so neither convergence of those branches nor optimality over general collaboration structures follows from the theorem; the neural experiments test that transfer empirically. The pFedGraph comparison in Section G is therefore a competitiveness test against a close graph-selective method, not an action-matched test of FedBHG's route.

H.3 Voluntary Sharing and Coalition Deviations

Model-sharing games are FedBHG's closest analytic ancestor. Donahue and Kleinberg derive expected MSE for mean estimation and linear regression and use it to study voluntary coalitions, stability, optimality, and fairness (Donahue & Kleinberg, 2021a;b; 2023). Their utilities average over sample and client-parameter draws rather than condition on a realized checkpoint summary. Their sample-size classification also coincides algebraically with FedBHG's half-threshold under common local variance (Donahue & Kleinberg, 2021a); neither the bias–variance tradeoff nor the threshold is new here. Blum et al. study a different ex-ante strategic variable—the sampling burden that agents contribute to a common learning objective—and characterize stable and envy-free allocations (Blum et al., 2021).

Later work already makes coalition utilities empirical or learned. Collaboration Equilibrium constructs a directed benefit graph using Pareto optimization and validation performance, then extracts collaboration coalitions through graph operations (Cui et al., 2022). Ota et al. define pair synergy through classification improvement or gradient cosine similarity and optimize a coalition structure (Ota et al., 2022). FedEgoists builds a directed data-complementarity graph from local-model information and combines it with reported competitive relationships to exclude free riders and conflicts (Chen et al., 2024). Zal et al. use trained local and coalition accuracy as utilities and exhaustively evaluate Nash equilibria on public datasets (Zal et al., 2025). Their subsequent study includes communication costs and incentives (Zal et al., 2026). Empirical real-data stability analysis is therefore not itself the distinction claimed here.

Other game-theoretic FL papers solve different strategic problems. CoreFed treats a core-stable fairness condition for a learned shared predictor rather than a partition among continuation actions (Chaudhury et al., 2022). FedEgoists addresses competition, Gong et al. permit overlapping coalitions and personalized mixtures, and LEAP jointly handles hierarchical association and communication resources (Chen et al., 2024; Gong et al., 2026; Lu et al., 2024). Hasan's recent transferable-surplus model is the closest direct stability comparison: coalition learning benefit, system and participant costs, and monetary transfers induce a pairwise exact-potential game with Nash/individual stability, budget-feasibility, and welfare guarantees; its experiments estimate pair values from validation gains (Hasan, 2026). That work designs affordable incentives after coalition values are specified or estimated. FedBHG instead derives a client learning-loss comparison for one fixed pair of actions and has no transfer, budget, or welfare theorem.

Information-dependent coalition evaluation is older than these FL formulations. Wilson (1978) defines communication-dependent blocking through conditional utility; Dutta & Vohra (2005) studies credibility and information inferred from objections. The disclosure protocol in Section B.3 excludes endogenous proposal selection and repeated revelation, and does not claim their strategic credibility guarantees. It defines which summaries enter each pre-report assessment of a common-model proposal. The resulting centroid argument excludes an all-weak, one-strict improvement under those assessments. A supplied benefit graph is unnecessary for this idealized comparison, but unrestricted privately informed objections and the fitted training algorithm remain outside the guarantee.

H.4 Participation Characterizations and Data Exchange

Theorem B.12 concerns a common action accepted by every member for every observed summary, within fixed positive linear aggregators. Bayes optimality for estimating the population mean does not itself answer this participation question, because each client cares about its own latent target. Conversely, the theorem does not characterize data-dependent aggregators or imply a new posterior estimator. The personalized weights in Donahue & Kleinberg (2021a) coincide with the scalar posterior coefficients on the overlapping variance model; their personalized actions and ex-ante comparisons do not establish the union-conditioned common-action theorem here.

Core-stable data exchange also has different feasible outcomes. Song et al. (2025) study bilateral fractional exchanges with sharing costs. Their outcome is an exchange matrix, whereas our outcome is a partition into common estimates and each comparison uses realized summaries. Their existence theorem therefore does not settle the participation construction or the covariance obstruction analyzed here.

I Limitations and Deployment Considerations

Scope. The no-deviation guarantee concerns posterior-risk comparisons with known variance components, prescribed common actions and the specified summary-disclosure rule. Expected-loss stability additionally requires shared quadratic curvature and Newton updates. FedBHG estimates its components from full-cluster summaries and continues with SGD or Adam; its accuracy, loss and client-harm outcomes are empirical claims. The experiments assess predictive performance under the tested protocols, not strategic stability of trained models or validity of the Gaussian hierarchy.

I.1 From Newton Updates to Multi-Round Training

Theorem C.3 compares the full local and common Newton updates. Theorem C.7 extends the sign when the quadratic decision margin dominates a local Taylor remainder. Remark C.10 shows that even one-dimensional quadratic trajectories can reverse their ordering before reaching those endpoints. A guarantee for arbitrary learning rates, optimizers or training horizons must therefore account for the trajectories, beyond their checkpoint gradients. The multi-horizon neural experiments measure this transfer directly.

I.2 Disclosure and Statistical Calibration

The single-round protocol in Section B.3 compares risks before any votes, using each incumbent/proposal summary union. Raw data, training history, previous messages and strategic proposals can convey additional information. Extending the theorem to these signals requires posterior sufficiency or a larger information model; strategic voting would require a corresponding equilibrium analysis. The pair-dependent comparisons need not form a transitive ranking of coalitions.

The server's fitted components use all base-cluster summaries and can reveal information outside an oracle client's comparison transcript. Holding fitted weights fixed during continuation does not remove this data dependence. Applying the rule within upstream clusters also leaves cross-cluster deviations outside the guarantee. Proposition E.1 shows that even the ideal profile maximum can concentrate on a biased variance and the wrong route. Bootstrap refitting measures sensitivity around that fitted rule, rather than accuracy relative to the oracle. The ANOVA expression corrects the equal-variance degrees of freedom; the reported training runs retain the profile estimator.

I.3 Empirical Scope and Selection

The repeated-run intervals describe optimizer variation in four cells selected after the ten-cell seed-1 sweep, on fixed client partitions. The four additional observations use prospectively specified new partitions for two vision tasks. Two draws per task support a robustness check, but not a partition-level confidence interval, and do not assess new language partitions. The score-specific advantage is concentrated in Fashion S3/CFL; the new EMNIST routes again match simpler coverage-controlled routes. Section G.7 records the evaluation-selection protocol and the reported development, repeated-seed, and prospective-partition stages.

End-to-end comparisons match 210 communication rounds, while trajectories and optimizer work differ. Ditto and pFedMe cover all three tasks; Self-FL and pFedGraph use small validation grids and one fixed partition on each vision task. Broader claims require more partitions, partial participation, larger architectures and additional selective-collaboration baselines such as FedCollab and FedGain. Common-checkpoint comparisons isolate continuation choices; end-to-end comparisons assess final performance under their respective training procedures.

I.4 Cost and Scale

The cost study measures summary collection, fitting, communication payloads, optimizer work and continuation CUDA allocation in single-machine executions. Network latency, serialization, transport overhead, routing-phase process RSS, stragglers and distributed critical-path latency remain unmeasured. A full-dimensional summary costs approximately one model upload per client, limiting savings when nearly everyone shares. The current Gram-based fitter uses $O(Md + M^2)$ storage and

$O(M^2d + GM^2)$ work for G profile evaluations. Measurements at $M = 100$ establish feasibility at that scale; larger client populations require separate evaluation.

I.5 Covariance and Variance Estimation

The scalar implementation replaces gradient covariances by trace averages. Direction agreement suffices for the general-covariance theorem, but the scalar score does not test that condition. The counterexample establishes that conflicting directions can defeat every partition; it does not estimate how frequently this happens in practice or classify all intermediate covariance families. A computationally feasible diagnostic of directional disagreement in high-dimensional neural models remains open.

The observed gradients reject the fitted spherical model. Upstream clustering can reduce heterogeneity without ensuring Gaussianity or unimodality. Although average route sensitivity is low, near-threshold clients can change assignments in roughly half of bootstrap refits. Few independent clients or effective coordinates also limit variance estimation. The synthetic experiments enumerate deviations under a known posterior and numerical tolerance; their frequencies are empirical, while Proposition C.8 states deterministic margin conditions. Neither establishes a finite-sample concentration guarantee for the fitted neural router. The largest common gain Γ measures severity in the simulation's posterior-risk units; other utilities and disclosure rules require their own approximation criterion.

I.6 Fairness and Voluntary Participation

Posterior-risk stability leaves resource inequality, bargaining power, representation and demographic fairness uncontrolled. Precision weighting can concentrate influence on clients with reliable estimates. Deployment objectives involving influence, group robustness, monetary cost, communication latency or institutional constraints require corresponding utilities or allocation constraints. A participation theorem under posterior MSE is not legal or regulatory assurance.

I.7 Privacy of Routing Statistics

FedBHG transmits $(\widehat{\boldsymbol{\theta}}_j, \mathsf{SSE}_j, n_j)$. The mean gradient reveals directional information about a local objective, and the dispersion statistic reports within-client gradient variability. These summaries are not privacy neutral. Differential-privacy noise would change the estimated variances precisely where routing is most sensitive, near $B_j = 1/2$. A private version requires recalibrating the route and aggregation weights for those perturbed summaries.

J Discussion and Conclusion

Under the specified disclosed-summary comparison, a singleton-or-pool partition excludes every all-weak, one-strict improvement. Its proof aligns posterior coefficients with the aggregate's centroid. For at least three pooling-eligible clients, participation for every observation forces precision weights within fixed positive linear aggregation. Direction agreement extends the construction to general covariances; conflicting directions can instead make every partition vulnerable, already with three clients and two dimensions. These statements leave the intermediate covariance classes and data-dependent common actions open.

FedBHG applies the scalar construction to checkpoint-based selective continuation. Synthetic experiments show that no-deviation is sensitive to route estimation, directional agreement, and aggregation weights in the controlled models. Neural experiments test predictive transfer under fitted routing and ordinary training: across four selected task/clusterer settings on fixed client partitions, FedBHG improves on both universal continuations in accuracy in 19/20 runs and loss in 17/20. Four fresh-partition observations improve accuracy in all four and loss in three.

The clearest incremental accuracy gains over simpler routes occur on Fashion S3/CFL. Client-level ranking and threshold agreement remain informative after multi-round training, including at checkpoints whose gradients depart from the spherical working model. The experiments thus connect the participation construction to predictive benefits in the tested settings. Section I details the

statistical, training and deployment limitations. Extending the participation guarantee to estimated routes and ordinary multi-round training is an open direction.