A DECENTRALIZED LENS ON CENTRALIZED FEDERATED LEARNING: UNIFYING FL VIA INEXACT DOUGLAS–RACHFORD SPLITTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative training without pooling client data, but data heterogeneity remains a major challenge, especially when combined with partial participation and inexact local computation, whose joint effects are difficult to capture in a unified analysis of existing methods. To analyze these interactions, we develop an operator-splitting framework on the full client product space with a consensus constraint. Within this framework, a target field defines the residual and energy, while a conditional mean-square error criterion connects the target dynamics to actual client computations. This connection enables convex convergence guarantees under an additional summable conditional-mean condition, as well as nonconvex stationarity guarantees through expected forward–backward-envelope descent. Under local Kurdyka–Łojasiewicz (KL) geometry, almost-sure boundedness, and explicit process and error-tail conditions, the nonconvex analysis further yields finite length and convergence of the full sequence. Applying the abstract algorithm to federated optimization yields the FedExSplit (FES) family, with variants defined by the target potential. We then use an exact heterogeneous quadratic construction to explain how its mean-anchored variant (FES-MA) can suppress the effect of client disagreement on the mean update in a specified parameter regime. Controlled one-step experiments support this mechanism, while training experiments show that FES-MA improves validation accuracy over FedAvg by approximately 1.7 percentage points on FEMNIST and 2.6 percentage points on CIFAR-100 under high heterogeneity ($\alpha_{\mathrm{LDA}} = 0.1$).

1 INTRODUCTION

Federated learning trains a shared model without pooling client data. FedAvg alternates local training with server aggregation (McMahan et al., 2017), but heterogeneous objectives cause client drift (Karimireddy et al., 2020). FedOpt adapts server updates (Reddi et al., 2021), and FedExP uses extrapolation (Jhunjhunwala et al., 2023). With persistent client iterates, the shared average is insufficient: states with the same average can produce different responses.

The product-space formulation (Pierra, 1984) retains individual variables and their disagreement under a consensus constraint, providing a *decentralized lens on centralized federated learning*. We ask how to retain client history while controlling its interaction with heterogeneous objectives under partial participation and inexact computation.

SCAFFOLD and FedDyn counter drift through control variates and dynamic regularization (Karimireddy et al., 2020; Acar et al., 2021). FedSplit treats convex problems with inexact solves (Pathak & Wainwright, 2020); FedDR gives nonconvex stationarity bounds under sampling and inexact proximal evaluations (Tran-Dinh et al., 2021); iFedDR uses relative resolvent errors and correction under maximal monotonicity (Pethick et al., 2025). We control full-update errors beyond proximal approximation.

DANE uses gradient-corrected local models at a shared iterate (Shamir et al., 2014). INEXACT-DANE permits approximate solves (Reddi et al., 2016), and FedDANE combines sampled corrections and clients with inexact computation (Li et al., 2019). These developments motivate a comple-

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.

Table 1: Selected guarantees of related methods. Assumptions differ by row; full conditions and proof qualifications appear in Appendix A.

Method	Participation	Computation / error	Setting	Selected guarantee
DANE (2014)	Full	Exact local solves	Strongly convex	Linear objective-gap decay (Thm. 4)
FedDANE (2019)	Random sampling	Sampled correction; relative distance error	Nonconvex[d]	Expected stationarity bound (Thms. 5–6)
iFedDR (2025)	Full	Relative resolvent residual + correction	Convex	Whole sequence (Thm. 3.1)
FedExProx (2026)	Uniform sampling	Exact proximal solves	Convex + PL[a]	Linear gap decay; a.s. finite length / whole sequence[†] (Thm. D.1)
FIAELT (2024)	Full	Pathwise relative proximal error	Nonconvex + KL[b]	Finite length; whole sequence; rates (Thm. 2)
FedADMM-MM (2025)	Random sampling	Summable objective-gap error; server MM	Nonconvex	A.s. stationary cluster points; stationarity bound (Thms. 1–2)
Ours	Budget-admissible sampling	Conditional mean-square full-update error	Nonconvex	Vanishing residual; stationary cluster points; stationarity bound (Thm. 2.1)
Ours	Same	Same; stronger tolerance decay	Nonconvex + KL-type[c]	**A.s. finite length; whole sequence** (Thm. 2.3)

[a]Convex client objectives, interpolation, and envelope PL. [b]Bounded trajectories and convex regularization; proof qualifications in Appendix A. [c]Almost-sure boundedness, conditional concavity, and stronger error-tail decay; Section 2.3. [d]Bounded gradient dissimilarity and joint restrictions on sampling, accuracy, and parameters. All rows retain the cited parameter conditions; FedADMM-MM uses Assumptions 1–4. [†]Direct consequences of reported bounds. Stochastic gap rates are in expectation. Cluster points are conditional on existence. "a.s." means almost surely. Our sampling and solver choices must satisfy the joint error budget.

mentary question: how does persistent client disagreement affect local responses and their aggregate under stochastic, inexact updates?

Our framework builds on Douglas–Rachford splitting (DRS) (Lions & Mercier, 1979; Eckstein & Bertsekas, 1992) and its relative-error variants (Eckstein & Yao, 2018; Svaiter, 2019). A target field defines the consensus dynamics; a conditional mean-square criterion connects them to actual outputs. The analysis combines stochastic quasi-Fejér and supermartingale techniques (Robbins & Siegmund, 1971; Combettes & Pesquet, 2015) with forward–backward-envelope (FBE) descent (Themelis et al., 2018). For nonconvex sequence convergence, we adapt the stochastic Kurdyka–Łojasiewicz (KL) conditioning step of Fest et al. (2026) and use local FBE geometry in the finite-length argument (Attouch et al., 2013).

Applying the framework yields FedExSplit (FES): FES-DR recovers FedDR under matching update rules; mean-anchored FES-MA evaluates gradients at the client-state average. FES-MA exposes DANE as a special case in a persistent product space, where natural-residual relaxation retains history and controls its coupling to heterogeneous responses. This interpretation and joint error analysis follow from the framework, with FES-MA as an application.

Table 1 compares selected guarantees; Appendix A gives the full comparison.

Contributions. We establish convex sequence convergence under a summable conditional-mean bias bound, nonconvex stationarity through expected FBE descent, and nonconvex finite length and whole-sequence convergence under local KL geometry, almost-sure boundedness, and the process and error-tail conditions in Section 2.3. We derive the full-state implicit-error identity, verify the budget for sampled convex FES-DR with exact solves, and give explicit FES-MA error constants with a feasible regime. An exact quadratic construction identifies when FES-MA reduces disagreement-induced aggregate perturbations while retaining history. Controlled one-step validation supports this mechanism, complemented by FEMNIST and CIFAR-100 experiments.

Algorithm 1 Unified stochastic inexact Douglas–Rachford splitting method.

1: Initialize $x^0 \in \mathcal{H}$; fix $\beta > 0$, $a_0 \in (0, 1]$, and the budget in (2.2).
2: **for** $k = 0, 1, \ldots$ **do**
3: Choose predictable $\alpha_k \in [a_0, 1]$ and evaluate $e_k = e_\beta(x^k)$.
4: Generate x^{k+1} such that $\Theta_k := M_\beta(x^{k+1}) - M_\beta(x^k) + \alpha_k e_k$ satisfies (2.2).
5: **end for**

Our guarantees concern server aggregation; fully decentralized protocols (Zehtabi et al., 2025) require additional communication and consensus-error bounds.

2 A UNIFIED FRAMEWORK FOR FEDERATED OPTIMIZATION

2.1 PROBLEM AND ABSTRACT ALGORITHM

Let $\mathcal{H} = \mathbb{R}^{Nd}$, let $\Omega = \{\mathbf{1}_N \otimes w : w \in \mathbb{R}^d\}$ be the consensus subspace, and let Π and $Q = I - \Pi$ be the consensus and disagreement projectors. Following the product-space formulation of Pierra (1984), we consider

$$\min_{x \in \mathcal{H}} P(x) + \delta_\Omega(x). \tag{2.1}$$

Here δ_Ω is the indicator of Ω; the standard FL objective corresponds to $P(x) = \sum_i f_i(x_i)$. Throughout, $P \in C^{1,1}$, $g = \nabla P$ is L-Lipschitz, and $\inf_\Omega P > -\infty$. The client state x^k need not be in consensus; no trajectory boundedness is included in these basic conditions.

For fixed $\beta > 0$, write $M_\beta(x) = x + \beta g(x)$, $T_\beta(x) = \Pi(x - \beta g(x))$, and $e_\beta(x) = x - T_\beta(x)$. Given a filtration $(\mathcal{F}_k)$, let x^k and $\alpha_k \in [a_0, 1]$ be $\mathcal{F}_k$-measurable, with $a_0 > 0$, and write $\mathbf{E}_k = \mathbf{E}[\cdot \mid \mathcal{F}_k]$ and $e_k = e_\beta(x^k)$. Algorithm 1 implements $M_\beta(x^{k+1}) = M_\beta(x^k) - \alpha_k e_k + \Theta_k$ under the conditional error criterion

$$\mathbf{E}_k\|\Theta_k\|^2 \le \rho\|e_k\|^2 + \varepsilon_k^2, \qquad \rho \ge 0, \quad \varepsilon_k \ge 0, \quad \sum_k \varepsilon_k^2 < \infty. \tag{2.2}$$

The budgets are deterministic. This criterion concerns the distribution of the actual output, not a pathwise relative bound or a zero-mean requirement.

For convex P, or for $\beta L < 1$, set $J_\beta = M_\beta^{-1} = \mathrm{prox}_{\beta P}$ and $z^k = M_\beta(x^k)$. The update becomes perturbed relaxed Douglas–Rachford splitting (DRS) for $P + \delta_\Omega$ (Bauschke & Combettes, 2017, Chapter 26):

$$z^{k+1} = z^k + \alpha_k \left[\Pi(2J_\beta z^k - z^k) - J_\beta z^k\right] + \Theta_k. \tag{2.3}$$

Appendix B.2 gives the change of variables and its nonconvex justification.

The residual satisfies $\|e_\beta(x)\|^2 = \|Qx\|^2 + \beta^2\|\Pi g(x)\|^2$, so its zeros are consensus critical points. The nonconvex analysis uses the full-space forward–backward envelope (FBE), specialized to the consensus indicator (Themelis et al., 2018)

$$\mathcal{E}_\beta(x) = P(x) - \langle g(x), e_\beta(x)\rangle + \frac{1}{2\beta}\|e_\beta(x)\|^2. \tag{2.4}$$

2.2 STATIONARITY AND CONVEX CONVERGENCE

For the FBE analysis, a sufficient parameter range is

$$t = \beta L \le 1/16, \qquad c_1(t)\sqrt{\rho} + c_2(t)\rho \le a_0/8, \tag{2.5}$$

where $c_1(t) = 1 + 2t/(1-t) + (1+t)/(1-t)^2$ and $c_2(t) = (1+t)/(2(1-t)^2)$.

Theorem 2.1 (Stationarity)**.** *Under the basic conditions, Algorithm 1,* (2.2)*, and* (2.5)*, assume* $\mathbf{E}|\mathcal{E}_\beta(x^0)| < \infty$*. Then the FBE has a finite almost-sure limit,* $\sum_k \mathbf{E}\|e_k\|^2 < \infty$*, and* $e_k \to 0$ *almost surely. Every existing cluster point is critical for* (2.1)*, and* $\min_{0 \le k \le n} \mathbf{E}\|e_k\|^2 = O((n+1)^{-1})$.

The descent estimate and proof appear in Appendices B–C. Neither KL nor bounded trajectories are required, and cluster-point existence is not guaranteed.

Scope of the parameter conditions. Equation (2.5) is conservative and does not characterize stability; empirical choices may be more aggressive. The asymptotic conclusions permit a finite initial phase: if the corresponding theorem's assumptions hold from a deterministic index K, with fixed P, β and the required integrability at x^K, reindexing its proof suffices. The empirical settings below are not certified by this observation.

Theorem 2.2 (Convex convergence)**.** *Suppose the basic conditions and* (2.2) *hold, P is convex, and $\mathcal{X}^\star = \operatorname{argmin}_\Omega P \neq \varnothing$. Require the additional deterministic mean budget*

$$\|\mathbf{E}_k \Theta_k\| \leq \delta_k, \qquad \delta_k \geq 0, \qquad \sum_k \delta_k < \infty. \tag{2.6}$$

For any $\beta > 0$, assume $\rho < a_0(2 - a_0)$ and $\mathbf{E}\|M_\beta(x^0) - M_\beta(x^\star)\|^2 < \infty$ for some deterministic $x^\star \in \mathcal{X}^\star$. Then x^k converges almost surely to a point in $\mathcal{X}^\star$ and $\sum_k \|e_k\|^2 < \infty$ almost surely. With $y^k = T_\beta(x^k)$ and $P^\star = \min_\Omega P$,

$$\min_{0 \leq k \leq n} \mathbf{E}\|e_k\|^2 = O((n+1)^{-1}), \qquad \min_{0 \leq k \leq n} \mathbf{E}[P(y^k) - P^\star] = O((n+1)^{-1/2}). \tag{2.7}$$

See Appendix C for the proof. Square-summable tolerances suffice, without (2.5) or KL. The mean condition (2.6) does not follow from (2.2); centered errors give $\delta_k = 0$.

2.3 Nonconvex convergence

The conditioning step of Fest et al. (2026, proof of Lemma 4.2) motivates the following condition on an integrable supermartingale $V_k \to V_\infty$:

$$\varphi(V_k - V_\infty) - \mathbf{E}_k \varphi(V_{k+1} - V_\infty) \geq \varphi'(V_k - V_\infty)(V_k - \mathbf{E}_k V_{k+1}). \tag{2.8}$$

Here we require nonnegative gaps $V_k - V_\infty$ and this inequality almost surely for every k, with $\varphi(t) = t^{1-\theta}/(1-\theta)$. At zero gaps we use $\varphi'(0) = +\infty$ and $0 \cdot (+\infty) = 0$; thus the condition also forces $V_k - \mathbf{E}_k V_{k+1} = 0$ there. For the process below, the global FBE lower bound and supermartingale property ensure integrability of the quantities on the left (Appendix D.4). The limiting energy may be random.

Interpretation and a verifiable local case. A sufficient case is almost-sure containment in a deterministic compact region $\mathcal{K}$ whose critical set lies in a connected set of local minimizers with common FBE value ℓ, with $\mathcal{E}_\beta \geq \ell$ on $\mathcal{K}$. Then $V_\infty = \ell$ and ordinary conditional concavity verifies (2.8); Appendix D.5 gives the justification. Nonisolated minimizers are allowed. The general conditions below also allow multiple critical components.

For the general result, let $\mathcal{K} \subseteq \mathcal{H}$ be a deterministic closed region, possibly $\mathcal{H}$ itself, and assume

$$\mathbb{P}\left(x^k \in \mathcal{K} \text{ for all } k,\ \sup_k \|x^k\| < \infty \right) = 1, \qquad \mathcal{S} = \{z \in \mathcal{K} : e_\beta(z) = 0\}. \tag{2.9}$$

The trajectory bound may be random; no deterministic compact set need contain all paths. The region may contain critical components with different energies. Require the local KL property of the FBE with a common deterministic exponent $\theta \in [1/2, 1)$: for every $z \in \mathcal{S}$ there are a neighborhood $\mathcal{U}_z$ and constants $c_z, \eta_z > 0$ such that

$$\operatorname{dist}(0, \partial \mathcal{E}_\beta(x)) \geq c_z \big(\mathcal{E}_\beta(x) - \mathcal{E}_\beta(z)\big)^\theta \quad \text{if } x \in \mathcal{U}_z, \quad 0 < \mathcal{E}_\beta(x) - \mathcal{E}_\beta(z) < \eta_z. \tag{2.10}$$

The additive tolerances satisfy

$$\sum_k \varepsilon_k < \infty, \qquad \sum_k \left(\sum_{j=k}^{\infty} \varepsilon_j^2 \right)^\theta < \infty. \tag{2.11}$$

Geometric tolerances suffice. Semialgebraic geometry supplies local power KL estimates and compact-set uniformization (Bolte et al., 2007; 2014); it does not itself imply (2.8).

With the constants $a, c_\varepsilon > 0$ of Appendix B, define $r_k = c_\varepsilon \sum_{j \geq k} \varepsilon_j^2$ and $V_k = \mathcal{E}_\beta(x^k) + r_k$. Expected FBE descent gives

$$V_k - \mathbf{E}_k V_{k+1} \geq a\|e_k\|^2. \tag{2.12}$$

Theorem 2.1 identifies its limit V_∞ with the limiting FBE value. We impose (2.8) on this process; no KL property is assigned to the error-tail correction.

Algorithm 2 FedExSplit family.

1: Initialize $(x_i^0)_{i=1}^N$, their average, and $\beta, \gamma_k, \mu_k > 0$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Select $\mathcal{S}_k$ and provide the average and correction information.
4: Each selected client forms v_i^k, computes u_i^{k+1}, and sets $x_i^{k+1} = u_i^{k+1}$.
5: Keep unselected states unchanged; set $\bar{x}^{k+1} = \bar{x}^k + N^{-1} \sum_{i \in \mathcal{S}_k} (u_i^{k+1} - x_i^k)$.
6: **end for**

Theorem 2.3 (Convergence of the full sequence)**.** *Under the conditions of Theorem 2.1, assume* (2.9)–(2.11) *and* (2.8) *for the process and power function specified above. Then*

$$\sum_k \|e_k\| < \infty, \qquad \sum_k \|x^{k+1} - x^k\| < \infty \quad \text{a.s.} \tag{2.13}$$

Consequently $x^k \to x^\infty \in \mathcal{S}$ *almost surely, with* $0 \in g(x^\infty) + N_\Omega(x^\infty)$ *and* $\mathcal{E}_\beta(x^\infty) = P(x^\infty) = V_\infty$.

Appendix D proves the result using KL uniformization, conditional concavity, and the additive-tail estimate.

3 Federated algorithms under the unified scheme

3.1 Persistent-state template and operator choices

FES denotes the FL algorithm family induced by the abstract DRS scheme; target potentials specify its variants, with the corresponding implementation fields given below. Let $F(x) = \sum_i f_i(x_i)$, $\Phi(w) = N^{-1} \sum_i f_i(w)$, and $\bar{x}^k = N^{-1} \sum_i x_i^k$. A selection projector D_k identifies the participating clients. Each selected client forms a center using an implementation field B_k and a correction estimate $\widehat{C}_k$, then solves a local subproblem:

$$v_i^k = x_i^k + \beta(B_k(x^k))_i - \gamma_k\big(x_i^k - \bar{x}^k + \beta(\widehat{C}_k)_i\big), \qquad u_i^{k+1} \approx \operatorname{prox}_{\beta\mu_k f_i}(v_i^k).$$

The two variants, FES-DR and mean-anchored FES-MA, use (G_1, H_1) and (G_2, H_2), respectively:

Method	$(B_k(x))_i$	Correction target $(H_j(x))_i$	Target potential $P_j(x)$
FES-DR	$\nabla f_i(x_i)$	$N^{-1} \sum_j \nabla f_j(x_j)$	$F(x)$
FES-MA	$\nabla f_i(\bar{x})$	$\nabla\Phi(\bar{x})$	$N\Phi(\bar{x})$

Thus $\nabla P_1 = G_1$ and $\nabla P_2 = H_2$, although G_2 need not be conservative. The potentials agree on the consensus subspace,

$$P_1(\mathbf{1}_N \otimes w) = P_2(\mathbf{1}_N \otimes w) = N\Phi(w), \tag{3.1}$$

so they define the same constrained FL task despite differing away from consensus. The server average is maintained from actual client-state increments, as in Algorithm 2.

FES-MA, residual relaxation, and DANE. For P_2, write $w = \bar{x}$, $q_i = x_i - w$, and $h = \nabla\Phi(w)$. With exact correction, $(e_\beta(x))_i = q_i + \beta h$ and

$$v_i = w + \beta\nabla f_i(w) - \gamma\beta h + (1 - \gamma)q_i.$$

Thus γ relaxes the natural residual, acting on both disagreement and the consensus gradient; DANE's gradient scale acts only on the latter. At $\gamma = \mu_k = 1$, full participation, exact solves, and matching equal-weight aggregation, the mean trajectory recovers DANE with gradient scale one and proximal penalty $1/\beta$ (Shamir et al., 2014). For $\gamma \neq 1$, the center retains client history, whose interaction with heterogeneous responses is analyzed in Section 4. Under sampling, Algorithm 2 also retains unselected states in the full-state mean, whereas FedDANE aggregates sampled outputs (Li et al., 2019). The framework accounts jointly for selection, correction, and local-solve errors.

Write $\eta_{i,k} = u_i^{k+1} + \beta\mu_k\nabla f_i(u_i^{k+1}) - v_i^k$ on selected clients and zero otherwise. The local objective is strongly convex for convex f_i, and also for L-smooth nonconvex f_i when $\beta\mu_k L < 1$. These equation residuals quantify approximate solution; a fixed number of local steps alone does not certify (2.2).

Verification within the abstract framework. For convex FES-DR, exact local solves, $\mu_k = 1$, exact H_1, and equal conditional inclusion probability p, the actual error is $\Theta_k = \gamma(pI - D_k)e_k$ with $\alpha_k = p\gamma$. Its conditional mean is zero and its conditional second moment is $\gamma^2 p(1-p)\|e_k\|^2$. For $0 < \gamma \leq 1$ this verifies the error and parameter conditions of Theorem 2.2; initial integrability is still required.

For FES-MA, Proposition 3.1 bounds the full implicit error, including changes in every block of H_2 when the average moves. Theorems 2.1 and 2.3 apply under their respective error, parameter, and local conditions.

Connections to existing algorithms. FedDR. With $\mu_k = 1$, exact local solves and correction, consistent initialization, and matching aggregation, the FES-DR change of variables $y_i = x_i + \beta\nabla f_i(x_i)$ recovers FedDR (Tran-Dinh et al., 2021) without server regularization. Here γ_k corresponds to its existing relaxation parameter. Other proximal scales and inexact solves require the error accounting below. **FedProx.** Setting $B_k = \widehat{C}_k = 0$ and $\gamma_k = 1$ centers the proximal solve at $\bar{x}^k$; recovering FedProx (Li et al., 2020) also requires its reset and aggregation rules. **FedAvg and SABER.** Local SGD responses and variance-reduced corrections (Mishchenko et al., 2024) can be considered through their complete implicit error.

3.2 A CONCRETE FES-MA ERROR BOUND

Proposition 3.1 (Specified FES-MA implementation)**.** *Let the local gradients be L-Lipschitz, $t = \beta L < 1$, and use $P = P_2$, $\mu_k = 1$, $B_k = G_2$, and fixed $\gamma_k = \gamma > 0$. Suppose $\mathbf{E}_k D_k = pI$ and $\alpha_k = p\gamma \in [a_0, 1]$. Define $z_k = \eta_k - \gamma\beta D_k(\widehat{C}_k - H_2(x^k))$. There are constants $R, A \geq 0$, depending only on t, p, γ, such that*

$$(\mathbf{E}_k\|\Theta_k\|^2)^{1/2} \leq R\|e_k\| + A(\mathbf{E}_k\|z_k\|^2)^{1/2}.$$

For fixed p, γ, $R = \gamma\sqrt{p(1-p)} + O(t)$ and $A = 1 + O(t)$ as $t \to 0$. A bound $(\mathbf{E}_k\|z_k\|^2)^{1/2} \leq u\|e_k\| + v_k$ with deterministic budgets and $\sum_k v_k^2 < \infty$ therefore yields (2.2)*, with ρ proportional to $(R + Au)^2$ and ε_k proportional to v_k. No independence between selection, correction, and solver errors is required.*

Feasibility and scope. For $p = 1$, exact H_2, $\gamma = a_0 = 1/2$, and $\beta L = 1/256$, local solves with $(\mathbf{E}_k\|\eta_k\|^2)^{1/2} \leq v_k$ give a budget satisfying (2.5). Square-summable v_k suffice for stationarity; geometric v_k also satisfy (2.11), with boundedness, KL, and conditional concavity still required for whole-sequence convergence. Smoothness controls the local gradient increment without a client gradient-variance bound, but correction sampling variance remains in z_k. The selection bound is not claimed tight: failure to certify moderate participation does not imply instability. Other proximal scales add error terms. The empirical configurations below explore performance outside this sufficient certificate.

4 DISAGREEMENT AND HETEROGENEOUS LOCAL RESPONSES

4.1 A COMMON-STATE COMPARISON

We compare FES-DR and FES-MA from the same persistent state, using the centers in Section 3.1, full participation, equal weights, exact gradients and local solves, and $\mu = 1$. Write $w = N^{-1}\sum_i x_i$, $q_i = x_i - w$, and $h = \nabla\Phi(w)$. Both methods use the same β, γ without resetting client states. With $d_i = \nabla f_i(x_i) - \nabla f_i(w)$ and $\bar{d} = N^{-1}\sum_i d_i$, their centers differ by $v_i^{(2)} - v_i^{(1)} = -\beta(d_i - \gamma\bar{d})$, where superscripts $1, 2$ denote FES-DR and FES-MA. The following calculation identifies the effect on their output means.

4.2 AN EXACT MEAN-UPDATE DECOMPOSITION

Consider $f_i(u) = \frac{1}{2}u^\top A_i u + b_i^\top u + c_i$, with symmetric A_i and $I + \beta A_i \succ 0$. Set

$$B_i = (I + \beta A_i)^{-1}, \qquad \bar{B} = \frac{1}{N}\sum_i B_i, \qquad C_q = \frac{1}{N}\sum_i B_i q_i, \qquad \bar{d} = \frac{1}{N}\sum_i A_i q_i.$$

Proposition 4.1 (Mean-update decomposition)**.** *The output means satisfy*

$$w_2^+ - w = -\gamma\beta\bar{B}h + (1-\gamma)C_q,$$
$$w_1^+ - w = -\gamma\beta\bar{B}h - \gamma C_q - \gamma\beta\bar{B}\bar{d}.$$

Define each method's disagreement-induced mean perturbation by

$$\Delta_j(q) = w_j^+(w+q_1,\dots,w+q_N) - w_j^+(w,\dots,w).$$

Proposition 4.1 gives $\Delta_2 = (1-\gamma)C_q$ and $\Delta_1 = -\gamma C_q - \gamma\beta\bar{B}\bar{d}$. If all A_i are equal, both vanish despite arbitrary gradient offsets b_i. Otherwise, the mean-zero disagreement couples to differences in A_i and B_i; the ordering depends on their alignment and possible cancellation within Δ_1.

4.3 A TWO-CLIENT EXAMPLE WITH AN EXPLICIT ADVANTAGE REGIME

Example 4.2 (Curvature heterogeneity at a fixed global objective)**.** Let $N = 2$, $d = 1$, and

$$f_1(u) = \frac{a+\delta}{2}u^2 + bu, \qquad f_2(u) = \frac{a-\delta}{2}u^2 - bu,$$

where $a > 0$, $\delta > 0$, and $1 + \beta(a-\delta) > 0$. The global objective is $\Phi(u) = au^2/2$, independent of δ and b. Start from $x_1 = r$, $x_2 = -r$, with $r > 0$. Thus $w = 0$ is globally optimal, while the client state is not in consensus. With $D = (1+\beta a)^2 - \beta^2\delta^2 > 0$, the exact output means are

$$w_2^+ = -\frac{(1-\gamma)\beta\delta r}{D}, \qquad w_1^+ = -\frac{\gamma\beta^2 a\delta r}{D}.$$

Consequently,

$$\frac{|w_2^+|}{|w_1^+|} = \frac{|1-\gamma|}{\gamma\beta a}, \qquad \Phi(w_j^+) - \Phi(0) = \frac{a}{2}|w_j^+|^2.$$

FES-MA has a strictly smaller mean perturbation and increase in the global objective evaluated at the client-state mean whenever $|1-\gamma| < \gamma\beta a$. This is a nonempty neighborhood of $\gamma = 1$ and includes values with retained client history, $\gamma \neq 1$.

Controlled one-step validation. We implement Example 4.2 directly in NumPy: each method constructs its own proximal centers from the same client state and solves $(I + \beta A_i)x_i^+ = v_i - \beta b_i$. The closed-form means above are evaluated independently for comparison. Unless varied, we use $a = 2$, $\delta = 1$, $\beta = 0.2$, $r = 1$, $b = 1$, and $\gamma = 0.9$, retaining client history. We measure $S_j = |\Delta_j(q)|$, which equals $|w_j^+|$ here because the consensus-input response is zero. At the default setting, direct solves give $w_1^+ = -0.0375$ and $w_2^+ \approx -0.0104167$: FES-MA reduces the mean perturbation to 27.78% of FES-DR's and the increase in $\Phi(w_j^+)$ to $1/12.96$ of its value.

Figure 1(a,c) confirms the predicted growth with curvature heterogeneity and client disagreement. Panel (b) also includes regimes favoring FES-DR, and all $9{,}900$ parameter combinations in panel (d) agree with the theoretical comparison away from equality. The scalar scans agree with the closed-form means to within 3×10^{-16}. This isolates the disagreement–response interaction; the FEMNIST and CIFAR-100 experiments in Section 5 assess the full algorithms under sampled participation and inexact local computation.

Remark 4.3 (Interpretation of increasing heterogeneity). In this example the two nonzero perturbation magnitudes increase with δ over the admissible range, but their ratio does not depend on δ. The example therefore supports an increasing absolute separation in the favorable regime, not an increasing relative advantage. Outside that regime FES-DR can be better. It compares $\Phi(w_j^+)$, not the full-space FBE, and the illustrative parameters are not a certificate of (2.5).

Beyond quadratics. At $\gamma = 1$, the FES-MA center is $v_i^{(2)} = w + \beta\nabla f_i(w) - \beta\nabla\Phi(w)$, independent of the individual initial state, recovering the DANE case identified in Section 3.1. Thus, whenever the local response is unique, the output depends only on w and $\Delta_2(q) = 0$. For L-smooth nonconvex losses, $\beta L < 1$ ensures uniqueness. At $\gamma \neq 1$ client history remains, and the quadratic example identifies a regime of reduced coupling without asserting universal dominance for nonlinear losses.

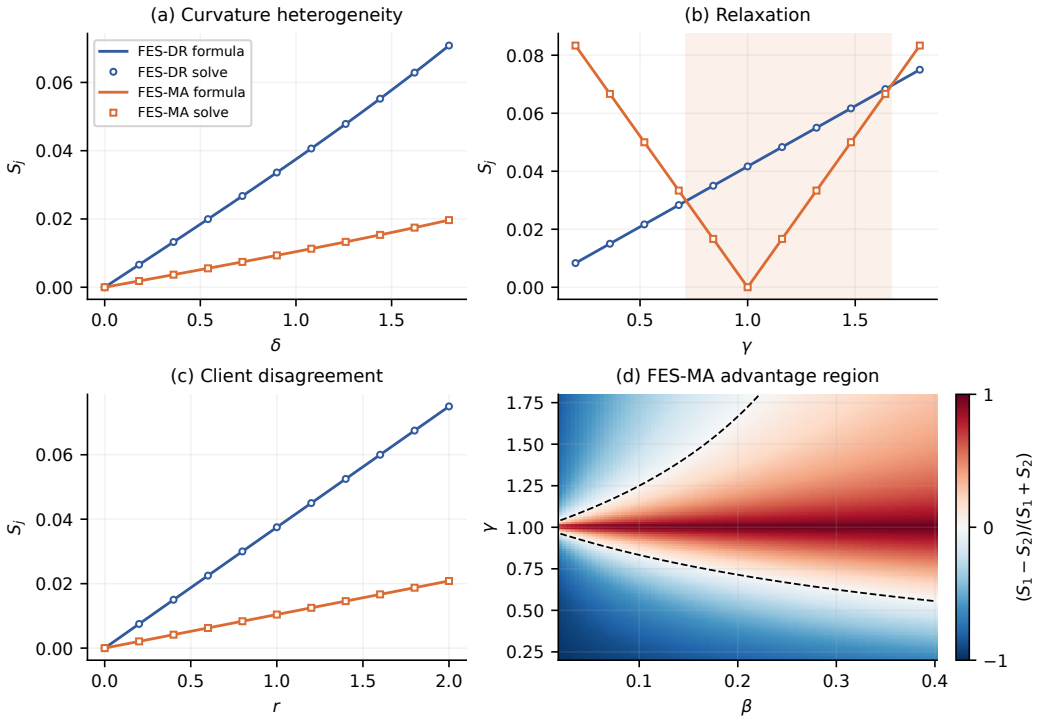


Figure 1: Controlled quadratic parameter scans for Example 4.2. Lines in (a)–(c) are closed-form predictions; markers are direct proximal solves. (a) Curvature heterogeneity $\delta \in [0, 1.8]$ varies at a fixed global objective. (b) Relaxation $\gamma \in [0.2, 1.8]$ reveals the favorable interval (shaded). (c) Disagreement amplitude $r \in [0, 2]$ gives a linear response. (d) Direct solves on a 90×110 grid in (β, γ); positive contrast means a smaller FES-MA perturbation. Dashed curves mark $|1-\gamma| = \gamma\beta a$. All panels use full participation, exact gradients, equal weights, and $\mu = 1$.

5 Numerical Experiments

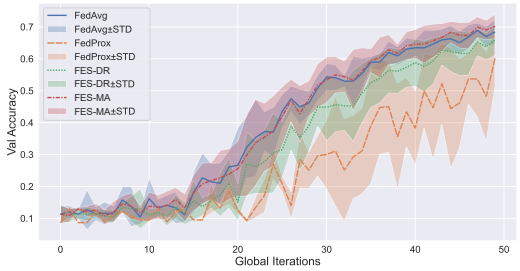
5.1 Experimental Setup

We compare the FES-DR and FES-MA variants with FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020) on FEMNIST and CIFAR-100 using `fl-sim` on two NVIDIA RTX 3090 GPUs. FES-DR and FES-MA use $(B_k, \widehat{C}_k) = (G_j^{\mathcal{S}_k}, H_j^{\mathcal{S}_k})$, $j = 1, 2$. All methods use learning rate 0.05, client sampling ratio 0.5, and three local epochs per round, with dataset-specific logarithmic grid searches for algorithm parameters. Local proximal scales are μ_{prox} for FedProx and $\beta\mu$ for FES-DR and FES-MA.

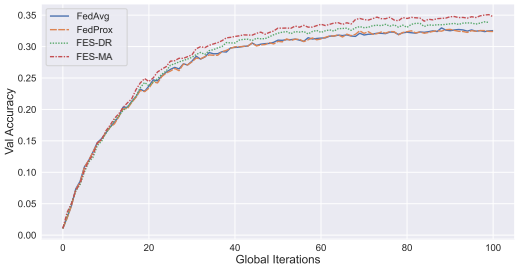
FES-DR and FES-MA both use $\gamma = 1.1$, where γ is the relaxation parameter in FedDR. Both training implementations average participating clients' returned models, agreeing with the template's full-state mean under full participation. Their (β, μ) pairs, ordered as FES-DR/FES-MA, are $(0.04, 250)/(0.04, 125)$ on FEMNIST, $(0.005, 4000)/(0.05, 2000)$ on CIFAR-100 with $\alpha_{\text{LDA}} = 0.1$, and $(0.005, 4000)/(0.005, 20000)$ with $\alpha_{\text{LDA}} = 1.0$. FedProx uses $\mu_{\text{prox}} = 2$, 100, and 200, respectively.

5.2 Datasets and Training Results

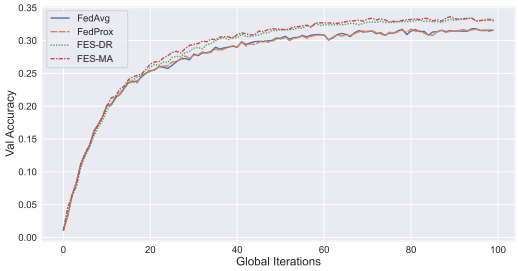
FEMNIST. We use a 10-class subset (a–j) extracted from the original 62-class FEMNIST dataset, resampled and repartitioned to increase heterogeneity, with min–max normalization to $[0, 1]$. The model is `cnn_femnist_tiny`. We run each algorithm five times, with 50 communication rounds per run. At the last plotted iteration, FES-MA reaches approximately 70.2% mean validation accuracy, exceeding FedAvg (68.4%), FedProx (60.3%), and FES-DR (66.0%) by about 1.7, 9.9, and 4.2



(a) FEMNIST.



(b) CIFAR-100, $\alpha_{\mathrm{LDA}} = 0.1$.



(c) CIFAR-100, $\alpha_{\mathrm{LDA}} = 1.0$.

Figure 2: Validation accuracy on FEMNIST (top) and CIFAR-100 under high and moderate heterogeneity (bottom). FES-DR and FES-MA both use relaxation $\gamma = 1.1$. The controlled quadratic study separately isolates the disagreement mechanism.

percentage points, respectively (Figure 2a). Accuracy values and gains here and below are estimated from the plotted curves at their last common iteration; FEMNIST curves average the five runs.

CIFAR-100. We use Dirichlet-based label-skew partitioning (Hsu et al., 2019) with $\alpha_{\mathrm{LDA}} \in \{0.1, 1.0\}$, the `cnn_cifar` model, and 100 rounds. Under high heterogeneity ($\alpha_{\mathrm{LDA}} = 0.1$), at plotted iteration 99, FES-MA reaches approximately 35.1% accuracy, compared with 32.5% for FedAvg, 32.4% for FedProx, and 33.8% for FES-DR: gains of about 2.6, 2.7, and 1.2 percentage points, respectively (Figure 2b). At $\alpha_{\mathrm{LDA}} = 1.0$, FES-MA reaches about 33.0%, exceeding FedAvg and FedProx (both 31.6%) by 1.4 percentage points, while FES-DR reaches 33.1% (Figure 2c). The larger advantage of FES-MA over FedAvg and FedProx under stronger heterogeneity supports its practical benefit; the quadratic study separately tests the proposed mechanism.

6 Conclusion

We developed a product-space framework for stochastic, inexact federated optimization, with convex convergence under conditional error and mean-bias budgets, and nonconvex stationarity through expected FBE descent. Under local KL geometry, almost-sure boundedness, and the stated process and error-tail conditions (Section 2.3), we further obtain finite length and whole-sequence convergence. A concrete error bound for the FES-MA variant connects the abstract conditions to persistent client updates. The quadratic analysis and controlled scans identify a regime where FES-MA suppresses disagreement-induced mean perturbations; FEMNIST and CIFAR-100 results provide complementary evidence on complete training. Sharper sampling bounds and nonlinear mechanism tests remain directions for extending this connection.

AI use statement

Starting from the authors' existing manuscript, we used generative AI tools to assist with checking and revising mathematical claims and proofs, translating and editing text. AI assistance also supported the code work of the controlled quadratic validation study and the preparation of its scripts and visualizations. The federated learning training implementation and the FEMNIST and CIFAR-

100 training results were developed independently of this assistance. Additionally, AI tools helped identify relevant literature, check bibliographic information, and format references. Checks of the AI-assisted revisions included examining mathematical assumptions and proof steps, comparing bibliographic details with source publications, and compiling and inspecting the manuscript. The authors are responsible for reviewing all AI-assisted material and for the final content of this work, including its text, mathematical claims, code, and experimental results.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Anonymous authors. Inexact alternating direction method of multipliers with efficient local termination criterion for cross-silo federated learning, 2024. URL https://openreview.net/forum?id=MZU09jacLd. TMLR submission; cited anonymous manuscript.

Anonymous authors. Mitigating data heterogeneity effect in client-reshuffling-based federated learning, 2026. URL https://openreview.net/pdf/f568890f84bde1a4f5e841ef01179eb7ff0998e4.pdf. ICLR 2026 submission; cited anonymous manuscript.

Wojciech Anyszka, Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Tighter performance theory of FedExProx, 2026. URL https://arxiv.org/abs/2410.15368v2. arXiv:2410.15368, version 2.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1–2):91–129, 2013. doi: 10.1007/s10107-011-0484-9.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, Switzerland, second edition, 2017. doi: 10.1007/978-3-319-48311-5.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. doi: 10.1137/060670080.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014. doi: 10.1007/s10107-013-0701-9.

Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015. doi: 10.1137/140971233.

Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1–3):293–318, 1992. doi: 10.1007/BF01581204.

Jonathan Eckstein and Wang Yao. Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. *Mathematical Programming*, 170(2):417–444, 2018. doi: 10.1007/s10107-017-1160-5.

Jean-Baptiste Fest, Audrey Repetti, and Emilie Chouzenoux. A stochastic use of the Kurdyka–Łojasiewicz property: Investigation of optimization algorithms behaviours in a non-convex differentiable framework. *Foundations of Data Science*, 9:164–191, 2026. doi: 10.3934/fods.2025016.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *NeurIPS Workshop on Federated Learning*, 2019. URL https://research.google/pubs/measuring-the-effects-of-non-identical-data-distribution-for-federated-visual-classification/.

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

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Hanmin Li and Peter Richtárik. On the convergence of FedProx with extrapolation and inexact prox. In *NeurIPS Workshop on Optimization for Machine Learning (OPT)*, 2024. URL https://opt-ml.org/papers/2024/paper76.pdf.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. FedDANE: A federated Newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019. URL https://arxiv.org/abs/2001.01920.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka–Łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023. doi: 10.1137/21M1468048.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. doi: 10.1137/0716071.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Konstantin Mishchenko, Rui Li, Hongxiang Fan, and Stylianos Venieris. Federated learning under second-order data heterogeneity. Preprint, OpenReview, 2024. URL https://openreview.net/forum?id=71YWF6ZuPD.

Reese Pathak and Martin J. Wainwright. FedSplit: An algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7057–7066, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4ebd440d99504722d80de606ea8507da-Abstract.html.

Thomas Pethick, Puya Latafat, Basile Lewandowski, Zheng Xu, Peter Kairouz, Panagiotis Patrinos, and Volkan Cevher. iFedDR: Auto-tuning local computation with inexact Douglas–Rachford splitting in federated learning, 2025. URL https://openreview.net/pdf?id=XFpb3T5Zc9. Preprint.

Duy Nhat Phan, Patrick Hytla, Andrew Rice, and Thuy Ngoc Nguyen. Federated learning with randomized alternating direction method of multipliers and application in training neural networks. *Neural Networks*, 189:107501, 2025. doi: 10.1016/j.neunet.2025.107501.

Guy Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28(1):96–115, 1984. doi: 10.1007/BF02612715.

Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczós, and Alex Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016. URL https://arxiv.org/abs/1608.06879.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, London, 1971. doi: 10.1016/B978-0-12-604550-5.50015-8.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 1998. doi: 10.1007/978-3-642-02431-3.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1000–1008, 2014. URL https://proceedings.mlr.press/v32/shamir14.html.

Yongcun Song, Ziqi Wang, and Enrique Zuazua. FedADMM-InSa: An inexact and self-adaptive ADMM for federated learning, 2024. URL https://arxiv.org/abs/2402.13989v3. arXiv:2402.13989, version 3.

Benar Fux Svaiter. A weakly convergent fully inexact Douglas–Rachford method with relative error tolerance. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:57, 2019. doi: 10.1051/cocv/2018063.

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018. doi: 10.1137/16M1080240.

Quoc Tran-Dinh, Nhan H. Pham, Dzung T. Phan, and Lam M. Nguyen. FedDR: Randomized Douglas–Rachford splitting algorithms for nonconvex federated composite optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30326–30338, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/fe7ee8fc1959cc7214fa21c4840dff0a-Abstract.html.

Jiao Xue and Chundong Wang. EF-Feddr: Communication-efficient federated learning with Douglas–Rachford splitting and error feedback. *Frontiers in Artificial Intelligence*, 9:1699896, 2026. doi: 10.3389/frai.2026.1699896.

Bicheng Ying, Zhe Li, and Haibo Yang. Exact and linear convergence for federated learning under arbitrary client participation is attainable. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/395409679270591fd2a70abc694cf5a1-Abstract-Conference.html.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher G. Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9521b6e7f33e039e7d92e23f5e37bbf4-Abstract-Conference.html.

Shenglong Zhou and Geoffrey Ye Li. Federated learning via inexact ADMM, 2023. URL https://arxiv.org/abs/2204.10607v4. arXiv:2204.10607, version 4.

A Detailed comparison of convergence guarantees

Tables 2–3 expand Table 1, separating participation, error, and geometric conditions whenever they change the conclusion. Entries concern the specified analyses, rather than every variant of an algorithm.

Table 2: Reported guarantees and marked direct consequences under specified participation and error models (Part I). Convexity refers to the original objective. Point convergence, stationarity measures, and objective-gap rates are distinct.

Method	Client participation	Computation / error model	Convexity	Convergence guarantee	Source theorem
DANE (Shamir et al., 2014)	Full	Exact local solves	Strongly convex	Linear objective-gap decay	Thm. 4
INEXACTDANE (Reddi et al., 2016)	Full	Relative local-solve error	Nonconvex[6]	Stationarity bound	Thm. 4
FedDANE (Li et al., 2019)	Random sampling	Sampled correction; relative distance error	Nonconvex[5]	Expected stationarity bound	Thms. 5–6
FedSplit (Pathak & Wainwright, 2020)	Full	Fixed absolute proximal error	Convex[1]	Error neighborhood	Thm. 1
FedSplit	Full	Vanishing absolute proximal error	Convex[1]	Whole-sequence convergence†	Thm. 3
iFedDR (Pethick et al., 2025)	Full	Relative resolvent residual + correction	Convex[2]	Residual bound; whole-sequence convergence	Thm. 3.1
FedADMM-InSa (Song et al., 2024)	Full	Relative local optimality residual	Convex[1]	Whole-sequence convergence	Thm. 3.3
Inexact FedExProx (Li & Richtárik, 2024)	Full	Fixed absolute proximal error	Convex[3]	Error neighborhood	Thm. 1
Inexact FedExProx	Full	Relative proximal error	Convex[3]	Linear rate; finite length and whole sequence†	Thms. 2–3
Inexact FedExProx	Uniform w/o replacement	Relative proximal error	Convex[3]	Expected linear rate; a.s. finite length and whole sequence†	Thm. 11
FedExProx (revised analysis) (Anyszka et al., 2026)	Uniform w/o replacement	Exact proximal evaluations	Convex[4]	Expected linear gap decay; a.s. finite length and whole sequence†	Thm. D.1
FedExProx (revised analysis)	Full	Fixed absolute proximal error	Convex[4]	Error neighborhood	Thm. E.1
FedProx (Li et al., 2020)	Random sampling	Relative local gradient residual	Nonconvex[5]	Stationarity complexity	Thms. 4, 6
FedDR (Tran-Dinh et al., 2021)	iid proper sampling	Square-summable absolute proximal error	Nonconvex	Stationarity bound; stationary cluster points†	Thm. 3.1
FedDR	iid proper sampling	Relative proximal error	Nonconvex	Stationarity bound; stationary cluster points†	Thm. A.1

Conditions. All rows retain the cited initialization, smoothness, and parameter conditions. [1]Each client objective is strongly convex; FedADMM-InSa refers to its fixed-penalty FedADMM-In theorem. [2]Convex optimization specialization of a maximally monotone framework. [3]Convex client objectives, a strongly convex global objective, and interpolation; sampling changes the relative-error threshold. [4]Convex client objectives, interpolation, and average Moreau-envelope PL. The sampled row uses Thm. D.1 with $0 < \alpha\gamma < 2/L_{\gamma,S}$, rather than the differing stepsize printed in Thm. 7.5. [5]Dissimilarity and joint restrictions on participation, accuracy, and parameters. [6]The nonconvex INEXACTDANE theorem selects one local output instead of averaging; its strongly convex results also give linear convergence (Thm. 3). Relative errors use different reference quantities: the ideal proximal displacement in Inexact FedExProx and FedDANE, the actual local displacement in FedDR, and a local subproblem gradient in FedProx.

Reading both tables. "a.s." means almost surely; "w/o" means without. FIAELT and FedCDR refer to the cited submission versions; the following † convention applies to both tables. †Direct consequences of reported bounds: vanishing-error contraction gives the FedSplit limit; a uniform finite sum of expected squared stationarity measures gives a.s. vanishing measures. For Inexact FedExProx, linear distance decay to the unique solution gives finite length. For sampled FedExProx, the update norm is bounded by a constant times distance to the common solution set; for FOCUS, its squared expectation is bounded by a constant times the reported Lyapunov function. Geometric decay thus gives $\sum_k \mathbf{E}\|x^{k+1} - x^k\| < \infty$, hence a.s. finite length and point convergence. FOCUS implications are conditional on its reported bound (qualification D below). Cluster points are conditional on existence; rates use each paper's measure and iteration unit.

Table 3: Reported guarantees and marked direct consequences (Part II). Additional conditions and proof qualifications apply; our entries require verified joint error budgets.

Method	Client participation	Computation / error model	Convexity	Convergence guarantee	Source theorem
FedADMM (Zhou & Li, 2023)	Prescribed window coverage	Geometrically decaying local gradient residual	Nonconvex[A]	Vanishing gradients; stationary cluster points; stationarity bound	Thms. 4.1–4.3
FedADMM	Same	Same	Nonconvex[A]	Whole-sequence convergence	Thm. 4.2(b)
FIAELT (Anonymous authors, 2024)	Full	Conditional mean-square relative proximal error	Nonconvex[B]	Stationarity bound; a.s. vanishing stationarity measure[†]	Thm. 1
FIAELT	Full	Pathwise relative proximal error	Nonconvex[B]	Stationary cluster points	Prop. 6; Cor. 3
FIAELT	Full	Pathwise relative proximal error	Nonconvex[B]	Finite length; whole-sequence convergence; KL-dependent rates	Thm. 2
FedADMM-MM (Phan et al., 2025)	History-independent, positive probabilities	Summable local objective-gap error; server MM	Nonconvex[C]	A.s. vanishing consensus residual; stationary cluster points	Thm. 1
FedADMM-MM	Same	Same; surrogate subgradient condition	Nonconvex[C]	Stationarity bound; a.s. vanishing stationarity measure[†]	Thm. 2
FedCDR (Anonymous authors, 2026)	Client reshuffling	Exact proximal evaluations	Nonconvex	Stationarity bound; vanishing gradients; stationary cluster points[†]	Thm. 2
FedCDR	Client reshuffling	Square-summable absolute proximal error	Nonconvex	Same	Thm. 1
FOCUS (Ying et al., 2025)	Positive, nonuniform probabilities	Prescribed exact-gradient updates	Nonconvex[D]	Stationarity bound; vanishing gradients; stationary cluster points[†]	Thm. 1
FOCUS	Same	Same	Nonconvex[D]	Expected linear gap decay; a.s. finite length and whole sequence[†]	Thm. 1
SG-FOCUS	Same	Unbiased, bounded-variance gradients; fixed step size	Nonconvex[D]	Stationarity bound with a noise floor	Thm. 6
SG-FOCUS	Same	Same	Nonconvex[D]	Linear objective-gap decay to a noise neighborhood	Thm. 5
EF-Feddr (Xue & Wang, 2026)	iid proper sampling	Absolute proximal error + compression error feedback	Nonconvex[E]	Stationarity bound with a compression residual	Thm. 1; Eq. (28)
Ours	Budget-admissible sampling	Conditional mean-square full-update error	Convex[F]	**A.s. whole-sequence convergence**	Thm. 2.2
Ours	Same	Same; square-summable tolerances	Nonconvex[F]	Vanishing residual; stationary cluster points; stationarity bound	Thm. 2.1
Ours	Same	Same; stronger tolerance decay	Nonconvex[F]	**A.s. finite length; whole-sequence convergence**	Thm. 2.3

[A]Coercivity for cluster-point/sequence results; an isolated stationary cluster point additionally for whole-sequence convergence. Complexity does not require coercivity. [B]Reported results with convex regularization; pathwise rows require bounded trajectories (App. C.3.1), and the last also requires KL. The cited version has descent-parameter inconsistencies (Prop. 5; Eqs. (39)–(40)). [C]Exact server surrogate minimization and Assumptions 1–3; complexity additionally requires Assumption 4. Function-value error summability is distinct from distance-error summability. [D]Objective-gap rows require PL; conclusions concern the server trajectory. PL/stochastic bounds are reported results with printed descent/noise-coefficient inconsistencies (App. E.6; Thm. 6). General Markov-dependent participation is not covered. [E]The printed M_4 scales with K, so $M_4\nu^2/K$ in Eq. (28) need not vanish at fixed compression precision. [F]Convex convergence additionally requires summable conditional-mean bias. Nonconvex whole-sequence convergence requires almost-sure boundedness, local KL geometry with a common exponent, conditional concavity, and stronger tolerance-tail decay (Section 2.3). Sampling and solver choices must satisfy the joint budget; arbitrary participation rates are not certified.

PROOF GUIDE

This appendix provides the technical statements and proofs supporting the main results. Conditions are restated where needed.

Main-paper result or omitted material	Proof location
DRS equivalence, residual geometry, FBE slope and descent	Appendix B
Theorems 2.1, 2.2	Appendix C
Conditional summability; Theorem 2.3	Appendix D
Sampling and implicit-error propositions; Proposition 3.1	Appendix E
Proposition 4.1, Example 4.2, cancellation	Appendix F

B NOTATION AND COMMON ESTIMATES

The projection, resolvent, and monotonicity facts used below are standard in convex and monotone-operator analysis (Bauschke & Combettes, 2017). We use the Fréchet and limiting subdifferentials in the sense of Rockafellar & Wets (1998). The FBE construction follows the envelope literature (Themelis et al., 2018); the residual, slope, and conditional-descent estimates below are proved for the present update and its $C^{1,1}$ target. In particular, the slope estimate is not obtained by assuming a twice continuously differentiable FBE.

References to Theorems 2.1–2.3, Proposition 3.1, Proposition 4.1, and Example 4.2 refer to the main paper.

B.1 STANDING NOTATION AND ASSUMPTIONS

Let $\mathcal{H} = \mathbb{R}^{Nd}$ with its Euclidean inner product, $\Omega = \{\mathbf{1}_N \otimes w : w \in \mathbb{R}^d\}$, Π its orthogonal projector, and $Q = I - \Pi$. We consider $\min_x P(x) + \delta_\Omega(x)$, with $P \in C^{1,1}$, an L-Lipschitz gradient $g = \nabla P$, and $P_{\inf} := \inf_\Omega P > -\infty$. No boundedness of the trajectory is included in these standing assumptions. For fixed $\beta > 0$ define

$$M_\beta(x) = x + \beta g(x), \quad T_\beta(x) = \Pi(x - \beta g(x)), \quad e_\beta(x) = x - T_\beta(x),$$
$$\mathcal{E}_\beta(x) = P(x) - \langle g(x), e_\beta(x)\rangle + \frac{\|e_\beta(x)\|^2}{2\beta}.$$

Write $s_k = x^{k+1} - x^k$, $e_k = e_\beta(x^k)$, and $\mathbf{E}_k = \mathbf{E}[\cdot \mid \mathcal{F}_k]$. The state x^k and step $\alpha_k \in [a_0, 1]$ are $\mathcal{F}_k$-measurable, where $a_0 > 0$. The actual implicit error is defined by

$$M_\beta(x^{k+1}) = M_\beta(x^k) - \alpha_k e_k + \Theta_k. \tag{B.1}$$

The main-paper criterion (2.2) means

$$\mathbf{E}_k\|\Theta_k\|^2 \le \rho\|e_k\|^2 + \varepsilon_k^2, \qquad \rho \ge 0, \quad \varepsilon_k \ge 0, \quad \sum_k \varepsilon_k^2 < \infty,$$

with deterministic budgets. It neither assumes conditional centering nor bounds realized errors pathwise. In the nonconvex branch we use the sufficient parameter condition (2.5), explicitly

$$t = \beta L \le 1/16, \quad q = (1-t)^{-1}, \quad c_1 = 1 + 2tq + (1+t)q^2,$$
$$c_2 = (1+t)q^2/2, \quad c_1\sqrt{\rho} + c_2\rho \le a_0/8.$$

The convex branch instead allows any $\beta > 0$ and uses its separate mean budget and parameter inequality.

B.2 EQUIVALENCE WITH DOUGLAS–RACHFORD SPLITTING

For convex P, the standard proximal–resolvent identity gives $J_\beta = (I + \beta\nabla P)^{-1} = \mathrm{prox}_{\beta P}$, while $\mathrm{prox}_{\beta\delta_\Omega} = \Pi$ (Bauschke & Combettes, 2017, Chapter 26). For an L-smooth nonconvex P and $\beta L < 1$, the proximal objective $P(u) + \|u - z\|^2/(2\beta)$ is strongly convex with modulus

$\beta^{-1} - L > 0$. It has a unique minimizer, characterized by $z = u + \beta g(u)$, so the same inverse identity holds. With $z = M_\beta(x)$, this gives

$$J_\beta z = x, \qquad 2J_\beta z - z = x - \beta g(x), \qquad \Pi(2J_\beta z - z) - J_\beta z = -e_\beta(x).$$

Substitution into (B.1) yields (2.3). This algebraic equivalence also holds with $\Theta_k \neq 0$; nonconvex convergence is established by the envelope analysis below, rather than by a convex monotone-operator convergence theorem.

B.3 Residual geometry and step control

Lemma B.1 (Residual geometry and step control)**.** *The residual admits the orthogonal decomposition*

$$\begin{aligned} e_\beta(x) &= Qx + \beta \Pi g(x), \\ \|e_\beta(x)\|^2 &= \|Qx\|^2 + \beta^2 \|\Pi g(x)\|^2. \end{aligned} \tag{B.2}$$

Consequently, $e_\beta(x) = 0$ is equivalent to $x \in \Omega$ and $\Pi g(x) = 0$, or equivalently $0 \in g(x) + N_\Omega(x)$. If $t := \beta L < 1$, then M_β is bijective and its inverse is $q := (1-t)^{-1}$-Lipschitz. Moreover,

$$\begin{aligned} \|s_k\| &\le q(\alpha_k \|e_k\| + \|\Theta_k\|), \\ (\mathbf{E}_k \|s_k\|^2)^{1/2} &\le q\big((1 + \sqrt{\rho})\|e_k\| + \varepsilon_k\big). \end{aligned} \tag{B.3}$$

The same right-hand side bounds $\mathbf{E}_k \|s_k\|$. If P is convex, bijectivity and these estimates hold for every $\beta > 0$, with $q = 1$.

Proof. The residual decomposition follows from $I = \Pi + Q$. Since $Qx \in \Omega^\perp$ and $\Pi g(x) \in \Omega$ are orthogonal, their squared norms add. The criticality equivalence follows from $N_\Omega(x) = \Omega^\perp$ for $x \in \Omega$.

$$\langle M_\beta(x) - M_\beta(y), x - y \rangle \ge (1 - \beta L)\|x - y\|^2.$$

Thus the inverse is q-Lipschitz on the image. For every v, the map $u \mapsto v - \beta g(u)$ is a contraction and has a unique fixed point, proving surjectivity of M_β. The first step estimate follows from (B.1). The conditional L^2 triangle inequality and (2.2) give the second estimate; conditional Cauchy–Schwarz gives the first-moment version.

If P is convex, monotonicity of its gradient improves the strong monotonicity constant to 1. The strongly convex coercive function $u \mapsto \frac{1}{2}\|u - v\|^2 + \beta P(u)$ has a unique minimizer satisfying $M_\beta(u) = v$, without a restriction on βL. This proves the claim. □

B.4 Envelope geometry and expected descent

Proposition B.2 (FBE descent and criticality)**.** *If $\beta L < 1$, then the FBE is locally Lipschitz and satisfies*

$$\begin{aligned} \mathcal{E}_\beta(x) &\ge P(T_\beta(x)) + \frac{1 - \beta L}{2\beta}\|e_\beta(x)\|^2 \\ &\ge P_{\inf}. \end{aligned} \tag{B.4}$$

$$\begin{aligned} (\beta^{-1} - L)\|e_\beta(x)\| &\le \operatorname{dist}(0, \partial \mathcal{E}_\beta(x)), \\ \operatorname{dist}(0, \partial \mathcal{E}_\beta(x)) &\le (\beta^{-1} + L)\|e_\beta(x)\|. \end{aligned} \tag{B.5}$$

Here ∂ denotes the limiting subdifferential. Thus $0 \in \partial \mathcal{E}_\beta(x)$ is equivalent to criticality for the original constrained problem. Suppose further that condition (2.5) *holds.*

Set $\kappa = a_0/(4\beta)$, $a = \kappa/4$, and $c_\varepsilon = c_2/\beta + c_1^2/(\kappa\beta^2)$. Then

$$\mathbf{E}_k \mathcal{E}_\beta(x^{k+1}) \le \mathcal{E}_\beta(x^k) - a\|e_k\|^2 + c_\varepsilon \varepsilon_k^2. \tag{B.6}$$

The parameter range is sufficient and conservative. Estimates (B.3)*,* (B.5)*, and* (B.6) *are verified at the abstract level and need not be re-established separately for each client implementation.*

Proof. **Representation and lower bound.** Projection optimality and (B.2) yield

$$\begin{aligned} \mathcal{E}_\beta(x) &= \min_{u\in\Omega} \left\{ P(x) + \langle g(x), u - x\rangle + \frac{1}{2\beta}\|u - x\|^2 \right\} \\ &= P(x) - \langle g(x), Qx\rangle + \frac{1}{2\beta}\|Qx\|^2 - \frac{\beta}{2}\|\Pi g(x)\|^2. \end{aligned} \tag{B.7}$$

The minimizer is $T_\beta(x)$. Smoothness gives $P(T_\beta(x)) \le P(x) - \langle g(x), e_\beta(x)\rangle + (L/2)\|e_\beta(x)\|^2$, proving (B.4). Formula (B.7) also shows that the FBE is locally Lipschitz.

Limiting subdifferential estimate. Set $h = y - x$ and $\Delta g = g(x + h) - g(x)$. Direct expansion gives

$$\mathcal{E}_\beta(x + h) - \mathcal{E}_\beta(x) = \beta^{-1}\langle e_\beta(x), h\rangle - \langle e_\beta(x), \Delta g\rangle + R(x, h), \tag{B.8}$$

$$\begin{aligned} R(x, h) &= P(x + h) - P(x) - \langle g(x), h\rangle - \langle \Delta g, Qh\rangle \\ &\quad + \frac{1}{2\beta}\|Qh\|^2 - \frac{\beta}{2}\|\Pi\Delta g\|^2, \\ |R(x, h)| &\le \left(\frac{3L}{2} + \frac{1}{2\beta} + \frac{\beta L^2}{2}\right)\|h\|^2. \end{aligned}$$

For any Fréchet subgradient $v \in \widehat{\partial}\mathcal{E}_\beta(x)$, the supporting inequality and (B.8), evaluated along arbitrary directions, imply $\|v\| \le (\beta^{-1} + L)\|e_\beta(x)\|$. Taking $h = -se_\beta(x)$ with $s \downarrow 0$ in the same inequality gives

$$\langle v, e_\beta(x)\rangle \ge (\beta^{-1} - L)\|e_\beta(x)\|^2.$$

Both estimates pass to limiting subgradients. Local Lipschitz continuity, Rademacher's theorem, and a bounded subsequence of gradients at nearby differentiability points ensure that the limiting subdifferential is nonempty. Taking the infimum of subgradient norms proves (B.5). No Hessian or everywhere differentiability of the FBE is used.

Conditional expected descent. Temporarily write $x = x^k$, $y = x^{k+1}$, $s = y - x$, $e = e_k$, $d = g(y) - g(x)$, $\alpha = \alpha_k$, and $\Theta = \Theta_k$. In the minimization representation of $\mathcal{E}_\beta(y)$, use $T_\beta(x) = x - e$ as a feasible comparison point to obtain

$$\begin{aligned} \mathcal{E}_\beta(y) - \mathcal{E}_\beta(x) &\le P(y) - P(x) - \langle g(y), s\rangle - \langle d, e\rangle \\ &\quad + \beta^{-1}\langle e, s\rangle + (2\beta)^{-1}\|s\|^2 \\ &\le -\frac{\alpha}{\beta}\|e\|^2 + 2L\|e\|\|s\| + \beta^{-1}\|e\|\|\Theta\| + \frac{1 + t}{2\beta}\|s\|^2. \end{aligned}$$

The last inequality uses $s + \beta d = -\alpha e + \Theta$ and smoothness. Substituting the pathwise estimate in (B.3), the coefficient of $\|e\|^2$ becomes

$$-\frac{\alpha}{\beta}\left(1 - 2tq - \frac{1 + t}{2}\alpha q^2\right) \le -\frac{\alpha}{4\beta} \le -\kappa,$$

because the bracket is at least $59/225 > 1/4$ under (2.5). The remaining coefficients are bounded by c_1/β and c_2/β, giving

$$\mathcal{E}_\beta(y) - \mathcal{E}_\beta(x) \le -\kappa\|e\|^2 + \frac{c_1}{\beta}\|e\|\|\Theta\| + \frac{c_2}{\beta}\|\Theta\|^2. \tag{B.9}$$

By (2.2), $\mathbf{E}_k\|\Theta_k\| \le \sqrt{\rho}\|e_k\| + \varepsilon_k$. Taking conditional expectations, applying (2.5), and using

$$\frac{c_1}{\beta}\varepsilon_k\|e_k\| \le \frac{\kappa}{4}\|e_k\|^2 + \frac{c_1^2}{\kappa\beta^2}\varepsilon_k^2$$

proves (B.6). For this conditional argument, initial FBE integrability implies $\mathbf{E}\|e_0\|^2 < \infty$ by (B.4). Inductively, the error criterion makes Θ_k square integrable, (B.3) does the same for s_k, and (B.9) bounds the next energy above by an integrable variable while (B.4) bounds it below by $P_{\inf}$. Hence every conditional expectation used here is well defined. Conditional centering and bounded trajectories are not needed. This completes the proof. □

C Stationarity and convex convergence

The probabilistic tools are the almost-supermartingale principle (Robbins & Siegmund, 1971) and the stochastic quasi-Fejér approach (Combettes & Pesquet, 2015). We supply the update-specific estimates and the passage to point convergence explicitly, rather than invoking a block-sampling theorem whose hypotheses need not match Algorithm 1.

C.1 Proof of main Theorem 2.1

Statement (main Theorem 2.1). Under the basic conditions, Algorithm 1, (2.2), and (2.5), assume $\mathbf{E}|\mathcal{E}_\beta(x^0)| < \infty$. Then the FBE has a finite almost-sure limit, $\sum_k \mathbf{E}\|e_k\|^2 < \infty$, and $e_k \to 0$ almost surely. Every existing cluster point is critical for (2.1), and $\min_{0\le k\le n} \mathbf{E}\|e_k\|^2 = O((n+1)^{-1})$.

Use a and c_ε from Proposition B.2, and define the deterministic error tail and corrected nonnegative process

$$r_k = c_\varepsilon \sum_{j=k}^{\infty} \varepsilon_j^2, \qquad Z_k = \mathcal{E}_\beta(x^k) - P_{\inf} + r_k. \tag{C.1}$$

The following explicit rate implies the main-paper order bound:

$$\min_{0\le k\le n} \mathbf{E}\|e_k\|^2 \le \frac{\mathbf{E}Z_0}{a(n+1)}. \tag{C.2}$$

Proof. By (B.4) and (B.6), $Z_k \ge 0$ and $\mathbf{E}_k Z_{k+1} \le Z_k - a\|e_k\|^2$. Initial integrability and induction ensure that every Z_k is integrable. The nonnegative supermartingale convergence theorem (a special case of Robbins & Siegmund (1971)) yields a finite almost-sure limit for Z_k. Since $r_k \to 0$, the FBE also converges. Taking expectations and summing gives

$$a\sum_{k=0}^{n} \mathbf{E}\|e_k\|^2 + \mathbf{E}Z_{n+1} \le \mathbf{E}Z_0.$$

Nonnegativity, monotone convergence, and averaging establish square summability and (C.2). Continuity of the residual implies that every cluster point has zero residual and is therefore critical by Lemma B.1. Estimate (B.3) additionally yields $\sum_k \mathbf{E}\|s_k\|^2 < \infty$. This proves the theorem. □

C.2 Proof of main Theorem 2.2

Statement (main Theorem 2.2). Suppose the basic conditions and (2.2) hold, P is convex, and $\mathcal{X}^\star = \operatorname{argmin}_\Omega P \ne \varnothing$. Require the additional deterministic mean budget

$$\|\mathbf{E}_k\Theta_k\| \le \delta_k, \qquad \delta_k \ge 0, \qquad \sum_k \delta_k < \infty.$$

For any $\beta > 0$, assume $\rho < a_0(2 - a_0)$ and $\mathbf{E}\|M_\beta(x^0) - M_\beta(x^\star)\|^2 < \infty$ for some deterministic $x^\star \in \mathcal{X}^\star$. Then x^k converges almost surely to a point in $\mathcal{X}^\star$ and $\sum_k \|e_k\|^2 < \infty$ almost surely. With $y^k = T_\beta(x^k)$ and $P^\star = \min_\Omega P$,

$$\min_{0\le k\le n} \mathbf{E}\|e_k\|^2 = O((n+1)^{-1}), \qquad \min_{0\le k\le n} \mathbf{E}[P(y^k) - P^\star] = O((n+1)^{-1/2}).$$

For the fixed deterministic solution $x^\star$, write

$$\mathcal{D}_k(x^\star) = \|M_\beta(x^k) - M_\beta(x^\star)\|^2, \qquad A_0 = \mathbf{E}\mathcal{D}_0(x^\star), \qquad \lambda = a_0(2 - a_0) - \rho > 0. \tag{C.3}$$

Define finite constants

$$b_k = \delta_k + \varepsilon_k^2, \qquad B = \left(\prod_{k\ge 0}(1 + \delta_k)\right)\left(A_0 + \sum_{k\ge 0} b_k\right), \qquad C = A_0 + B\sum_{k\ge 0}\delta_k + \sum_{k\ge 0} b_k.$$

Besides almost-sure convergence, the proof gives

$$\min_{0\le k\le n} \mathbf{E}\|e_k\|^2 \le \frac{C}{\lambda(n+1)}, \qquad \min_{0\le k\le n} \mathbf{E}[P(y^k) - P^\star] \le \frac{2\sqrt{BC}}{\beta\sqrt{\lambda(n+1)}} + \frac{LC}{2\lambda(n+1)}.$$

Proof. **Convex geometry.** For $x^\star \in \mathcal{X}^\star$, one has $Qx^\star = 0$ and $\Pi g(x^\star) = 0$. Write $v = x - x^\star$ and $d = g(x) - g(x^\star)$. Orthogonal decomposition gives

$$\begin{aligned}\langle M_\beta(x) - M_\beta(x^\star), e_\beta(x)\rangle &= \langle v + \beta d, Qv + \beta\Pi d\rangle \\ &= \|e_\beta(x)\|^2 + \beta\langle v, d\rangle \ge \|e_\beta(x)\|^2.\end{aligned} \tag{C.4}$$

The last inequality uses gradient monotonicity. Consequently, $\|e_\beta(x)\| \le \|M_\beta(x) - M_\beta(x^\star)\|$. Set $z_k = M_\beta(x^k) - M_\beta(x^\star)$ and $w_k = z_k - \alpha_k e_k$. Then

$$\|w_k\|^2 \le \mathcal{D}_k(x^\star) - \alpha_k(2 - \alpha_k)\|e_k\|^2 \le \mathcal{D}_k(x^\star).$$

Conditional distance recursion. Since w_k is $\mathcal{F}_k$-measurable, (B.1), (2.2), (2.6), and $2\sqrt{u} \le u + 1$ yield

$$\begin{aligned}\mathbf{E}_k\mathcal{D}_{k+1}(x^\star) &= \|w_k\|^2 + 2\langle w_k, \mathbf{E}_k\Theta_k\rangle + \mathbf{E}_k\|\Theta_k\|^2 \\ &\le \mathcal{D}_k(x^\star) - \lambda\|e_k\|^2 + 2\delta_k\sqrt{\mathcal{D}_k(x^\star)} + \varepsilon_k^2 \\ &\le (1 + \delta_k)\mathcal{D}_k(x^\star) - \lambda\|e_k\|^2 + b_k.\end{aligned} \tag{C.5}$$

Only the cross term involving the conditional mean is controlled here. The realized random error is not replaced by its conditional mean.

Boundedness and square summability. Set $p_k = \prod_{j=0}^{k-1}(1 + \delta_j)$, with the empty product equal to 1. By (C.5),

$$\frac{\mathcal{D}_k(x^\star)}{p_k} + \sum_{j=k}^{\infty}\frac{b_j}{p_{j+1}}$$

is a nonnegative supermartingale with conditional decrement at least $\lambda\|e_k\|^2/p_{k+1}$. Since p_k converges to a finite positive value, $\mathcal{D}_k(x^\star)$ has a finite almost-sure limit and $\sum_k \|e_k\|^2 < \infty$. The convex inverse estimate in Lemma B.1 gives $\|x^k - x^\star\|^2 \le \mathcal{D}_k(x^\star)$. Hence the trajectory is bounded and all its cluster points belong to $\mathcal{X}^\star$.

Convergence of the full sequence. Choose a countable deterministic dense subset of $\mathcal{X}^\star$. Lipschitz continuity of M_β ensures initial distance integrability at every such point. Apply the preceding convergence argument to these points and take the countable intersection of the resulting probability-one events. Fix a trajectory in this intersection. Density and the triangle inequality imply that

its distance to $M_\beta(s)$ has a limit for every $s \in \mathcal{X}^\star$. Choose a cluster point $\widehat{x} \in \mathcal{X}^\star$. Along the corresponding subsequence, the distance to $M_\beta(\widehat{x})$ tends to zero, so the full distance sequence tends to zero. The inverse estimate yields $x^k \to \widehat{x}$. No random cluster point is substituted into a conditional expectation recursion.

Rates. Taking expectations in (C.5), then using induction and summation, gives

$$\sup_k \mathbf{E}\mathcal{D}_k(x^\star) \le B, \qquad \lambda \sum_{k=0}^{n} \mathbf{E}\|e_k\|^2 \le C.$$

Averaging gives the first rate. Since $y^k \in \Omega$, convexity and smoothness imply

$$\begin{aligned} 0 \le P(y^k) - P^\star &\le \langle g(x^k), y^k - x^\star \rangle + \frac{L}{2}\|e_k\|^2 \\ &= \frac{1}{\beta}\langle \Pi e_k, y^k - x^\star \rangle + \frac{L}{2}\|e_k\|^2. \end{aligned}$$

Moreover, $\|y^k - x^\star\| \le 2\sqrt{\mathcal{D}_k(x^\star)}$. Taking expectations and applying Cauchy–Schwarz yields

$$\mathbf{E}[P(y^k) - P^\star] \le \frac{2\sqrt{B}}{\beta}\sqrt{\mathbf{E}\|e_k\|^2} + \frac{L}{2}\mathbf{E}\|e_k\|^2.$$

Choose a deterministic index minimizing the expected squared residual to obtain the second rate. This completes the proof. □

D Conditional concavity and nonconvex sequence convergence

The deterministic finite-length framework (Attouch et al., 2013) provides the KL geometry, while the conditioning argument of Fest et al. (2026, proof of Lemma 4.2) motivates (2.8). We assume that inequality for the specified process and prove the additive-error extension below; we do not invoke Fest's full theorem without checking its additional hypotheses.

D.1 Conditional concavity with a random limiting level

Write $D_k = V_k - \mathbf{E}_k V_{k+1} \ge aR_k^2$ and $U_k = V_k - V_\infty$. The gap U_k need not be $\mathcal{F}_k$-measurable. Consequently neither V_∞ nor $\varphi'(U_k)$ is moved outside a conditional expectation. Instead, the assumed inequality (2.8) directly yields

$$\varphi(U_k) - \mathbf{E}_k\varphi(U_{k+1}) \ge U_k^{-\theta} D_k \quad (U_k > 0).$$

At $U_k = 0$, its extended-value interpretation forces $D_k = 0$ and hence $R_k = 0$, and also $\mathbf{E}_k\varphi(U_{k+1}) = 0$. This handles zero gaps without assuming that a random zero-gap event belongs to $\mathcal{F}_k$.

D.2 Residual summability with an additive error tail

Lemma D.1 (Residual summability)**.** *Let V_k be an integrable adapted process converging to an integrable V_∞, and let $R_k \ge 0$ be finite and adapted. Suppose $D_k \ge aR_k^2$, $a > 0$, and (2.8) holds with nonnegative gaps $U_k = V_k - V_\infty$, $\theta \in [1/2, 1)$, and the zero-gap convention above. Assume every $\varphi(U_k)$ is integrable. Let $r_k \ge 0$ be deterministic with $\sum_k r_k^\theta < \infty$, and set $v_k = U_k - r_k$. Suppose almost surely there exist $\gamma > 0$ and a finite integer N, both allowed to depend on the trajectory, such that $R_k \ge \gamma v_k^\theta$ for $k \ge N$ whenever $v_k > 0$. Then $\sum_k R_k < \infty$ almost surely.*

Proof. Set $W_k = R_k^2/U_k^\theta$ on $U_k > 0$, and $W_k = 0$ otherwise. Conditional concavity, including its zero-gap convention, gives

$$\mathbf{E}_k\varphi(U_{k+1}) + aW_k \le \varphi(U_k). \tag{D.1}$$

All terms are nonnegative. Taking expectations and summing, using the tower property only on $\varphi(U_{k+1})$, yields

$$a\sum_{k=0}^{n}\mathbf{E}W_k + \mathbf{E}\varphi(U_{n+1}) \le \mathbf{E}\varphi(U_0), \qquad \mathbf{E}\sum_{k\ge 0} W_k \le a^{-1}\mathbf{E}\varphi(U_0) < \infty. \tag{D.2}$$

This step does not require W_k or U_k to be adapted. Fix a trajectory where the weighted sum is finite and the eventual residual bound holds. If $v_k \ge r_k$ and $U_k > 0$, then $U_k = v_k + r_k \le 2v_k$, so $W_k \ge \gamma 2^{-\theta} R_k$. If $v_k < r_k$ and $U_k > 0$, then $U_k < 2r_k$ and Young's inequality gives

$$R_k \le \frac{R_k^2}{U_k^\theta} + \frac{U_k^\theta}{4} \le W_k + \frac{2^\theta}{4} r_k^\theta.$$

This second case includes negative uncorrected gaps v_k. At $U_k = 0$, $R_k = 0$. Thus, for $k \ge N$,

$$R_k \le \max\{1, 2^\theta/\gamma\} W_k + \frac{2^\theta}{4} r_k^\theta. \tag{D.3}$$

The coefficient is finite on the fixed trajectory, so the conclusion follows. No expectation of the random coefficient and no stopping-time property of N are needed. □

D.3 From conditional to realized summability

Lemma D.2 (From conditional summability to realized summability)**.** *Let $Y_k \ge 0$ be $\mathcal{F}_{k+1}$-measurable and set $a_k = \mathbf{E}_k Y_k$. If $\sum_k a_k < \infty$ almost surely, then $\sum_k Y_k < \infty$ almost surely.*

Proof. Write $A_k = \sum_{j=0}^{k} a_j$. For every positive integer m, the variable $\mathbf{1}_{\{A_k \le m\}}$ is $\mathcal{F}_k$-measurable. Conditional expectation and Tonelli's theorem give

$$\mathbf{E}\sum_{k\ge 0} \mathbf{1}_{\{A_k\le m\}} Y_k = \mathbf{E}\sum_{k\ge 0} \mathbf{1}_{\{A_k\le m\}} a_k \le m.$$

Each truncated sum is therefore finite almost surely. Take the intersection of these probability-one events over all positive integers m. On a trajectory with $\sum_k a_k < \infty$, choose an integer $m > \sum_k a_k$; every indicator is then equal to 1. This proves the claim. □

D.4 Proof of main Theorem 2.3

We use precisely the containment and almost-sure boundedness (2.9), the local KL property of the FBE (2.10), error schedule (2.11), and conditional concavity (2.8) stated in the main theorem.

Proof. **Integrability.** Set $m = \inf_\Omega P > -\infty$. Since $\beta L < 1$, the descent lemma and $T_\beta(x) = x - e_\beta(x) \in \Omega$ give

$$\mathcal{E}_\beta(x) \ge P(T_\beta(x)) + \frac{1-\beta L}{2\beta}\|e_\beta(x)\|^2 \ge m.$$

Thus $V_k \ge m$, and its integrable supermartingale property gives $\mathbf{E}(V_k - m) \le \mathbf{E}(V_0 - m) < \infty$. Theorem 2.1 and $r_k \to 0$ give $V_k \to V_\infty$ almost surely. Fatou's lemma yields $V_\infty \ge m$ and $\mathbf{E}(V_\infty - m) \le \mathbf{E}(V_0 - m)$, so V_∞ is integrable. The assumed nonnegative gaps satisfy $0 \le U_k = V_k - V_\infty \le V_k - m$. Therefore

$$0 \le \varphi(U_k) = \frac{U_k^{1-\theta}}{1-\theta} \le \frac{1 + V_k - m}{1-\theta},$$

which proves integrability of every $\varphi(U_k)$ without a deterministic bound on the trajectories.

Cluster sets and their energy. Theorem 2.1 gives $e_k \to 0$ and convergence of the FBE to V_∞. Fix a trajectory in the probability-one event where these properties, containment in the closed region $\mathcal{K}$, and boundedness hold, and denote its cluster set by $\mathcal{A}$. Boundedness gives a nonempty compact cluster set, and closedness of $\mathcal{K}$ ensures $\mathcal{A} \subset \mathcal{K}$. Consequently,

$$\varnothing \ne \mathcal{A} \subset \mathcal{S}, \qquad \operatorname{dist}(x^k, \mathcal{A}) \to 0, \qquad \mathcal{E}_\beta|_{\mathcal{A}} = V_\infty.$$

Indeed, a subsequence staying a positive distance from $\mathcal{A}$ would have a further convergent subsequence by boundedness, contradicting the definition of $\mathcal{A}$. Residual continuity gives criticality, and energy continuity gives the common value. Other critical components may have different energies.

KL uniformization. For each $z \in \mathcal{A}$, use (2.10). A finite collection of the neighborhoods $\mathcal{U}_z$ covers $\mathcal{A}$. Take the minima of their positive constants c_z and η_z, denoted $c_{\mathcal{A}}$ and $\eta_{\mathcal{A}}$. Their union contains a neighborhood of the compact set $\mathcal{A}$. Since all centers have energy V_∞, the common exponent yields

$$\operatorname{dist}(0, \partial\mathcal{E}_\beta(x^k)) \geq c_{\mathcal{A}}\big(\mathcal{E}_\beta(x^k) - V_\infty\big)^\theta$$

eventually whenever the energy gap is positive. This is the compact-set uniformization argument (Bolte et al., 2014, Lemma 6) applied to the trajectory's cluster set; no additional expected-KL property is assumed. The resulting constants and entry index may depend on the trajectory.

Residual summability. Set $R_k = \|e_k\|$, $U_k = V_k - V_\infty \geq 0$ and $v_k = \mathcal{E}_\beta(x^k) - V_\infty = U_k - r_k$. Equation (2.12) gives $D_k \geq aR_k^2$. The slope estimate (B.5) and the preceding KL bound give $R_k \geq \gamma v_k^\theta$ eventually on positive-gap steps, with $\gamma = c_{\mathcal{A}}/(\beta^{-1} + L) > 0$. Lemma D.1 applies by (2.11) and the assumed conditional concavity, proving $\sum_k \|e_k\| < \infty$ almost surely. In particular, the proof never subtracts a random V_∞ from a conditional descent inequality as though it were predictable.

Finite length and point convergence. Equations (B.3), (2.2), and (2.11) imply

$$\sum_k \mathbf{E}_k\|s_k\| \leq q(1+\sqrt{\rho})\sum_k \|e_k\| + q\sum_k \varepsilon_k < \infty \quad \text{a.s.}$$

Lemma D.2 gives realized finite length. Thus x^k converges to $x^\infty \in \mathcal{S}$. Residual continuity, Lemma B.1, and the FBE identity give criticality and $\mathcal{E}_\beta(x^\infty) = P(x^\infty) = V_\infty$. □

D.5 Geometry, tolerance schedules, and scope

KL analyses combine geometry with control of the iterates: local convergence near minimizers is studied by Attouch et al. (2013), compact constant-energy sets underpin KL uniformization (Bolte et al., 2014), and stochastic random reshuffling uses almost-sure boundedness and local KL conditions (Li et al., 2023). These motivate the boundedness condition; conditional concavity remains the separate requirement (2.8).

Verification for a connected set of local minimizers. Suppose the iterates remain almost surely in a deterministic compact region $\mathcal{K}$ whose nonempty critical set $\mathcal{S}_{\mathcal{K}} = \{z \in \mathcal{K} : e_\beta(z) = 0\}$ lies in a connected set of local minimizers with common energy ℓ, and $\mathcal{E}_\beta \geq \ell$ on $\mathcal{K}$. Theorem 2.1 and compactness ensure cluster points in $\mathcal{S}_{\mathcal{K}}$; continuity therefore identifies $V_\infty = \ell$. This does not presuppose point convergence. Now $U_k = V_k - \ell \geq 0$ is adapted, and ordinary supporting-line concavity gives, on $U_k > 0$,

$$\varphi(U_k) - \mathbf{E}_k\varphi(U_{k+1}) \geq \varphi'(U_k)(U_k - \mathbf{E}_k U_{k+1}) = \varphi'(U_k)(V_k - \mathbf{E}_k V_{k+1}).$$

At $U_k = 0$, nonnegativity and the supermartingale property force $\mathbf{E}_k U_{k+1} = 0$, $D_k = 0$, and $U_{k+1} = 0$ conditionally almost surely. The endpoint convention in (2.8) is therefore satisfied as well. This verifies the general theorem's conditioning requirement. The decisive facts are the common level, the lower bound, and almost-sure containment of the iterates in $\mathcal{K}$; connectedness alone does not supply them. This sufficient case is not used in the general proof above, and conditioning on the future event that all iterates remain in $\mathcal{K}$ would not preserve the original conditional estimates automatically.

KL geometry and tolerances. For semialgebraic $C^{1,1}$ targets, the FBE is locally Lipschitz and semialgebraic by (B.7). Semialgebraic KL theory (Bolte et al., 2007) and compact-set uniformization (Bolte et al., 2014) provide the local power estimates used here. The theorem states a common exponent explicitly; general KL geometry is not substituted for this power requirement. Exponents below $1/2$ can be increased to $1/2$ on gaps at most one.

If $\varepsilon_k \leq C_\varepsilon \tau^k$, $0 < \tau < 1$, then

$$r_k \leq \frac{c_\varepsilon C_\varepsilon^2}{1-\tau^2}\tau^{2k}, \qquad \sum_k r_k^\theta \leq \frac{(c_\varepsilon C_\varepsilon^2/(1-\tau^2))^\theta}{1-\tau^{2\theta}} < \infty.$$

A polynomial schedule $\varepsilon_k = O((k+1)^{-p})$ suffices if $p > 1$ and $\theta(2p-1) > 1$. Mere square summability does not imply these conditions. Finally, residual summability implies $\sum_k \mathbf{E}_k\|\Theta_k\| < \infty$; Lemma D.2 then gives $\sum_k \|\Theta_k\| < \infty$ almost surely. This is a consequence of the analysis, not a pathwise error assumption.

E Proofs for persistent-client implementations

E.1 Implementation notation and local solvability

FES-DR and FES-MA use (G_1, H_1) and (G_2, H_2), respectively. Let $F(x) = \sum_i f_i(x_i)$, $\Phi(w) = N^{-1}\sum_i f_i(w)$, and $\bar{x} = N^{-1}\sum_i x_i$. The two target potentials are $P_1(x) = F(x)$ and $P_2(x) = N\Phi(\bar{x})$, with gradients $G_1(x) = (\nabla f_i(x_i))_i$ and $H_2(x) = \mathbf{1}_N \otimes \nabla\Phi(\bar{x})$. The implementation field $G_2(x) = (\nabla f_i(\bar{x}))_i$ need not be conservative; $H_1 = \Pi G_1$ and $H_2 = \Pi G_2$. If the f_i have L-Lipschitz gradients, so do G_1 and H_2, since

$$\|H_2(x) - H_2(y)\| \le \sqrt{N}L\|\bar{x} - \bar{y}\| \le L\|x - y\|.$$

At consensus, either projected target gradient is $\mathbf{1}_N \otimes \nabla\Phi(w)$, so the residual zeros correspond to the intended constrained critical points.

For selected clients, the algorithm forms

$$v_i^k = x_i^k + \beta(B_k(x^k))_i - \gamma_k\big(x_i^k - \bar{x}^k + \beta(\widehat{C}_k)_i\big), \qquad u_i^{k+1} \approx \operatorname*{argmin}_u \left\{ f_i(u) + \frac{\|u - v_i^k\|^2}{2\beta\mu_k} \right\}.$$

The actual state is $x_i^{k+1} = u_i^{k+1}$ on selected blocks and $x_i^{k+1} = x_i^k$ elsewhere. D_k denotes this block-selection projector. Set $\eta_{i,k} = u_i^{k+1} + \beta\mu_k\nabla f_i(u_i^{k+1}) - v_i^k$ on selected blocks and zero elsewhere. The server maintains the mean using actual increments, not merely the mean of selected responses.

Lemma E.1 (Local solvability and equation residuals)**.** *For convex f_i, the local minimizer exists uniquely for every $\beta\mu_k > 0$. For L-smooth nonconvex f_i, the same holds if $\beta\mu_k L < 1$. For the exact response $u^\star$ to the same center and an approximate response with residual η, their distance is bounded by $\|\eta\|$ in the convex case and by $(1 - \beta\mu_k L)^{-1}\|\eta\|$ in the nonconvex case.*

Proof. If f_i is convex, its affine lower bound shows that $f_i(u) + \|u - v\|^2/(2\beta\mu)$ is coercive and strongly convex. If f_i is L-smooth and nonconvex, its quadratic lower bound shows strong convexity and coercivity whenever $\beta\mu L < 1$. Thus the corresponding local minimizer exists and is unique. Strong monotonicity of $u \mapsto u + \beta\mu\nabla f_i(u)$ gives, for an approximate response and the exact response to the same center,

$$\|u - u^\star\| \le \begin{cases} \|\eta\|, & f_i \text{ convex}, \\ (1 - \beta\mu L)^{-1}\|\eta\|, & \beta\mu L < 1. \end{cases}$$

This local response bound is not itself a bound on the total implicit error of a coupled target. □

E.2 Exact-solve sampling for FES-DR

Proposition E.2 (Exact persistent-state sampling)**.** *Suppose each f_i is convex and L-smooth and* $\operatorname{argmin}\Phi \neq \varnothing$. *In Algorithm 2, take $P = P_1$, $\mu_k = 1$, $B_k = G_1$, $\widehat{C}_k = H_1(x^k)$, and exact local solutions. Suppose every client has the same conditional inclusion probability $p \in (0, 1]$, and let $\gamma_k = \gamma \in (0, 1]$. No independence between client indicators within a round is required. With $\alpha_k = p\gamma$, the actual error is*

$$\Theta_k = \gamma(pI - D_k)e_k.$$

Consequently, (2.2) *and* (2.6) *hold with*

$$\rho = \gamma^2 p(1-p), \qquad \varepsilon_k = \delta_k = 0.$$

Taking $a_0 = p\gamma$ gives $a_0(2 - a_0) - \rho = p\gamma(2 - \gamma) > 0$. If the initial mapped distance is integrable, all conclusions of Theorem 2.2 apply for every $\beta > 0$.

Proof. For a selected client, exact local optimality gives

$$x_i^{k+1} + \beta \nabla f_i(x_i^{k+1}) = x_i^k + \beta \nabla f_i(x_i^k) - \gamma\big(x_i^k - \bar{x}^k + \beta(H_1(x^k))_i\big).$$

For an unselected client, both the state and its own block of G_1 remain unchanged. Stacking the equations yields

$$M_\beta(x^{k+1}) = M_\beta(x^k) - \gamma D_k e_k.$$

With $\alpha_k = p\gamma$, this is (B.1) with the stated Θ_k. Since e_k is measurable at the beginning of the round and $\mathbf{E}_k D_k = pI$, its conditional mean is zero. Writing $d_{i,k}$ for the client indicator and using orthogonality of different blocks,

$$\begin{aligned}\mathbf{E}_k\|\Theta_k\|^2 &= \gamma^2 \sum_{i=1}^N \mathbf{E}_k (d_{i,k} - p)^2 \, \|(e_k)_i\|^2 \\ &= \gamma^2 p(1-p)\|e_k\|^2.\end{aligned}$$

Only the marginal conditional inclusion probabilities were used. The constrained solution set is nonempty because $P_1(\mathbf{1}_N \otimes w) = N\Phi(w)$. The stated computation of the positive convex decrement verifies the remaining parameter requirement of Theorem 2.2. □

E.3 The full-state implicit-error identity

Proposition E.3 (Full-state implicit-error identity)**.** *Let $g = \nabla P$ be either target above, and let the actual persistent transition be as in Algorithm 2. For any predictable α_k, its total error in* (B.1) *satisfies*

$$\begin{aligned}\Theta_k = {} & (\alpha_k I - \gamma_k D_k) e_k \\ & + \beta D_k (B_k(x^k) - g(x^k)) \\ & - \gamma_k \beta D_k (\widehat{C}_k - \Pi g(x^k)) \\ & + \beta D_k \big(g(x^{k+1}) - \mu_k \nabla F(x^{k+1})\big) \\ & + \beta (I - D_k)\big(g(x^{k+1}) - g(x^k)\big) + \eta_k.\end{aligned}$$

Proof. For each selected block, the definition of the local equation residual gives

$$u_i^{k+1} - x_i^k = \beta(B_k(x^k))_i - \gamma_k\big((Qx^k)_i + \beta(\widehat{C}_k)_i\big) - \beta\mu_k \nabla f_i(u_i^{k+1}) + \eta_{i,k}.$$

The state difference is zero outside the selected blocks, so

$$s_k = D_k\Big[\beta B_k(x^k) - \gamma_k(Qx^k + \beta\widehat{C}_k) - \beta\mu_k \nabla F(x^{k+1})\Big] + \eta_k.$$

Add $\beta(g(x^{k+1}) - g(x^k)) + \alpha_k e_k$ to both sides. Using $e_k = Qx^k + \beta\Pi g(x^k)$ and grouping selected and unselected gradient increments gives the asserted identity. In particular,

$$D_k g(x^k) + g(x^{k+1}) - g(x^k) = D_k g(x^{k+1}) + (I - D_k)(g(x^{k+1}) - g(x^k)),$$

which explains the term for the unselected blocks. The calculation is an identity for the actual output and makes no exchange between conditional expectation and a nonlinear map. □

For $g = H_2$, the unselected states stay fixed, but their target-gradient blocks change with the global average. The $(I - D_k)$ term must therefore be retained. This identity alone does not certify a conditional error budget.

E.4 Proof of main Proposition 3.1

Suppose $\mu_k = 1$, $B_k = G_2$, $g = H_2$, $t = \beta L < 1$, and $\gamma_k = \gamma > 0$. Let $\mathbf{E}_k D_k = pI$, with $p \in (0, 1]$ and $\alpha_k = p\gamma \in [a_0, 1]$. Define

$$\xi_k = \widehat{C}_k - H_2(x^k), \quad z_k = \eta_k - \gamma\beta D_k \xi_k, \quad \zeta_k = (\mathbf{E}_k\|z_k\|^2)^{1/2},$$

$$R = \gamma\sqrt{p(1-p)} + \frac{t\sqrt{p}(1 + t + 2\gamma)}{1 - t}, \quad A = \frac{1+t}{1-t}.$$

We prove $(\mathbf{E}_k\|\Theta_k\|^2)^{1/2} \le R\|e_k\| + A\zeta_k$. If $\zeta_k \le u\|e_k\| + v_k$ with deterministic budgets, then for any $\lambda > 0$ one may take

$$\rho = (1+\lambda)(R+Au)^2, \qquad \varepsilon_k = \sqrt{1+\lambda^{-1}}\, Av_k,$$

provided $\sum_k v_k^2 < \infty$.

Proof. Let $(\Delta_k)_i = \nabla f_i(x_i^{k+1}) - \nabla f_i(\bar{x}^k)$. The local equation with $\mu_k = 1$ and the unchanged unselected states give

$$s_k = -\gamma D_k e_k - \beta D_k \Delta_k + z_k.$$

Since $s_k = D_k s_k$ and the gradients are L-Lipschitz,

$$\|D_k\Delta_k\| \le L\big(\|D_k Q x^k\| + \|s_k\|\big).$$

Consequently, the conditional L^2 triangle inequality and $\mathbf{E}_k D_k = pI$ imply

$$S_k := (\mathbf{E}_k\|s_k\|^2)^{1/2} \le \frac{\sqrt{p}(\gamma+t)}{1-t}\|e_k\| + \frac{\zeta_k}{1-t}.$$

Here $\|Qx^k\| \le \|e_k\|$, and for every predictable vector v, $\mathbf{E}_k\|D_k v\|^2 = p\|v\|^2$.

By the definition of the actual error,

$$\Theta_k = \gamma(pI - D_k)e_k - \beta D_k\Delta_k + \beta\big(g(x^{k+1}) - g(x^k)\big) + z_k.$$

The target $g = H_2$ is L-Lipschitz on the product space. Moreover,

$$\mathbf{E}_k\|(pI - D_k)e_k\|^2 = p(1-p)\|e_k\|^2.$$

It follows that

$$\begin{aligned}(\mathbf{E}_k\|\Theta_k\|^2)^{1/2} &\le \big(\gamma\sqrt{p(1-p)} + t\sqrt{p}\big)\|e_k\| + 2tS_k + \zeta_k \\ &\le R\|e_k\| + A\zeta_k.\end{aligned}$$

The last claim follows from $(a+b)^2 \le (1+\lambda)a^2 + (1+\lambda^{-1})b^2$. All estimates concern conditional moments of the actual random output; no pathwise relative bound on Θ_k was assumed. □

E.5 Verification of the feasible parameter example

Take $p = 1$, exact H_2, $\gamma = a_0 = 1/2$, $t = 1/256$, and $\zeta_k \le v_k$. With $u = 0$ and $\lambda = 1$,

$$R = \frac{171}{21760}, \qquad A = \frac{257}{255}, \qquad \rho = 2R^2 \approx 1.2351066 \times 10^{-4}, \qquad \varepsilon_k = \sqrt{2}Av_k.$$

Substitution into the main-paper constants gives $c_1\sqrt{\rho} + c_2\rho \approx 0.022508 < 0.0625 = a_0/8$ and $t < 1/16$. Thus this instance meets the sufficient expected-descent regime while retaining client history. First-order and tail summability of geometric v_k follow from Appendix D.5.

The increment bounded above compares the same client's gradients at two points and needs only smoothness. Sampling variance in the correction remains inside z_k. The coefficient R treats centered selection conservatively, so the calculation does not certify arbitrary participation rates or the empirical hyperparameters. If $\mu_k \ne 1$, additional local-gradient terms remain in Proposition E.3; this proof cannot be applied unchanged. A fixed number of local SGD epochs likewise does not imply the stated residual schedule.

F Proofs for the heterogeneity mechanism

F.1 Complete local and mean responses

We use the comparison in main Section 4: FES-DR and FES-MA start from the same x, with full participation, equal weights, exact gradients, $\mu = 1$, and exact local responses. Write $w = N^{-1}\sum_i x_i$, $q_i = x_i - w$, and $h = \nabla\Phi(w)$. For $j = 1, 2$ the centers are

$$\begin{aligned}v_i^{(1)} &= x_i + \beta\nabla f_i(x_i) - \gamma(q_i + \beta h_1), \\ v_i^{(2)} &= x_i + \beta\nabla f_i(w) - \gamma(q_i + \beta h), \\ h_1 &= N^{-1}\sum_i \nabla f_i(x_i).\end{aligned}$$

Let $f_i(u) = u^\top A_i u/2 + b_i^\top u + c_i$, with A_i symmetric and $I + \beta A_i \succ 0$. Set $B_i = (I + \beta A_i)^{-1}$, $\bar{B} = N^{-1}\sum_i B_i$, $C_q = N^{-1}\sum_i B_i q_i$, and $\bar{d} = N^{-1}\sum_i A_i q_i$.

Proposition F.1 (Full quadratic response formulas)**.** *The local and mean responses satisfy*

$$\begin{aligned} x_i^{+,(2)} - w &= B_i((1-\gamma)q_i - \gamma\beta h), \\ x_i^{+,(1)} - w &= q_i - \gamma B_i(q_i + \beta h + \beta\bar{d}), \\ w_2^+ - w &= -\gamma\beta\bar{B}h + (1-\gamma)C_q, \\ w_1^+ - w &= -\gamma\beta\bar{B}h - \gamma C_q - \gamma\beta\bar{B}\bar{d}. \end{aligned}$$

The last two equalities are main Proposition 4.1; the first two restore its local-response details.

Proof. For either method, exact local optimality gives $(I + \beta A_i)x_i^{+,(j)} = v_i^{(j)} - \beta b_i$. For FES-MA, subtract $(I + \beta A_i)w$ and substitute the center to obtain

$$(I + \beta A_i)(x_i^{+,(2)} - w) = (1-\gamma)q_i - \gamma\beta h.$$

For FES-DR, use $h_1 = h + \bar{d}$ to obtain

$$(I + \beta A_i)(x_i^{+,(1)} - w) = (I + \beta A_i)q_i - \gamma q_i - \gamma\beta(h + \bar{d}).$$

Multiply by B_i and average. The identity $\sum_i q_i = 0$ gives the asserted mean formulas. Setting every $q_i = 0$ gives the common consensus-input response and hence the expressions for Δ_j. □

F.2 Disagreement coupling and the removed intermediate bounds

Define $\Delta_j(q) = w_j^+(w + q_1, \ldots, w + q_N) - w_j^+(w, \ldots, w)$. Proposition F.1 gives $\Delta_2 = (1-\gamma)C_q$ and $\Delta_1 = -\gamma C_q - \gamma\beta\bar{B}\bar{d}$.

Lemma F.2 (Disagreement–response bounds)**.** *Write* $\bar{A} = N^{-1}\sum_i A_i$, $r_q^2 = N^{-1}\sum_i \|q_i\|^2$, $\sigma_A^2 = N^{-1}\sum_i \|A_i - \bar{A}\|_{\mathrm{op}}^2$, *and* $\sigma_B^2 = N^{-1}\sum_i \|B_i - \bar{B}\|_{\mathrm{op}}^2$. *Then*

$$\bar{d} = \frac{1}{N}\sum_i (A_i - \bar{A})q_i, \quad C_q = \frac{1}{N}\sum_i (B_i - \bar{B})q_i, \qquad \|\bar{d}\| \le \sigma_A r_q, \quad \|C_q\| \le \sigma_B r_q.$$

In particular, equal curvature matrices imply $\Delta_1 = \Delta_2 = 0$ *for arbitrary linear offsets* b_i.

Proof. The identities follow from $\sum_i q_i = 0$. Bound each product by the operator norm times $\|q_i\|$, sum, and apply Cauchy–Schwarz over clients. If the A_i coincide, so do the B_i, and both centered sums vanish. □

The bounds identify curvature–disagreement coupling, but do not order $\|\Delta_1\|$ and $\|\Delta_2\|$, because cancellation within Δ_1 is possible.

F.3 Derivation of main Example 4.2

Take $f_1(u) = (a+\delta)u^2/2 + bu$, $f_2(u) = (a-\delta)u^2/2 - bu$, $a, \delta, r, \beta, \gamma > 0$, $1 + \beta(a - \delta) > 0$, and $(x_1, x_2) = (r, -r)$. Then $\Phi(u) = au^2/2$ and $w = 0$, although the full state is not in consensus. The two inverse local responses are $B_1 = [1 + \beta(a+\delta)]^{-1}$ and $B_2 = [1 + \beta(a-\delta)]^{-1}$. Set $D = (1 + \beta a)^2 - \beta^2\delta^2$. Their denominators are positive, so $D > 0$. At the specified initial state,

$$h = 0, \qquad \bar{d} = \delta r, \qquad \bar{B} = \frac{1 + \beta a}{D}, \qquad C_q = \frac{r}{2}(B_1 - B_2) = -\frac{\beta\delta r}{D}.$$

Substitution in Proposition 4.1 gives the stated means. The perturbation ratio follows because $a, \beta, \delta, r, \gamma$ are positive. Since $\Phi(u) = au^2/2$, the same strict ordering holds for the increases in $\Phi(w_j^+)$.

For fixed a, β, r, γ, each nonzero magnitude is a positive constant times δ/D. Its derivative is

$$\frac{d}{d\delta}\frac{\delta}{(1 + \beta a)^2 - \beta^2\delta^2} = \frac{(1 + \beta a)^2 + \beta^2\delta^2}{D^2} > 0.$$

The ratio is constant in δ. In particular, increasing heterogeneity enlarges the absolute difference in the favorable regime without changing the relative factor. The parameter b cancels from both responses. If $\delta > a$, one client loss is nonconvex while the global objective remains strongly convex; the formula remains valid under the same positive-denominator condition. The resulting expressions are

$$w_2^+ = -\frac{(1-\gamma)\beta\delta r}{D}, \qquad w_1^+ = -\frac{\gamma\beta^2 a\delta r}{D}, \qquad \frac{|w_2^+|}{|w_1^+|} = \frac{|1-\gamma|}{\gamma\beta a}.$$

Thus $|1-\gamma| < \gamma\beta a$ characterizes the strict advantage in this example. At $a = 2$, $\delta = r = 1$, $\beta = 0.2$, and $\gamma = 0.9$, the means are -0.0375 and $-1/96$ for FES-DR and FES-MA. Their objective increases are 0.00140625 and $1/9216$, giving the factor 12.96 reported in the main text. These compare the global objective evaluated at the client-state mean, not FBE comparisons or a certification of the sufficient convergence constants.

F.4 Cancellation beyond quadratic losses

Proposition F.3 (FES-MA response at unit local relaxation)**.** *Suppose each local proximal subproblem has a unique solution. At $\gamma = 1$, the exact FES-MA output depends on the initial full client state only through its average w. In particular $\Delta_2(q) = 0$ for every mean-zero disagreement q. For L-smooth nonconvex losses, $\beta L < 1$ is sufficient for the required uniqueness.*

Proof. At $\gamma = 1$, $v_i^{(2)} = w + \beta\nabla f_i(w) - \beta\nabla\Phi(w)$, independent of q_i. Uniqueness of each response proves independence of the whole output from the initial disagreement. For L-smooth nonconvex f_i, the local objective is $(\beta^{-1} - L)$-strongly convex when $\beta L < 1$, as established in Appendix E.1. □

For $\gamma \neq 1$, the individual history term $(1-\gamma)x_i$ remains in the center. The quadratic ordering therefore cannot be generalized to all nonlinear objectives without additional analysis. This proposition restores the formal statement behind the main-paper "Beyond quadratics" paragraph.