ASYMPTOTIC ANYTIME-VALID INFERENCE FOR FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated models are often inspected or released repeatedly during training, making uncertainty quantification at a single fixed horizon insufficient. We develop asymptotic anytime-valid inference for smooth, locally identifiable federated M-estimation with full-participation Local SGD and deterministic intermittent communication. For the Polyak–Ruppert average of synchronized iterates, we establish an almost-sure Gaussian approximation that accounts for both total local computation and the variance clock of communication-round averages. The approximation retains the exact finite-schedule normalization and controls the accumulated local drift and noise-freezing errors. Combining it with Gaussian time-uniform boundaries and a strongly consistent covariance estimator yields confidence sequences with asymptotic time-uniform coverage after any fixed, prespecified multiplicative inflation. We also prove strong consistency of a federated sandwich estimator based on fresh samples. Experiments on generalized linear models and a public-health dataset examine coverage, precision, and sensitivity to communication schedules.

1 INTRODUCTION

Federated learning (FL) trains models from decentralized data while limiting communication and data sharing (Konecny et al., 2016; McMahan et al., 2017). In FedAvg, clients perform several local updates before the server aggregates their parameters (McMahan et al., 2017). Analyses of Local SGD show that infrequent synchronization can preserve the statistical efficiency of centralized mini-batch methods under suitable smoothness or convexity conditions (Stich, 2019; Khaled et al., 2020). Heterogeneous client distributions introduce local drift, motivating corrections such as FedProx and SCAFFOLD (Li et al., 2020; Karimireddy et al., 2020); Kairouz et al. (2021) survey these and related developments. Most of this literature studies optimization error. Statistical uncertainty at the models released during training is less well understood.

Inference for stochastic optimization often uses the Polyak–Ruppert average rather than the final iterate. Averaging yields asymptotically efficient estimators under standard stochastic-approximation conditions (Polyak & Juditsky, 1992; Toulis & Airoldi, 2017). Plug-in covariance estimators and covariance-cancelling batch-means procedures provide practical inference for SGD-based M-estimation (Chen et al., 2020; Zhu & Dong, 2021). For federated learning, Li et al. (2022, Theorem 4.2) establish a functional central limit theorem for averages of synchronized Local-SGD iterates under varying communication intervals, with fixed-horizon plug-in and random-scaling inference in their Corollaries 4.1 and 4.3; see also Cai et al. (2026). A region calibrated at a preassigned round need not retain its coverage when that round is selected from the observed training path.

Repeated monitoring arises naturally at synchronization rounds, when a global model is available for inspection or release. A practitioner may stop once a coefficient or prediction contrast reaches a target precision, select a checkpoint based on its uncertainty, or use the current interval width to decide whether another communication round is worthwhile. These decisions call for simultaneous coverage across release rounds. We study asymptotic time-uniform coverage, defined in Definition 1, for the Polyak–Ruppert average of synchronized iterates.

Confidence sequences provide time-uniform coverage under adaptive stopping (Darling & Robbins, 1967). Nonnegative supermartingales, e-processes, and time-uniform concentration yield nonasymptotic constructions for a broad range of statistical problems (Howard et al., 2021; Ramdas et al.,

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.

2023). Strong invariance principles provide an asymptotic alternative: Waudby-Smith et al. (2024) develop time-uniform central limit theory and asymptotic confidence sequences. Xie et al. (2024, Theorem 3.1) give an almost-sure Gaussian approximation for centralized averaged stochastic approximation and derive plug-in asymptotic confidence sequences in their Theorem 4.4. Dalal et al. (2024) develop related inference for double/debiased machine learning. Bonnerjee et al. (2025, Theorems 3.1 and 3.2) establish covariance-matching Gaussian approximations for partial sums of decentralized Local-SGD trajectories under a fixed synchronization period, with maximal errors controlled in probability. Our Theorem 1 concerns an almost-sure Brownian approximation for averages of synchronized iterates under deterministic, varying communication intervals.

Intermittent communication creates two difficulties for this analysis. Gradients are evaluated along client-specific trajectories, producing within-round drift. In addition, the total local computation $t_T = \sum_{m<T} E_m$ and the variance clock $v_T = \sum_{m<T} E_m^{-1}$ evolve at different rates, where E_m is the number of local steps in round m. A strong approximation must control the accumulated drift and state-dependent noise on the inference scale $T/\sqrt{v_T}$. Fixed-time weak convergence alone does not provide this pathwise control. The synchronized decomposition, the clocks t_T and v_T, and the covariance $S = \sum_k p_k^2 S_k$ have precedents in Local-SGD inference (Li et al., 2022).

We establish such an approximation under local regularity, moment, stability, and communication-schedule conditions. Theorem 1 couples the synchronized average to Brownian motion on the exact clock v_T, with an almost-sure error negligible at the inference scale. The communication-specific results are Lemmas 3 and 5: the former controls accumulated within-round drift, and the latter controls the state-dependent-to-frozen-noise error uniformly over release rounds. Combined with Polyak–Ruppert kernel bounds (Polyak & Juditsky, 1992), these estimates reduce the synchronized average to independent, generally non-identically distributed round noises up to an error $o_{\mathrm{a.s.}}(\sqrt{v_T}/T)$. We then apply an existing strong invariance principle (Philipp, 1986). A conditional-distribution construction places the coupled noise, iterates, remainders, and covariance estimates on one probability space.

Corollary 2 converts this approximation into asymptotic confidence sequences using Gaussian time-uniform boundaries, which have asymptotic time-uniform coverage. The construction accepts any strongly consistent positive-definite covariance estimator; we provide a fresh-sample federated sandwich estimator and prove its consistency. Experiments on generalized linear models and a public-health dataset compare coverage and region width, and examine the effects of client heterogeneity, dimension, and communication schedules.

Section 2 presents the estimator, assumptions, and main results. Section 3 reports the experiments. Proofs and additional results appear in the appendix.

2 Methodology

2.1 Problem setup

Consider K clients with local data distributions $\mathcal{P}_k$ and weights $p_k > 0$, where $\sum_{k=1}^K p_k = 1$. Client $k \in [K] = \{1, \dots, K\}$ has sample loss $f_k(x; \xi^k)$ for $\xi^k \sim \mathcal{P}_k$. The population objective is $f(x) := \sum_{k=1}^K p_k f_k(x)$, where $f_k(x) := \mathbb{E}[f_k(x; \xi^k)]$ and $x \in \mathbb{R}^d$. Let $x^\star$ denote the unique minimizer of f.

The arrays $(\xi_t^k)_{t \geq 0, k \in [K]}$ are independent across clients and times, with $\xi_t^k \sim \mathcal{P}_k$, and the stochastic gradients are unbiased at their arguments. All gradient evaluations are finite. Clients start from the same deterministic x_0. Let U denote the complete training array together with the optional fresh-sample array (ζ_m^k) used below, on a standard-Borel data space, and set

$$\mathcal{F}_t = \sigma(\xi_s^k : s < t,\ k \in [K]), \qquad \mathcal{H}_T = \mathcal{F}_{t_T} \vee \sigma(\zeta_m^k : m \leq T,\ k \in [K]).$$

The $\mathcal{H}_T$ are the release-time information sets; when fresh samples are not used, their term is omitted. The iterates and covariance estimates are measurable functions of U.

Communication occurs at deterministic times $0 = t_0 < t_1 < t_2 < \cdots$, where $t_{m+1} = t_m + E_m$, $E_m \in \{1, 2, \dots\}$. During communication round m, each client performs E_m local stochastic-gradient steps before synchronization using a roundwise constant step size η_m according to

$$x_{t+1}^k = x_t^k - \eta_m \nabla f_k(x_t^k; \xi_t^k), \qquad t = t_m, \dots, t_{m+1} - 1.$$

At the end of round m, the server aggregates the local models and broadcasts the synchronized iterate back to all clients: $\bar{x}_{t_{m+1}} := \sum_{k=1}^{K} p_k x_{t_{m+1}}^k$ and $x_{t_{m+1}}^k \leftarrow \bar{x}_{t_{m+1}}$ for all $k \in [K]$.

We estimate $x^\star$ by the Polyak–Ruppert average of synchronized iterates, $\bar{y}_T := T^{-1} \sum_{m=1}^{T} \bar{x}_{t_m}$, and seek simultaneous coverage over sufficiently late communication rounds.

Definition 1 (Asymptotic time-uniform coverage)**.** *A sequence of random regions $(\mathcal{C}_T)$ has asymptotic time-uniform coverage at level $1-\alpha$ for $x^\star$ if*

$$\lim_{T_0 \to \infty} \mathbb{P}(\forall T \geq T_0 : \; x^\star \in \mathcal{C}_T) \geq 1 - \alpha. \tag{1}$$

We refer to such a sequence as an asymptotic $(1-\alpha)$ confidence sequence. The limit is taken in the monitoring start T_0; it does not specify a finite start at which nominal coverage is guaranteed.

2.2 Gaussian approximation

Define $\epsilon_k(x) = \nabla f_k(x; \xi^k) - \nabla f_k(x)$ and $\epsilon(x) = \sum_{k=1}^{K} p_k \epsilon_k(x)$. Let $\gamma_m := E_m \eta_m$ be the effective step size and $s_m := \bar{x}_{t_m} - x^\star$. We use two clocks: $t_T := \sum_{m=0}^{T-1} E_m$ counts local steps per client, while $v_T := \sum_{m=0}^{T-1} E_m^{-1}$ measures the accumulated variance of roundwise noise averages. The following assumptions specify local regularity, noise moments, the communication schedule, and stability.

Assumption 1 (Local identification and linearization)**.** *The population objective $f = \sum_{k=1}^{K} p_k f_k$ is differentiable near $x^\star$, with $\nabla f(x^\star) = 0$. There exist a symmetric positive-definite matrix G with $\lambda_{\min}(G) \geq \mu > 0$, constants $r_0, L_R > 0$, and $\lambda \in (0, 1]$ such that*

$$\|\nabla f(x) - G(x - x^\star)\| \leq L_R \|x - x^\star\|^{1+\lambda}, \qquad \|x - x^\star\| \leq r_0. \tag{2}$$

Assumption 2 (Gradient noise and quadratic-mean regularity)**.** *The client/time samples are independent under the full-participation data model. For*

$$\epsilon_k(x) = \nabla f_k(x; \xi) - \nabla f_k(x), \qquad S_k = \mathbb{E}\left[\epsilon_k(x^\star)\epsilon_k(x^\star)^\top\right],$$

there are $B > 0$, $p > 2$, $M_{2p} < \infty$, and $L < \infty$ such that, for every client k,

$$\sup_{\|x - x^\star\| \leq B} \mathbb{E}\,\|\epsilon_k(x)\|^{2p} \leq M_{2p}. \tag{3}$$

$$\left\{\mathbb{E}\,\|\nabla f_k(x; \xi) - \nabla f_k(y; \xi)\|^2\right\}^{1/2} \leq L\,\|x - y\| \tag{4}$$

whenever $x, y \in \mathcal{B}_B(x^\star)$, where $\mathcal{B}_B(x^\star) = \{x : \|x - x^\star\| \leq B\}$. The matrix $S = \sum_{k=1}^{K} p_k^2 S_k$ is positive definite, and $\Sigma = G^{-1} S G^{-\top}$.

Assumption 3 (Step sizes and communication schedule)**.** *For the $p > 2$ and $\lambda \in (0, 1]$ specified above, choose*

$$a \in \left(\max\left\{\frac{1}{2}, \frac{1}{1+\lambda}\right\}, 1 - \frac{1}{p}\right), \qquad b \in [0, a(1+\lambda) - 1), \qquad \gamma_m \asymp (m+1)^{-a}, \tag{5}$$

where the interval for a is nonempty and γ_m is nonincreasing. The deterministic sequence E_m is either uniformly bounded or eventually nondecreasing with $E_m \asymp (m+1)^b$. Set $b = 0$ in the uniformly bounded case. Write

$$\nu_T = \frac{t_T v_T}{T^2}. \tag{6}$$

These schedules satisfy $v_T \to \infty$ and $1 \leq \nu_T \leq C$. The fixed-time CLT in Corollary 1 additionally assumes $\nu_T \to \nu \in [1, \infty)$.

Assumption 4 (Localization and mean-square stability)**.** *The synchronized recursion satisfies*

$$s_m \to 0 \quad \textit{almost surely}, \qquad \mathbb{E}\,\|s_m\|^2 \leq C_s \gamma_m \tag{7}$$

for all sufficiently large m and some $C_s < \infty$.

A nonempty range in (5) requires $p > (1+\lambda)/\lambda$. For $\lambda = 1$, this reduces to $p > 2$ and $b < 2a - 1$. The local expansion controls the nonlinear remainder, and the quadratic-mean increment condition controls within-round displacement and noise freezing (Li et al., 2022; Xie et al., 2024). The latter is weaker than samplewise Lipschitz continuity. Proposition 2 gives sufficient global strong-convexity and moment conditions for Assumption 4.

Positive curvature and diminishing step sizes yield the Polyak–Ruppert kernel bounds (Polyak & Juditsky, 1992; Zhu & Dong, 2021), while the noise moments support the strong approximation (Philipp, 1986). Sample-Hessian conditions are needed for the covariance estimator in Assumption 5 (Chen et al., 2020; Xie et al., 2024).

The strict polynomial inequalities also allow $E_m \asymp (m+1)^b d_m$ with $d_m + d_m^{-1} = O(\log^{\vartheta}(m+2))$ for some $\vartheta > 0$, subject to the same bounded-or-nondecreasing requirement and $\sup_T \nu_T < \infty$. This includes the logarithmic schedule used in the experiments. These factors add logarithmic terms to the bounds without changing their summability. We retain the exact factor ν_T throughout; its limit is used only in Corollary 1.

The remaining constants govern contraction, localization, and the approximation remainder; see Appendix A.3. The proof localizes synchronized iterates by Assumption 4 and controls within-round client-path exits by Lemma 1.

Client count and heterogeneity affect the leading covariance and drift remainder. Write $g_k^\star = \nabla f_k(x^\star)$, $\sum_{k=1}^K p_k g_k^\star = \nabla f(x^\star) = 0$, and $\zeta_\star^2 = \sum_{k=1}^K p_k \|g_k^\star\|^2$. The first-order covariance is

$$S = \sum_{k=1}^K p_k^2 S_k, \qquad \Sigma = G^{-1} S G^{-\top}. \tag{8}$$

Client count enters through the weights and client covariances. If $p_k = 1/K$ and $S_k = S_0$ for all k, then $S = S_0/K$ and marginal widths scale as $K^{-1/2}$ at fixed per-client local sample count E_m. For identical S_k and unequal weights, this dependence is summarized by the effective client count $K_{\text{eff}} = \frac{1}{\sum_{k=1}^K p_k^2}$.

For heterogeneous S_k, gains depend on the full covariance matrices. With $S_k = S_0$, $p_k = 1/K$, and fixed total gradient budget $N_m = KE_m$ per round, $\mathrm{Cov}(\widetilde{\varepsilon}_m) = \frac{S_0/K}{E_m} = \frac{S_0}{N_m}$, which is independent of K.

Under full participation, $\sum_k p_k g_k^\star = 0$ cancels deterministic heterogeneity from the leading synchronized gradient. Heterogeneity still affects S_k, G, and the within-round drift through $\zeta_\star$. The stopped drift variables in Lemma 1 satisfy

$$\mathbb{E}(\|\delta_m^{\mathrm{loc}}\|^2 \mid \mathcal{F}_{t_m}) \le C\{L^2, M_{2p}, \zeta_\star^2\}\gamma_m^2(1 + \|s_m\|^2).$$

The schedule restriction makes the accumulated drift negligible relative to $\sqrt{v_T}/T$. Finite-horizon optimization bounds can still be governed by heterogeneity-induced drift (Khaled et al., 2020), whereas the diminishing-step regime here makes its accumulated contribution to the first-order asymptotic law negligible. Appendix B.4 examines the effects of client count and heterogeneity at finite horizons.

Define the frozen round noises by

$$\varepsilon_m^\circ := \widetilde{\varepsilon}_m = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \sum_{k=1}^K p_k \{\nabla f_k(x^\star; \xi_t^k) - \nabla f_k(x^\star)\}. \tag{9}$$

These noises are independent across rounds, with covariance S/E_m, and need not be identically distributed. Let $F(U) = (G^{-1}\widetilde{\varepsilon}_m)_{m \ge 0}$.

Theorem 1 (Strong Gaussian approximation at release rounds)**.** *Suppose Assumptions 1–4 hold and the raw sample space is standard Borel. On an extension there exist (i) a raw-data array $U^\dagger$ with the same law as the original array U, (ii) the entire Local-SGD process reconstructed from $U^\dagger$, including $\bar{y}_T^\dagger$, and (iii) a standard d-dimensional Brownian motion W, such that*

$$\sqrt{t_T}(\bar{y}_T^\dagger - x^\star) = \frac{\sqrt{t_T}}{T} \Sigma^{1/2} W(v_T) + o_{\mathrm{a.s.}}(1). \tag{10}$$

Equivalently, with $\nu_T = t_T v_T / T^2$,

$$\Sigma^{-1/2}\sqrt{\frac{t_T}{\nu_T}}(\bar{y}_T^\dagger - x^\star) = \frac{W(v_T)}{\sqrt{v_T}} + o_{\mathrm{a.s.}}(1), \qquad \sqrt{\frac{t_T}{\nu_T}} = \frac{T}{\sqrt{v_T}}. \tag{11}$$

The Brownian increments also give i.i.d. $Z_1, Z_2, \ldots \sim N(0, I_d)$ such that

$$\sqrt{t_T}(\bar{y}_T^\dagger - x^\star) = \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T-1}\frac{1}{\sqrt{E_m}}\Sigma^{1/2}Z_{m+1} + o_{\mathrm{a.s.}}(1). \tag{12}$$

The sign of the leading Polyak–Ruppert noise is absorbed by replacing W with $-W$.

The exact factor ν_T accounts for schedule normalization and the iterated-logarithm fluctuations of $W(v_T)/\sqrt{v_T}$. Replacing it by ν in the almost-sure approximation requires a rate such as $|\nu_T/\nu - 1|\sqrt{\log\log v_T} \to 0$; the exact normalization $T/\sqrt{v_T}$ does not.

Lemma 8 reconstructs the iterates, remainders, and covariance estimates from the same $U^\dagger$. Equality in law transfers the resulting distributional statements to the original algorithm.

Corollary 1 (Fixed-time asymptotic normality)**.** *Under Theorem 1, if $\nu_T \to \nu \in (0, \infty)$, then*

$$\sqrt{t_T}(\bar{y}_T - x^\star) \Rightarrow N(0, \nu\Sigma).$$

This follows from $W(v_T)/\sqrt{v_T} \sim N(0, I_d)$ at deterministic $v_T > 0$ and Slutsky's theorem.

For confidence-sequence construction, let $\widetilde{\Sigma}_T$ be positive definite for every T and satisfy $\widetilde{\Sigma}_T \to \Sigma$ almost surely. An eventually positive-definite raw estimator $\bar{\Sigma}_T \to \Sigma \succ 0$ can be extended to every round by setting

$$H_T = \frac{\bar{\Sigma}_T + \bar{\Sigma}_T^\top}{2}, \qquad \widetilde{\Sigma}_T = \begin{cases} H_T, & \lambda_{\min}(H_T) > 0, \\ I_d, & \lambda_{\min}(H_T) \le 0. \end{cases}$$

The fallback is used only finitely often almost surely and preserves almost-sure consistency. For a fresh-sample sandwich estimator, let $\zeta_m^k \sim \mathcal{P}_k$ be independent of the training array and across clients and rounds. Define

$$\widehat{G}_T = \frac{1}{T}\sum_{m=1}^{T}\sum_{k=1}^{K} p_k \nabla^2 f_k(\bar{x}_{t_m}; \zeta_m^k), \qquad g_m = \sum_{k=1}^{K} p_k \nabla f_k(\bar{x}_{t_m}; \zeta_m^k). \tag{13}$$

$$\widehat{S}_T = \frac{1}{T}\sum_{m=1}^{T} g_m g_m^\top, \qquad \widehat{\Sigma}_T = \begin{cases} \widehat{G}_T^{-1}\widehat{S}_T\widehat{G}_T^{-\top}, & \widehat{G}_T \text{ is nonsingular}, \\ 0, & \widehat{G}_T \text{ is singular}. \end{cases} \tag{14}$$

This defines the raw estimator at every round, including rounds in which $\widehat{G}_T$ is singular.

Assumption 5 (Fresh-sample covariance estimation)**.** *All gradients and sample Hessians in the construction are defined and finite at the evaluated parameters. For every client k, the sample Hessian exists near $x^\star$, and differentiation can be interchanged with expectation:*

$$\mathbb{E}\{\nabla^2 f_k(x; \zeta^k)\} = \nabla^2 f_k(x), \qquad G = \sum_{k=1}^{K} p_k \nabla^2 f_k(x^\star),$$

and

$$\mathbb{E}\left\|\nabla^2 f_k(x^\star; \zeta^k)\right\|^2 < \infty. \tag{15}$$

$$\left\{\mathbb{E}\left\|\nabla^2 f_k(x; \zeta^k) - \nabla^2 f_k(y; \zeta^k)\right\|^2\right\}^{1/2} \le L_H \left\|x - y\right\|. \tag{16}$$

In the same neighborhood, $\sup_x \mathbb{E}\left\|\nabla f_k(x; \zeta^k)\right\|^4 < \infty$.

Proposition 1 (Strong consistency and eventual positive definiteness)**.** *Under Assumptions 1–4 and 5,*

$$\widehat{G}_T \to G, \qquad \widehat{S}_T \to S, \qquad \widehat{\Sigma}_T \to \Sigma \quad \textit{almost surely}.$$

Moreover,

$$\mathbb{P}\left(\begin{array}{l} \exists T_0(\omega) < \infty \textit{ such that, for every } T \geq T_0(\omega), \\ \lambda_{\min}\left(\dfrac{\widehat{G}_T + \widehat{G}_T^\top}{2}\right) \geq \dfrac{1}{2}\lambda_{\min}(G) > 0, \quad \lambda_{\min}(\widehat{S}_T) \geq \dfrac{1}{2}\lambda_{\min}(S) > 0, \quad \widehat{\Sigma}_T \succ 0 \end{array} \right) = 1.$$

Applying the fallback to $\widehat{\Sigma}_T$ gives $\widetilde{\Sigma}_T \succ 0$ at every release while preserving almost-sure consistency.

The fixed-time Wald region, with exact normalization $T^2/v_T = t_T/\nu_T$, is

$$\mathrm{CI}_T^{\mathrm{fix}}(\alpha) = \left\{ \theta \in \mathbb{R}^d : \frac{T^2}{v_T}(\bar{y}_T - \theta)^\top \widetilde{\Sigma}_T^{-1}(\bar{y}_T - \theta) \leq \chi^2_{d,1-\alpha} \right\}. \tag{17}$$

Local-SGD training is first order; the optional fresh-sample sandwich estimator requires second-order post-processing. Hessian-vector products and iterative linear solves can replace dense inversion for directional or low-dimensional functional inference; full-vector regions require the corresponding covariance representation. Appendix B.6 evaluates sample-reuse accuracy and cost, whereas Proposition 1 establishes fresh-sample consistency. Iterate-only or batch-means alternatives (Chen et al., 2020; Zhu & Dong, 2021) require a separate strong-consistency analysis on the irregular release clock.

In particular,

$$\mathbb{P}\left(\exists T_0 < \infty : \ \sigma_{\min}(\widehat{G}_T) \geq \lambda_{\min}(G)/2 > 0 \text{ for every } T \geq T_0 \right) = 1. \tag{18}$$

Indeed, $\widehat{G}_T \to G$ almost surely in operator norm and $\sigma_{\min}(\widehat{G}_T) \geq \sigma_{\min}(G) - \left\| \widehat{G}_T - G \right\|$, with $\sigma_{\min}(G) = \lambda_{\min}(G) > 0$ since $G \succ 0$.

2.3 Asymptotic anytime-valid inference

We combine Theorem 1 with a consistent covariance estimator and a time-uniform boundary for Gaussian sums.

Definition 2 (Gaussian boundary)**.** *Fix $\alpha \in (0,1)$ and $q = 2, \infty$. A map $u^{(q)}_{\alpha,d} : \mathbb{N} \to \mathbb{R}_+$ is called a $(1-\alpha)$ Gaussian boundary in dimension d if, for i.i.d. $Z_1, Z_2, \cdots \sim \mathcal{N}(0, I_d)$,*

$$\mathbb{P}\left(\forall n \geq 1 : \ \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \right\|_q \leq u^{(q)}_{\alpha,d}(n) \right) \geq 1 - \alpha. \tag{19}$$

A fixed multiplicative inflation absorbs the approximation and covariance-estimation errors.

Corollary 2. *Assume the conditions of Theorem 1, and let an $\mathcal{H}_T$-measurable estimator $\widetilde{\Sigma}_T \succ 0$ satisfy $\widetilde{\Sigma}_T \to \Sigma$ almost surely. Fix $\alpha \in (0,1)$ and any pre-specified $\delta > 0$. If $u^{(q)}_{\alpha,d}$ satisfies Definition 2, define*

$$\mathcal{C}_T^{(\delta)} := \left\{ \theta \in \mathbb{R}^d : \left\| \widetilde{\Sigma}_T^{-1/2}(\bar{y}_T - \theta) \right\|_q \leq (1+\delta) u^{(q)}_{\alpha,d}(n_T) \frac{\sqrt{v_T}}{T} \right\}, \quad n_T := \lfloor v_T \rfloor \vee 1. \tag{20}$$

Then,

$$\lim_{T_0 \to \infty} \mathbb{P}\Big(\forall T \geq T_0 : x^\star \in \mathcal{C}_T^{(\delta)} \Big) \geq 1 - \alpha.$$

On the Gaussian-boundary event, δ eventually absorbs the coupling, Brownian interpolation, and covariance-estimation errors. A shrinking inflation would require deterministic rates for all three errors. In one dimension, the interval is

$$\bar{y}_T \ \pm \ (1+\delta)\sqrt{\widetilde{\Sigma}_T}\, u^{(q)}_{\alpha,1}(n_T) \frac{\sqrt{v_T}}{T}. \tag{21}$$

The experiments use $\delta = 0$ to evaluate the first-order construction, while the coverage result applies to any fixed $\delta > 0$. The following boundaries give explicit Euclidean and sup-norm regions.

Corollary 3 (Two explicit Gaussian boundaries). *Let $n_0 > 0$ be fixed and pre-specified. The function W_{-1} is the lower branch of the Lambert W function.*

For $q = 2$, define

$$u^{(2)}_{\alpha,d}(n) = \sqrt{d\left(1 + \frac{n_0}{n\lambda^*_{\alpha,d}}\right)\log\left(\frac{1 + n\lambda^*_{\alpha,d}/n_0}{\alpha^{2/d}}\right)}, \tag{22}$$

where

$$\lambda^*_{\alpha,d} = -W_{-1}\left(-\alpha^{2/d}/e\right) - 1. \tag{23}$$

For $q = \infty$, define, with natural logarithms,

$$u^{(\infty)}_{\alpha,d}(n) = 1.7\sqrt{\log\log(2n) + 0.72\log(10.4d/\alpha)}. \tag{24}$$

Both are $(1-\alpha)$ Gaussian boundaries and may be used in Corollary 2.

For $S_n = \sum_{i=1}^n Z_i$, Gaussian mixing gives the nonnegative martingale

$$M_n(\rho) = \left(\frac{\rho}{\rho + n}\right)^{d/2} \exp\left\{\frac{\|S_n\|_2^2}{2(\rho + n)}\right\}. \tag{25}$$

Ville's inequality yields the Euclidean boundary. Any pre-specified $n_0 > 0$ and $\lambda > 0$ give a valid boundary with $\rho = n_0/\lambda$; $\lambda^\star_{\alpha,d}$ minimizes the radius at n_0. The sup-norm result follows by allocating level α/d to each two-sided scalar finite-LIL boundary and taking a union bound. The derivations are given in the proof of Corollary 3. Appendix B.3 discusses design-point selection and monitoring-start sensitivity.

The choice $q = 2$ gives ellipsoids, while $q = \infty$ gives parallelepipeds aligned with whitened coordinates. Marginal studentization, scalar tail arguments, and a union bound give an axis-aligned simultaneous band under the same fixed-inflation conditions:

$$\theta_j \in \bar{y}_{T,j} \;\pm\; (1+\delta)\sqrt{\widetilde{\Sigma}_{T,jj}}\, u^{(\infty)}_{\alpha/d,1}(n_T)\frac{\sqrt{v_T}}{T}, \qquad j \in [d].$$

2.4 Examples

We specialize the covariance estimator to the three regression models in Section 3. The approximation and confidence-sequence results apply under the assumptions above. Throughout, $\xi^k = (a^k, b^k)$ and fresh samples are $\zeta^k_m = (\tilde{a}^k_m, \tilde{b}^k_m)$.

Example 1 (Linear regression). *Consider the squared loss $f_k(x;\xi^k) := \frac{1}{2}\left\{b^k - (a^k)^\top x\right\}^2$. Under the linear data model in Section 3.1, conditional on the realized local targets,*

$$f_k(x) = \frac{1}{2}\mathbb{E}\left\{b^k - (a^k)^\top x\right\}^2 = \frac{1}{2}\|x - x^\star_k\|^2 + \frac{1}{2},$$

so the global target is $x^\star = \sum_{k=1}^K p_k x^\star_k$. Writing $\delta_k := x^\star - x^\star_k$, we have $G = \nabla^2 f(x^\star) = I_d$ and $\epsilon_k(x^\star) = \left\{a^k(a^k)^\top - I_d\right\}\delta_k - a^k\varepsilon^k$, which yields $S_k = \mathbb{E}\left\{\epsilon_k(x^\star)\epsilon_k(x^\star)^\top\right\} = (1 + \|\delta_k\|^2)I_d + \delta_k\delta_k^\top$. Hence $\Sigma = \sum_{k=1}^K p_k^2\left\{(1+\|\delta_k\|^2)I_d + \delta_k\delta_k^\top\right\}$. The fresh-sample plug-in quantities are

$$\widehat{G}_T = \frac{1}{T}\sum_{m=1}^T\sum_{k=1}^K p_k\,\tilde{a}^k_m(\tilde{a}^k_m)^\top,\; g_m = \sum_{k=1}^K p_k\,\tilde{a}^k_m\left\{(\tilde{a}^k_m)^\top\bar{x}_{t_m} - \tilde{b}^k_m\right\}.$$

Example 2 (Logistic regression). *Use the logistic loss $f_k(x;\xi^k) := \log\left\{1 + e^{(a^k)^\top x}\right\} - b^k(a^k)^\top x$. The Hessian $G_k := \nabla^2 f_k(x^\star) = \mathbb{E}\left[\sigma\left((a^k)^\top x^\star\right)\left\{1 - \sigma\left((a^k)^\top x^\star\right)\right\}a^k(a^k)^\top\right]$, where $\sigma(u) :=$*

$(1+e^{-u})^{-1}$. *Moreover,* $\epsilon_k(x^\star) = \left\{\sigma\big((a^k)^\top x^\star\big) - b^k\right\} a^k - \mathbb{E}\left[\left\{\sigma\big((a^k)^\top x^\star\big) - b^k\right\} a^k\right]$. *The fresh-sample plug-in quantities are* $g_m = \sum_{k=1}^K p_k \left\{\sigma\big((\tilde{a}_m^k)^\top \bar{x}_{t_m}\big) - \tilde{b}_m^k\right\} \tilde{a}_m^k$, *and*

$$\widehat{G}_T = \frac{1}{T}\sum_{m=1}^{T}\sum_{k=1}^{K} p_k\, \sigma\big((\tilde{a}_m^k)^\top \bar{x}_{t_m}\big)\left\{1-\sigma\big((\tilde{a}_m^k)^\top \bar{x}_{t_m}\big)\right\} \tilde{a}_m^k (\tilde{a}_m^k)^\top.$$

Example 3 (Poisson regression)**.** *Take the Poisson negative log-likelihood* $f_k(x;\xi^k) := \exp\left\{(a^k)^\top x\right\} - b^k (a^k)^\top x$. *Then* $G_k := \nabla^2 f_k(x^\star) = \mathbb{E}\left[\exp\left\{(a^k)^\top x^\star\right\} a^k (a^k)^\top\right]$, *and* $\epsilon_k(x^\star) = \left[\exp\left\{(a^k)^\top x^\star\right\} - b^k\right] a^k - \mathbb{E}\left(\left[\exp\left\{(a^k)^\top x^\star\right\} - b^k\right] a^k\right)$. *The fresh-sample plug-in quantities are*

$$\widehat{G}_T = \frac{1}{T}\sum_{m=1}^{T}\sum_{k=1}^{K} p_k\, \exp\left\{(\tilde{a}_m^k)^\top \bar{x}_{t_m}\right\} \tilde{a}_m^k (\tilde{a}_m^k)^\top,\ g_m = \sum_{k=1}^{K} p_k \left[\exp\left\{(\tilde{a}_m^k)^\top \bar{x}_{t_m}\right\} - \tilde{b}_m^k\right] \tilde{a}_m^k.$$

Substituting these model-specific quantities into (14), with the positive-definite fallback above, gives the fixed-time region (17) and the confidence sequence (20).

3 Experiments

In this section, we evaluate the finite-sample performance of the proposed inference methods through simulation studies. Due to the page limit, the real-data application is presented in Appendix B.2.

3.1 Simulation setting

We use $K = 10$ clients, dimension $d = 5$, and equal weights $p_k = 1/K$. Data follow the three regression models in Section 2.4:

- **Linear** (adapted from Li et al. (2022)): $b_t^k = (a_t^k)^\top x_k^\star + \varepsilon_t^k$, where $x_k^\star \sim \mathcal{N}(0, I_d)$, $a_t^k \sim \mathcal{N}(0, I_d)$, and $\varepsilon_t^k \sim \mathcal{N}(0,1)$, with global target $x^\star = K^{-1}\sum_{k=1}^K x_k^\star$.
- **Logistic** (adapted from Li et al. (2022)): $\mathbb{P}(b_t^k = 1) = \left[1+\exp\{-(a_t^k)^\top x^\star\}\right]^{-1}$, where $x^\star = (0, 1/4, 1/2, 3/4, 1)^\top$ and $a_t^k \sim \mathcal{N}(0, I_d)$.
- **Poisson** (adapted from Fan & Li (2001)): $b_t^k \mid a_t^k$ follows a Poisson distribution with $\mathbb{E}(b_t^k \mid a_t^k) = \exp\,(a_t^k)^\top x^\star$, where $s = (1.2, 0.6, 0, 0, 0.8)^\top$, $x^\star = 0.25s$, $a_t^k \sim \mathcal{N}(c_k s/\,\|s\|\,, \Gamma)$, $\Gamma_{ij} = 0.2^{|i-j|}$, and $c_k = -1.25 + 2.5(k-1)/(K-1)$.

Heterogeneity enters through client-specific local targets $x_k^\star$ in the linear model and covariate means in the Poisson model. We run Local SGD for $T = 50{,}000$ communication rounds from $x_0^k = 0$ for all k. Following Li et al. (2022), we set $\gamma_m = c_0(m+1)^{-0.505}$ and $\eta_m = \gamma_m/E_m$, with $c_0 = 0.5$, 2.0, and 0.2 for the linear, logistic, and Poisson models, respectively. The first 5% of the run is used as a warm-up period with $E_m = 1$, and time-uniform type-I error is reported after the warm-up period. We compare three communication schedules: C_1, with $E_m = 1$; C_5, with $E_m = 5$; and *Log*, with $E_m = \lceil \log_2(m+1) \rceil$. We compare the oracle covariance matrix Σ with plug-in sandwich estimates. In the main simulations and the GM design-point experiment, each client reuses its first training observation in each round, with gradients and Hessians evaluated at the synchronized round-start parameter. Appendix B.6 describes this estimator and examines its accuracy and cost. We use the exact finite-schedule factor ν_T in (6), with $\sqrt{\nu_T/t_T} = \sqrt{v_T}/T$ (Li et al., 2022). We compare LIL-UB and GM, with $n_0 = 20{,}000$ for GM; for reference, we also include the fixed-time Gaussian benchmark described in (17). All reported regions use $\delta = 0$, as described in Section 2.3. Each experiment is replicated $R = 500$ times using 110 Intel® Xeon® Platinum 8352V with 36 cores and 72GB memory.

3.2 Simulation results

Figure 1 reports time-uniform type-I error and average confidence-set volume for the three communication schedules. The fixed-time Gaussian benchmark produces smaller regions but substantially

more pathwise failures. LIL-UB and GM have lower empirical failure rates, with GM generally producing smaller regions than LIL-UB over the displayed range. Oracle and plug-in covariance yield similar results. The logistic and Poisson experiments show similar patterns and appear in Appendix B.1.

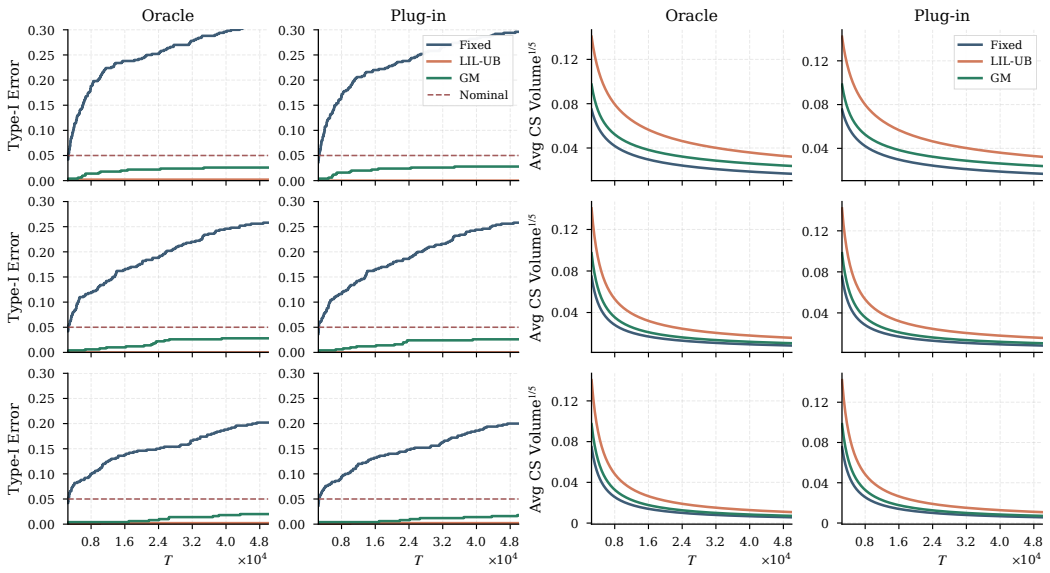


Figure 1: Linear regression: time-uniform type-I error and average confidence-set volume. Rows 1–3 correspond to the C_1, C_5, and *Log* communication schedules, respectively. The first two columns report time-uniform type-I error under oracle and plug-in covariance, and the last two columns report the corresponding average volume of the full $d = 5$ confidence sets. The dashed horizontal line marks the nominal 0.05 level.

Appendix B reports additional experiments on (1) the GM design point and monitoring start; (2) client count and heterogeneity under fixed per-client and total-system budgets; (3) partial participation; (4) covariance-estimation accuracy and cost across dimensions; and (5) communication schedules and schedule normalization. Monte Carlo average first-coordinate trajectories appear in Appendix B.1.

4 Conclusion

We develop asymptotic anytime-valid inference for Polyak–Ruppert averages of synchronized Local SGD iterates. The strong Gaussian approximation controls within-round drift and noise-freezing errors while retaining the exact relationship between total local computation and the variance clock. Combined with Gaussian boundaries and a strongly consistent covariance estimator, it yields computable confidence sequences with asymptotic time-uniform coverage after fixed multiplicative inflation. The experiments examine the resulting coverage–width trade-off and show the importance of finite-schedule normalization.

The analysis assumes smooth, locally identifiable objectives, full participation, deterministic communication, and the stated moment and stability conditions. Random client availability introduces selection variance, while random communication requires a predictable time change; extending the result to these settings calls for a martingale strong approximation and corresponding drift bounds. Privacy perturbations, nonsmooth objectives, and nonidentifiable neural-network parameters require additional modeling and analysis. Quantitative remainder bounds would also provide guidance on the monitoring start. Low-dimensional functionals and prediction heads offer further settings for federated sequential inference under suitable local regularity.

AI USE STATEMENT

Generative AI tools were used to revise wording and improve the clarity and readability of the manuscript; to identify potentially relevant literature, references, and search directions, and to assist with research execution, including drafting, debugging, and refining code. For mathematical proofs, AI was used primarily to review and check the authors arguments for potential errors, inconsistencies, or gaps. AI-assisted manuscript revisions, literature-search suggestions, feedback on research ideas and proofs, and generated or modified code were reviewed and verified by the authors before incorporation into the final work. The authors independently assessed the accuracy and appropriateness of these outputs and retain full responsibility for the methodology, proofs, implementation, analysis, and conclusions presented in the paper.

REPRODUCIBILITY STATEMENT

The simulation code and an accompanying `README` file are submitted with the manuscript. We provide a dedicated reproduction script for each figure and table, allowing the reported results to be reproduced individually. The `README` provides instructions for running the code.

REFERENCES

Soham Bonnerjee, Sayar Karmakar, and Wei Biao Wu. Sharp Gaussian approximations for Decentralized Federated Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5253. URL https://papers.nips.cc/paper_files/paper/2025/hash/e6a11b618402617342f38f5b49430937-Abstract-Conference.html.

Leheng Cai, Qirui Hu, and Shuyuan Wu. Federated learning of quantile inference under local differential privacy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a5bFKVtTyF.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020. doi: 10.1214/18-AOS1801. URL https://doi.org/10.1214/18-AOS1801.

Abhinandan Dalal, Patrick Blöbaum, Shiva Kasiviswanathan, and Aaditya Ramdas. Anytime-valid inference for double/debiased machine learning of causal parameters, 2024. URL https://arxiv.org/abs/2408.09598. arXiv:2408.09598.

D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967. doi: 10.1073/pnas.58.1.66. URL https://doi.org/10.1073/pnas.58.1.66.

R. Duan, Y. Ning, and Y. Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 2022.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Karen Hacker, Craig W Thomas, Guixiang Zhao, JNeka S Claxton, Paul Eke, and Machell Town. Social determinants of health and health-related social needs among adults with chronic diseases in the united states, behavioral risk factor surveillance system, 2022. *Preventing Chronic Disease*, 21:E94, 2024.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2019.

Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local SGD on identical and heterogeneous data. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4519–4529. PMLR, 2020. URL https://proceedings.mlr.press/v108/bayoumi20a.html.

Jakub Konecny, H. Brendan McMahan, Daniel Ramage, and Peter Richtarik. Federated optimization: Distributed machine learning for on-device intelligence, 2016. URL https://arxiv.org/abs/1610.02527. arXiv:1610.02527.

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, pp. 429–450, 2020. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local SGD. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of the Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1613–1661. PMLR, 2022. URL https://proceedings.mlr.press/v178/li22b.html.

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

Walter Philipp. A note on the almost sure approximation of weakly dependent random variables. *Monatshefte für Mathematik*, 102(3):227–236, 1986. doi: 10.1007/BF01294601. URL https://doi.org/10.1007/BF01294601.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894. URL https://doi.org/10.1214/23-STS894.

Herbert Robbins and David Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In Jagdish S. Rustagi (ed.), *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, New York, 1971. doi: 10.1016/B978-0-12-604550-5.50015-8. URL https://doi.org/10.1016/B978-0-12-604550-5.50015-8.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1g2JnRcFX.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017. doi: 10.1214/16-AOS1506. URL https://doi.org/10.1214/16-AOS1506.

Ian Waudby-Smith, David T. Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024. doi: 10.1214/24-AOS2408. URL https://doi.org/10.1214/24-AOS2408.

Chuhan Xie, Kaicheng Jin, Jiadong Liang, and Zhihua Zhang. Asymptotic time-uniform inference for parameters in averaged stochastic approximation, 2024. URL https://arxiv.org/abs/2410.15057. arXiv:2410.15057.

Yi Zhu and Jing Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *2021 Winter Simulation Conference (WSC)*, pp. 1–12. IEEE, 2021. doi: 10.1109/WSC52266.2021.9715437. URL https://doi.org/10.1109/WSC52266.2021.9715437.

A THEORETICAL DETAILS

A.1 CONSEQUENCES OF ASSUMPTION 3

Under Assumption 3, $b < 1$ and therefore $t_T = O(T^{1+b})$ and $v_T \asymp T^{1-b}$. Thus, $\sqrt{t_T}/T \to 0$ and $v_T \to \infty$. Moreover,

$$\frac{\sqrt{t_T}}{T} \sum_{m=0}^{T-1} \gamma_m \lesssim T^{(1+b)/2-a} \to 0,$$

which follows from $b < 2a - 1$.

A.2 DECOMPOSITIONS

Define the matrices

$$B_j := I - \gamma_j G, \qquad j \geq 0, \quad G \text{ as in (2)},$$

and for $0 \leq j \leq n$ define

$$A_j^n := \left(\sum_{l=j}^{n} \prod_{i=j+1}^{l} B_i \right) \gamma_j.$$

We use the convention $\prod_{i=j+1}^{j} B_i = I$ for $j \geq 0$.

Define $s_m = \bar{x}_{t_m} - x^\star$, $\gamma_m = \eta_m E_m$, and $s_{m+1} = s_m - \gamma_m v_m$, where

$$v_m := \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \sum_{k=1}^{K} p_k \nabla f_k(x_t^k; \xi_t^k).$$

Define

$$h_m := \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \sum_{k=1}^{K} p_k \nabla f_k(\bar{x}_{t_m}; \xi_t^k).$$

Let

$$\varepsilon_m := h_m - \nabla f(\bar{x}_{t_m}), \qquad \delta_m := v_m - h_m.$$

Then

$$v_m = \nabla f(\bar{x}_{t_m}) + \varepsilon_m + \delta_m,$$

and therefore

$$s_{m+1} = s_m - \gamma_m \big(\nabla f(\bar{x}_{t_m}) + \varepsilon_m + \delta_m \big).$$

Let

$$r_m := \nabla f(\bar{x}_{t_m}) - G s_m.$$

Then

$$s_{m+1} = B_m s_m - \gamma_m \big(r_m + \delta_m + \varepsilon_m \big), \qquad B_m := I - \gamma_m G.$$

Iterating the recursion gives

$$s_{m+1} = \left(\prod_{j=0}^{m} B_j \right) s_0 - \sum_{j=0}^{m} \left(\prod_{i=j+1}^{m} B_i \right) \gamma_j (v_j - G s_j),$$

and averaging yields

$$\begin{aligned}
\sqrt{t_T}(\bar{y}_T - x^\star) &= \frac{\sqrt{t_T}}{T} \sum_{m=0}^{T-1} s_{m+1} \\
&= \frac{\sqrt{t_T}}{T} \left\{ \frac{1}{\gamma_0} A_0^{T-1} B_0 s_0 - \sum_{m=0}^{T-1} A_m^{T-1} (r_m + \delta_m) - \sum_{m=0}^{T-1} A_m^{T-1} \varepsilon_m \right\} \\
&= \frac{\sqrt{t_T}}{T} \left\{ \frac{1}{\gamma_0} A_0^{T-1} B_0 s_0 - \sum_{m=0}^{T-1} A_m^{T-1} (r_m + \delta_m) - \sum_{m=0}^{T-1} \left(A_m^{T-1} - G^{-1} \right) \varepsilon_m - \sum_{m=0}^{T-1} G^{-1} \varepsilon_m \right\}.
\end{aligned}$$

A.3 TECHNICAL LEMMAS

Proposition 2 (Sufficient global conditions for stability)**.** *Suppose f is globally μ-strongly convex, the client quadratic-mean increment bound in* (4) *holds on $\mathbb{R}^d$, and, for some $C_\epsilon < \infty$,*

$$\mathbb{E}\|\epsilon_k(x)\|^2 \le C_\epsilon(1 + \|x - x^\star\|^2), \qquad x \in \mathbb{R}^d, \quad k \in [K].$$

For a deterministic common initializer and the step sizes in Assumption 3, the synchronized iterates satisfy Assumption 4.

Proof. All constants in this proof may depend on μ, L, C_ϵ, the fixed client weights, $\sum_k p_k\|\nabla f_k(x^\star)\|^2$, and the step-size envelopes. Conditional on $\mathcal{F}_{t_m}$, let $D_m = \max_{k,j\le E_m} \mathbb{E}(\|x^k_{t_m+j} - \bar{x}_{t_m}\|^2 \mid \mathcal{F}_{t_m})$ before the terminal synchronization. The update recursion, Cauchy–Schwarz, and the global second-moment bound give

$$D_m \le C\gamma_m^2(1 + \|s_m\|^2 + D_m).$$

For all sufficiently large m, absorption yields $D_m \le C\gamma_m^2(1 + \|s_m\|^2)$. The definitions of δ_m and ε_m then imply

$$\mathbb{E}(\|\delta_m\|^2 \mid \mathcal{F}_{t_m}) \le C\gamma_m^2(1 + \|s_m\|^2), \qquad \mathbb{E}(\|\varepsilon_m\|^2 \mid \mathcal{F}_{t_m}) \le CE_m^{-1}(1 + \|s_m\|^2). \tag{26}$$

Here $\mathbb{E}(\varepsilon_m \mid \mathcal{F}_{t_m}) = 0$. Expanding $s_{m+1} = s_m - \gamma_m\{\nabla f(\bar{x}_{t_m}) + \varepsilon_m + \delta_m\}$, using $\langle s_m, \nabla f(\bar{x}_{t_m})\rangle \ge \mu\|s_m\|^2$, and applying Young's inequality to the drift cross term, gives constants $c, C > 0$ such that

$$\mathbb{E}(\|s_{m+1}\|^2 \mid \mathcal{F}_{t_m}) \le (1 - c\gamma_m)\|s_m\|^2 + C\gamma_m^2 \tag{27}$$

for all sufficiently large m. The Robbins–Siegmund theorem (Robbins & Siegmund, 1971), together with $\sum_m \gamma_m^2 < \infty$ and $\sum_m \gamma_m = \infty$, implies that $\|s_m\|^2$ has an almost-sure limit and $\sum_m \gamma_m\|s_m\|^2 < \infty$. That limit must be zero.

Taking expectations and iterating (27), for a fixed sufficiently large m_0,

$$\mathbb{E}\|s_m\|^2 \le C\exp\left(-c\sum_{i=m_0}^{m-1}\gamma_i\right) + C\sum_{j=m_0}^{m-1}\gamma_j^2\exp\left(-c\sum_{i=j+1}^{m-1}\gamma_i\right).$$

The terms with $j < m/2$ decay exponentially in m^{1-a}. For $j \ge m/2$, $\gamma_j \asymp m^{-a}$, and summing the resulting geometric series bounds the second term by $Cm^{-2a}m^a = Cm^{-a}$. Thus $\mathbb{E}\|s_m\|^2 \lesssim \gamma_m$. The bound follows from the power envelopes in Assumption 3. □

Lemma 1 (Stopped within-round localization)**.** *Under Assumptions 1–4, every local gradient evaluation lies in $\mathcal{B}_B(x^\star)$ in all sufficiently late rounds almost surely. There are localized drift variables δ_m^{loc} equal to δ_m eventually almost surely and satisfying*

$$\mathbb{E}\|\delta_m^{\mathrm{loc}}\|^2 \le C\gamma_m^2. \tag{28}$$

The constant depends on the local moment envelope, L, B, the client weights, and the mean gradients at $x^\star$.

Proof. Fix $0 < r < \min(B, r_0)$ and set $J_m = \mathbf{1}\{\|s_m\| \le r\}$. For each client, stop its within-round path at its first exit from $\mathcal{B}_B(x^\star)$, including the step that causes the exit. Conditional on $\mathcal{F}_{t_m}$ and $J_m = 1$, all gradients before this exit are evaluated within the local neighborhood. Their second moments are uniformly bounded: the population gradients are bounded there by the quadratic-mean increment condition, and independence of the centered client noises gives

$$\mathbb{E}\|p_k\epsilon_k(x)\|^{2p} \le \mathbb{E}\Big\|\sum_{\ell=1}^K p_\ell\epsilon_\ell(x)\Big\|^{2p}, \qquad x \in \mathcal{B}_B(x^\star),$$

by conditional Jensen's inequality.

Let $D_{m,k}$ be the maximum displacement of this stopped path from $\bar{x}_{t_m}$. Cauchy–Schwarz applied to the sum of at most E_m updates yields

$$\mathbb{E}(J_m D_{m,k}^2 \mid \mathcal{F}_{t_m}) \le C J_m \eta_m^2 E_m^2 = C J_m \gamma_m^2. \tag{29}$$

Consequently,

$$\sum_m \mathbb{P}\{J_m = 1, \text{ some client exits in round } m\} \le \frac{CK}{(B-r)^2} \sum_m \gamma_m^2 < \infty.$$

Borel–Cantelli and $J_m = 1$ eventually almost surely prove the first assertion. Define δ_m^{loc} by multiplying the drift sum by J_m and, for each client, retaining only steps before its exit. The exit indicators are measurable before the next sample is drawn. The quadratic-mean increment condition and (29) therefore give

$$\mathbb{E}\|\delta_m^{\mathrm{loc}}\|^2 \le \frac{L^2}{E_m} \sum_{k=1}^{K} p_k \sum_{t=t_m}^{t_{m+1}-1} \mathbb{E}\{J_m \mathbf{1}_{\{t<\tau_{m,k}\}} \|x_t^k - \bar{x}_{t_m}\|^2\} \le C L^2 \gamma_m^2.$$

Here $\tau_{m,k}$ is the first exit time of the pre-synchronization client path. The stopped and original drift sums agree in every sufficiently late round, leaving a finite random prefix. □

Lemma 2. *Under Assumption 1 and 3,*

$$\frac{\sqrt{t_T}}{T\gamma_0} A_0^{T-1} B_0 s_0 \xrightarrow{\text{a.s.}} 0.$$

Proof. Recall that $B_i = I - \gamma_i G$. The deterministic kernel estimate in Lemma 4 gives uniform boundedness of A_0^{T-1},

$$\left\| \frac{\sqrt{t_T}}{T\gamma_0} A_0^{T-1} B_0 s_0 \right\| \lesssim \frac{\sqrt{t_T}}{T} \frac{\|s_0\|}{\gamma_0}.$$

The right-hand side tends to zero since $\sqrt{t_T}/T \to 0$ as $T \to \infty$. □

Lemma 3 (Accumulated nonlinear remainder and local drift)**.** *Under Assumptions 1–4,*

$$\frac{1}{\sqrt{v_T}} \sum_{m=0}^{T-1} (\|r_m\| + \|\delta_m\|) \xrightarrow{\text{a.s.}} 0, \qquad \frac{1}{T} \sum_{m=0}^{T-1} A_m^{T-1}(r_m + \delta_m) = o_{\text{a.s.}}(\sqrt{v_T}/T). \tag{30}$$

Proof. Use the indicator J_m from Lemma 1. The local expansion and mean-square stability imply

$$\mathbb{E}\|J_m r_m\| \le L_R \{\mathbb{E}\|s_m\|^2\}^{(1+\lambda)/2} \le C \gamma_m^{(1+\lambda)/2}, \qquad \mathbb{E}\|\delta_m^{\mathrm{loc}}\| \le C\gamma_m.$$

The constants depend on L_R, C_s and the displacement constant in (28). At $T_j = 2^j$, let $Q_T = \sum_{m<T}(\|J_m r_m\| + \|\delta_m^{\mathrm{loc}}\|)$.

For each $\epsilon > 0$,

$$\sum_{j\ge 1} \mathbb{P}\{Q_{T_j} > \epsilon \sqrt{v_{T_j}}\} \le \frac{C}{\epsilon} \sum_{j\ge 1} \left[2^{-j\{a(1+\lambda)-1-b\}/2} + 2^{-j(2a-1-b)/2} \right] < \infty. \tag{31}$$

Both exponents are positive under (5). Borel–Cantelli, monotonicity of Q_T, and $v_{2T}/v_T \asymp 1$ yield $Q_T/\sqrt{v_T} \to 0$ almost surely. The original and localized summands differ only over a finite random prefix by Lemma 1. Its sum is finite almost surely and $v_T \to \infty$, proving the first assertion.

Uniform boundedness of A_m^n, established in the proof of Lemma 4, proves the second. Logarithmic schedule factors multiply the summands in (31) by at most powers of j and preserve summability. □

Lemmas 3 and 5 reduce the synchronized average to independent round noises on the variance clock v_T. The local-drift term δ_m captures the effect of within-round client trajectories and vanishes when $E_m = 1$.

Lemma 4 (Kernel remainder)**.** *Under Assumptions 1–4,*

$$\frac{1}{\sqrt{v_T}} \sum_{m=0}^{T-1} (A_m^{T-1} - G^{-1})\varepsilon_m \xrightarrow{\text{a.s.}} 0. \tag{32}$$

Proof. To bound the kernel, diagonalize G, fix an eigenvalue $h > 0$, and write

$$P_{j,l} = \prod_{i=j+1}^{l} (1 - h\gamma_i), \qquad a_j^n = \gamma_j \sum_{l=j}^{n} P_{j,l}.$$

After a fixed initial index all factors are in $(0, 1)$. The power envelopes imply

$$\sum_{l=j}^{\infty} |P_{j,l}| \le C(j+1)^a, \qquad \sum_{j=0}^{r} E_j^{-1} |P_{j,r}| \le C(r+1)^{a-b}. \tag{33}$$

For the first bound, split $l \le 2j$ from $l > 2j$: the first sum is bounded by a geometric series with exponent $-c(l-j)(j+1)^{-a}$, and the second by a stretched-exponential tail. For the second, split $j \ge r/2$ from $j < r/2$ and use the analogous backward bounds. Finitely many initial factors affect only C. In particular, $|a_j^n| \le C$ and hence $\sup_{j \le n} \|A_j^n\| < \infty$.

For sufficiently large j, telescoping gives the exact identity

$$a_j^{T-1} - h^{-1} = \sum_{l=j}^{T-1} (\gamma_j - \gamma_{l+1}) P_{j,l} - h^{-1} P_{j,T}. \tag{34}$$

Put $\Delta\gamma_r = \gamma_r - \gamma_{r+1} \ge 0$. Since $\sum_{l=r}^{\infty} P_{j,l} \le C\gamma_r^{-1} P_{j,r}$, summing (34) with weights E_j^{-1} and using (33) yields

$$\sum_{j<T} \frac{|a_j^{T-1} - h^{-1}|}{E_j} \le C + C \sum_{r<T} \Delta\gamma_r (r+1)^{2a-b} + CT^{a-b} \le CT^{a-b}. \tag{35}$$

The last inequality follows by summation by parts using $\gamma_r \lesssim (r+1)^{-a}$ and $a - b > 0$. Uniform boundedness and the spectral decomposition therefore imply

$$\sum_{m<T} \frac{\|A_m^{T-1} - G^{-1}\|^2}{E_m} \le CT^{a-b}. \tag{36}$$

The constants depend on the eigenvalues of G and the step-size and communication envelopes. The bound follows from monotonicity and the power envelopes.

Set $\varepsilon_m^{\text{loc}} = J_m \varepsilon_m$. This is a martingale difference on the round filtration. Conditional moment inequalities for the average of E_m centered independent gradients give

$$\mathbb{E}(\|\varepsilon_m^{\text{loc}}\|^{2p} \mid \mathcal{F}_{t_m}) \le CE_m^{-p}.$$

Burkholder's inequality and Minkowski's inequality, valid also for noninteger $p > 1$, give

$$\mathbb{E}\left\| \frac{1}{\sqrt{v_T}} \sum_{m<T} (A_m^{T-1} - G^{-1})\varepsilon_m^{\text{loc}} \right\|^{2p} \le \frac{C}{v_T^p} \left(\sum_{m<T} \frac{\|A_m^{T-1} - G^{-1}\|^2}{E_m} \right)^p \le CT^{-p(1-a)}. \tag{37}$$

Thus, for each $\epsilon > 0$, the probabilities of exceeding ϵ have sum bounded by $C\epsilon^{-2p} \sum_{T \ge 1} T^{-p(1-a)} < \infty$ because $a < 1 - 1/p$. Borel–Cantelli proves (32) for the localized noises. Their difference from the original noises is confined to a finite random prefix, whose contribution vanishes by uniform kernel boundedness and $v_T \to \infty$. Logarithmic schedule factors only add logarithmic factors to the summable series. □

Recall the frozen round noise $\widetilde{\varepsilon}_m = \varepsilon^\circ_m$ defined in (9).

Lemma 5 (Noise freezing)**.** *Under Assumptions 1–4,*

$$\frac{1}{\sqrt{v_T}} \max_{0\le n<T} \left\| \sum_{m=0}^{n} G^{-1}(\varepsilon_m - \widetilde{\varepsilon}_m) \right\| \xrightarrow{\text{a.s.}} 0. \tag{38}$$

Proof. Let $Z_m = J_m G^{-1}(\varepsilon_m - \widetilde{\varepsilon}_m)$. These are martingale differences. The quadratic-mean increment condition implies

$$\mathbb{E}(\|Z_m\|^2 \mid \mathcal{F}_{t_m}) \le C\|G^{-1}\|^2 L^2 J_m \|s_m\|^2/E_m, \qquad \mathbb{E}\|Z_m\|^2 \le C\gamma_m/E_m.$$

The constants also depend on C_s and the client weights. Doob's maximal inequality gives

$$\mathbb{E}\max_{n<T} \Big\| \sum_{m=0}^{n} Z_m \Big\|^2 \le C H_{a+b}(T), \qquad H_c(T) = \begin{cases} T^{1-c}, & c<1, \\ \log(T+1), & c=1, \\ 1, & c>1. \end{cases} \tag{39}$$

At $T_j = 2^j$, Markov's inequality gives a summable bound because

$$\sum_{j\ge1} \frac{H_{a+b}(2^j)}{v_{2^j}} \lesssim \begin{cases} \sum_j 2^{-ja}, & a+b<1, \\ \sum_j j2^{-j(1-b)}, & a+b=1, \\ \sum_j 2^{-j(1-b)}, & a+b>1 \end{cases} < \infty. \tag{40}$$

Borel–Cantelli and dyadic interpolation prove the result for Z_m. The original and localized differences agree eventually almost surely; uniform boundedness of the partial sums over the finite exceptional prefix completes the argument. Powers of $\log T$ from the allowed logarithmic schedule factors preserve (40). □

Lemma 6 (Strong approximation for the frozen sequence)**.** *Under Assumptions 2 and 3, there exist on an extension a sequence $X^\dagger_m$ with*

$$(X^\dagger_m)_{m\ge0} \overset{d}{=} (G^{-1}\widetilde{\varepsilon}_m)_{m\ge0}$$

and a standard Brownian motion W such that

$$\left\| \sum_{m=0}^{T-1} X^\dagger_m - \Sigma^{1/2} W(v_T) \right\| = o_{\text{a.s.}}(\sqrt{v_T}). \tag{41}$$

Proof. Set

$$Y_m = G^{-1}\widetilde{\varepsilon}_m, \qquad A_T = \sum_{m<T} \mathrm{Cov}(Y_m) = v_T\Sigma, \qquad V_T = \mathrm{tr}(A_T) = v_T \,\mathrm{tr}(\Sigma).$$

Since $A_T/V_T = \Sigma/\,\mathrm{tr}(\Sigma)$, the covariance-discrepancy term in Theorem 2 of Philipp (1986) is zero. After deleting a finite prefix so that $V_{m+1} > e$, choose

$$f(x) = x^{1/p}(\log x)^{1+1/p}.$$

The required truncation series is finite because

$$\begin{aligned} \sum_m \mathbb{E}\left[\frac{\|Y_m\|^2 \mathbf{1}\{\|Y_m\|^2 > f(V_{m+1})\}}{f(V_{m+1})} \right] &\le C \sum_m \frac{E_m^{-p}}{V_{m+1}(\log V_{m+1})^{p+1}} \\ &\le C \sum_m \frac{E_m^{-1}}{V_{m+1}(\log V_{m+1})^{p+1}} < \infty. \end{aligned}$$

Indeed, $E_m \ge 1$ and $V_{m+1} - V_m = \mathrm{tr}(\Sigma)/E_m$, so the last sum is bounded by the convergent integral of $x^{-1}(\log x)^{-(p+1)}$. For every fixed $c>0$, $x^{-1}(\log x)^c f(x)$ is eventually decreasing, while f is

eventually increasing and diverges. A positive monotone extension defines f on the finite initial range. That theorem then gives, for constants $\kappa_1, \kappa_2 > 0$ determined by its parameters,

$$\left\| \sum_{m<T} X_m^\dagger - \Sigma^{1/2} W(v_T) \right\| = O_{\text{a.s.}}\Big(V_T^{1/2-\kappa_1} (\log V_T)^{\kappa_2} \Big) = o_{\text{a.s.}}(\sqrt{v_T}).$$

Lemma 8 embeds this coupled sequence in a copy of the full algorithm. □

Lemma 7. *Let $1 < A_0 < A_1 < A_2 < \cdots$ with $A_n \to \infty$, and write $\Delta A_n := A_n - A_{n-1} > 0$. Then, for every $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} \frac{\Delta A_n}{A_n (\log A_n)^{1+\epsilon}} < \infty. \tag{42}$$

Proof. The function $x \mapsto [x(\log x)^{1+\epsilon}]^{-1}$ is eventually decreasing. Therefore, for all sufficiently large n,

$$\int_{A_{n-1}}^{A_n} \frac{dx}{x(\log x)^{1+\epsilon}} \geq \frac{A_n - A_{n-1}}{A_n (\log A_n)^{1+\epsilon}}.$$

Summing over n and using

$$\int^{\infty} \frac{dx}{x(\log x)^{1+\epsilon}} < \infty$$

proves the claim. □

A.4 Proofs of main theorems

Lemma 8 (Lifting a frozen-noise coupling to the full algorithm)**.** *Let U be the infinite raw-data array, including training and fresh covariance-estimation samples, taking values in a standard Borel space, and let*

$$F(U) = (G^{-1}\widetilde{\varepsilon}_m)_{m \geq 0}.$$

Suppose $(X^\dagger, W)$ is a coupling supplied by Lemma 6, so $X^\dagger \overset{d}{=} F(U)$. Let $K(x, du)$ be a regular conditional distribution of U given $F(U) = x$. On a further extension, conditionally on $(X^\dagger, W)$ draw

$$U^\dagger \mid X^\dagger \sim K(X^\dagger, \cdot).$$

Then

$$U^\dagger \overset{d}{=} U, \qquad F(U^\dagger) = X^\dagger \quad \textit{almost surely}.$$

Consequently, all Local-SGD iterates and remainder terms reconstructed as measurable functions of $U^\dagger$ have the same joint law as the original algorithm and share the frozen sequence coupled to W.

Proof. For every bounded measurable h,

$$\mathbb{E}h(U^\dagger) = \mathbb{E}\left[\int h(u) K(X^\dagger, du) \right] = \mathbb{E}\left[\int h(u) K(F(U), du) \right] = \mathbb{E}h(U),$$

where the middle equality uses $X^\dagger \overset{d}{=} F(U)$ and the final equality is the defining property of a regular conditional distribution. Moreover, $K(x, \{u : F(u) = x\}) = 1$ for $F(U)$-almost every x, hence also for $X^\dagger$-almost every x. Therefore $F(U^\dagger) = X^\dagger$ almost surely. □

Proof of Theorem 1. Use the copy $U^\dagger$ constructed in Lemma 8. Let $R_{0,T}^\dagger$, $R_{1,T}^\dagger$, and $R_{2,T}^\dagger$ denote, respectively, the initial-condition, nonlinear/local-drift, and kernel-mismatch terms in the final display of Appendix A.2. The communication-round identity gives

$$\sqrt{t_T}(\bar{y}_T^\dagger - x^\star) = R_{0,T}^\dagger + R_{1,T}^\dagger + R_{2,T}^\dagger - \frac{\sqrt{t_T}}{T} \sum_{m<T} G^{-1} \varepsilon_m^\dagger.$$

The initial-condition, linearization/local-drift, and kernel-mismatch lemmas hold almost surely for $U^\dagger$ because $U^\dagger \overset{d}{=} U$. Lemma 5 therefore gives

$$\sqrt{t_T}(\bar{y}_T^\dagger - x^\star) = -\frac{\sqrt{t_T}}{T}\sum_{m<T} X_m^\dagger + o_{\mathrm{a.s.}}(1).$$

By Lemma 6,

$$\sum_{m<T} X_m^\dagger = \Sigma^{1/2} W(v_T) + o_{\mathrm{a.s.}}(\sqrt{v_T}).$$

Since $\sqrt{t_T v_T}/T = \sqrt{\nu_T} = O(1)$, substitution proves Theorem 1; replace W by $-W$ to absorb the sign. Brownian increments over $[v_m, v_{m+1}]$ give the i.i.d. representation, and $T/\sqrt{v_T} = \sqrt{t_T/\nu_T}$ gives the normalized form. □

Proof of Proposition 1. Let $H_m = \sum_k p_k \nabla^2 f_k(\bar{x}_{t_m}; \zeta_m^k)$, $Q_m = g_m g_m^\top$, and

$$\mathcal{K}_m = \mathcal{F}_{t_m} \vee \sigma(\zeta_j^k : j < m,\ k \in [K]),$$

so that $\mathcal{H}_{m-1} \subseteq \mathcal{K}_m \subseteq \mathcal{H}_m$ and the fresh samples in release m are independent of $\mathcal{K}_m$. Use $J_m = \mathbf{1}\{\|s_m\| \le r\}$ from Lemma 1, with r inside the common gradient and Hessian neighborhoods. Define

$$D_m^H = J_m H_m - \mathbb{E}(J_m H_m \mid \mathcal{K}_m), \qquad D_m^Q = J_m Q_m - \mathbb{E}(J_m Q_m \mid \mathcal{K}_m).$$

The local sample-Hessian second moment and gradient fourth moment give

$$\mathbb{E}\|D_m^H\|^2 + \mathbb{E}\|D_m^Q\|^2 \le C, \qquad \sum_{m\ge 1} \frac{\mathbb{E}\|D_m^H\|^2 + \mathbb{E}\|D_m^Q\|^2}{m^2} < \infty.$$

Both sequences are matrix martingale differences on $(\mathcal{H}_m)$ by the tower property. Entrywise martingale convergence for the series weighted by m^{-1}, followed by Kronecker's lemma, implies that their Cesàro averages converge to zero almost surely.

Decompose the localized Hessian average as

$$\frac{1}{T}\sum_{m=1}^T J_m(H_m - G) = \frac{1}{T}\sum_{m=1}^T D_m^H + \frac{1}{T}\sum_{m=1}^T \{\mathbb{E}(J_m H_m \mid \mathcal{K}_m) - J_m G\}.$$

On $J_m = 1$, the conditional mean is $\nabla^2 f(\bar{x}_{t_m})$; on $J_m = 0$, both terms vanish. The second average therefore converges to zero by $\bar{x}_{t_m} \to x^\star$, local continuity, and Cesàro averaging.

Similarly, let $g(x;\zeta) = \sum_k p_k \nabla f_k(x;\zeta^k)$ and $V(x) = \mathbb{E}\{g(x;\zeta)g(x;\zeta)^\top\}$ in the local neighborhood. Then

$$\frac{1}{T}\sum_{m=1}^T J_m(Q_m - S) = \frac{1}{T}\sum_{m=1}^T D_m^Q + \frac{1}{T}\sum_{m=1}^T \{\mathbb{E}(J_m Q_m \mid \mathcal{K}_m) - J_m S\}.$$

On $J_m = 1$, the conditional mean is $V(\bar{x}_{t_m})$. Independence of the fresh client samples and $\nabla f(x^\star) = 0$ give $V(x^\star) = S$. Quadratic-mean continuity and Cauchy–Schwarz imply continuity of V at $x^\star$, so the second average also converges to zero. Since $J_m = 1$ eventually almost surely and all evaluated matrices are finite, the omitted finite random prefix contributes $O_{\mathrm{a.s.}}(T^{-1})$. Thus $\widehat{G}_T \to G$ and $\widehat{S}_T \to S$ almost surely.

Continuity of inversion at $G \succ 0$ gives $\widehat{\Sigma}_T \to \Sigma$, because the singular-case value is used only finitely often almost surely. Weyl's inequality yields the quantified eventual eigenvalue event in Proposition 1. Symmetrization and the identity fallback make the implemented covariance positive definite for every T while preserving its limit. The fixed-time Wald region follows from the normalized Gaussian approximation and Slutsky's theorem. □

Lemma 9 (Brownian unit-interval modulus)**.** *For a standard d-dimensional Brownian motion,*

$$\sup_{0\le u\le 1} \|W(n+u) - W(n)\|_\infty = O(\sqrt{\log n}) \quad \textit{almost surely}.$$

Consequently,

$$\frac{1}{\sqrt{n}} \sup_{0\le u\le 1} \|W(n+u) - W(n)\|_q \longrightarrow 0 \quad \textit{almost surely}$$

for every fixed $q \in [1,\infty]$.

Proof. The reflection principle and a union bound give, for $x > 0$,

$$\mathbb{P}\left(\sup_{0\le u\le 1} \|W(n+u) - W(n)\|_\infty > x \right) \le 4d \exp(-x^2/2).$$

Taking $x_n = 2\sqrt{\log n}$ makes the probabilities summable. Borel–Cantelli proves the first display, and norm equivalence plus $\sqrt{\log n/n} \to 0$ proves the second. □

Proof of Corollary 2. Work first on the joint copy supplied by Theorem 1. We suppress daggers in this proof. The covariance consistency event and the coupling event both have probability one. With $Q_T = \widetilde{\Sigma}_T^{-1/2}\Sigma^{1/2}$, matrix continuity gives $Q_T \to I_d$, also in the induced q-norm, and the strong approximation gives

$$\frac{T}{\sqrt{v_T}} \widetilde{\Sigma}_T^{-1/2}(\bar{y}_T - x^\star) = Q_T \frac{W(v_T)}{\sqrt{v_T}} + o_{\mathrm{a.s.}}(1). \tag{43}$$

Lemma 9 gives $\|W(v_T) - W(\lfloor v_T \rfloor)\|_q / \sqrt{v_T} = o_{\mathrm{a.s.}}(1)$. Since $v_T \to \infty$, $n_T = \lfloor v_T \rfloor$ eventually.

Let $\mathcal{E}_\alpha = \{\|W(n)\|_q \le u^{(q)}_{\alpha,d}(n)\sqrt{n} \text{ for all integers } n \ge 1\}$. Definition 2 gives $\mathbb{P}(\mathcal{E}_\alpha) \ge 1 - \alpha$.

Every valid deterministic boundary also satisfies $u^{(q)}_{\alpha,d}(n) \ge c_{\alpha,d,q} > 0$: each marginal $W(n)/\sqrt{n}$ is standard normal, and its norm must fall below that boundary with probability at least $1 - \alpha$.

On $\mathcal{E}_\alpha$, (43) and $\sqrt{n_T/v_T} \le 1$ imply

$$\frac{T}{\sqrt{v_T}} \left\| \widetilde{\Sigma}_T^{-1/2}(\bar{y}_T - x^\star) \right\|_q \le (1 + e_T) u^{(q)}_{\alpha,d}(n_T) + r_T, \qquad e_T, r_T \to 0 \quad \text{a.s.}$$

The fixed $\delta > 0$ and the positive lower bound on u ensure that the right-hand side is at most $(1+\delta)u^{(q)}_{\alpha,d}(n_T)$ for every sufficiently large T. Consequently,

$$\mathcal{E}_\alpha \subseteq \bigcup_{N\ge 1} \bigcap_{T\ge N} \{x^\star \in \mathcal{C}_T^{(\delta)}\} \quad \text{up to a null set.}$$

Continuity of probability for the increasing union gives the claimed lower bound. Equality in law transfers it to the original sequence of regions. □

Proof of Corollary 3. Let $S_n = \sum_{i=1}^n Z_i$. For $\rho > 0$, Gaussian mixing of the exponential martingale gives

$$M_n(\rho) = \left(\frac{\rho}{\rho + n}\right)^{d/2} \exp\left\{ \frac{\|S_n\|_2^2}{2(\rho + n)} \right\},$$

a nonnegative martingale with $M_0 = 1$. Ville's inequality implies that, with probability at least $1 - \alpha$, for every $n \ge 1$,

$$\frac{\|S_n\|_2^2}{n} \le d\left(1 + \frac{\rho}{n}\right) \log\left(\frac{1 + n/\rho}{\alpha^{2/d}} \right).$$

Set $\rho = n_0/\lambda^*_{\alpha,d}$. The choice in (23) solves

$$\log(1+\lambda) + \frac{2}{d}\log(1/\alpha) = \lambda,$$

so the continuous-clock boundary multiplier $u(n)$ is minimized at $n = n_0$. This proves (22); other choices of $\rho > 0$ remain valid.

For (24), apply the two-sided scalar finite-LIL boundary of Howard et al. Howard et al. (2021) at error level α/d to each coordinate and use a union bound over d coordinates. □

A.5 Low-dimensional functionals

For larger models, one option is to fix or separately train the representation and update an identifiable, smooth, low-dimensional prediction head federatively, subject to the stated regularity and stability conditions. Another is to target a smooth functional $\psi(x) \in \mathbb{R}^r$. If $J_\psi = \nabla\psi(x^\star)$ has full row rank and the functional remainder is negligible at the coupling scale, then

$$\psi(\bar{y}_T) - \psi(x^\star) = J_\psi(\bar{y}_T - x^\star) + o_{\mathrm{a.s.}}(\sqrt{v_T}/T),$$

with covariance $J_\psi \Sigma J_\psi^\top$. A locally Lipschitz Jacobian is sufficient under the polynomial schedules here because the quadratic functional remainder is

$$O_{\mathrm{a.s.}}\left(\frac{v_T \log\log v_T}{T^2}\right) = o_{\mathrm{a.s.}}(\sqrt{v_T}/T).$$

Examples include fixed-covariate logits, calibrated risk contrasts, and a small set of head coefficients.

A.6 Participation covariance

Under full participation, $\sum_k p_k \nabla f_k(x^\star) = 0$ at every round, a cancellation used in Local-SGD and distributed-inference analyses (Li et al., 2022; Jordan et al., 2019; Duan et al., 2022). Client subsampling preserves this cancellation only in expectation and introduces additional selection noise.

Suppose client k is selected independently once per round with probability q_k, and the server uses Horvitz–Thompson weighting. At $x^\star$, the selected round gradient has covariance

$$\sum_{k=1}^{K} p_k^2 \left\{ \frac{S_k}{q_k E_m} + \left(\frac{1}{q_k} - 1\right) g_k^\star g_k^{\star\top} \right\}. \tag{44}$$

The first component accumulates on the $v_T = \sum_m E_m^{-1}$ clock. The selection–heterogeneity component accumulates on the T clock. Under irregular E_m, their cumulative covariance is generally

$$v_T A + TB,$$

where $A = \sum_k p_k^2 S_k/q_k$ and $B = \sum_k p_k^2(q_k^{-1} - 1) g_k^\star g_k^{\star\top}$. These matrices are generally nonproportional, requiring a matrix-valued covariance process rather than one fixed covariance matrix on a scalar clock. When all clients share the minimizer, $g_k^\star = 0$, the selection term vanishes and the noise covariance retains the v_T clock with S replaced by $\sum_k p_k^2 S_k/q_k$. Fixed-size sampling without replacement adds cross-client inclusion-covariance terms. Extending the strong approximation also requires bounds for local drift and noise freezing under the sampling scheme.

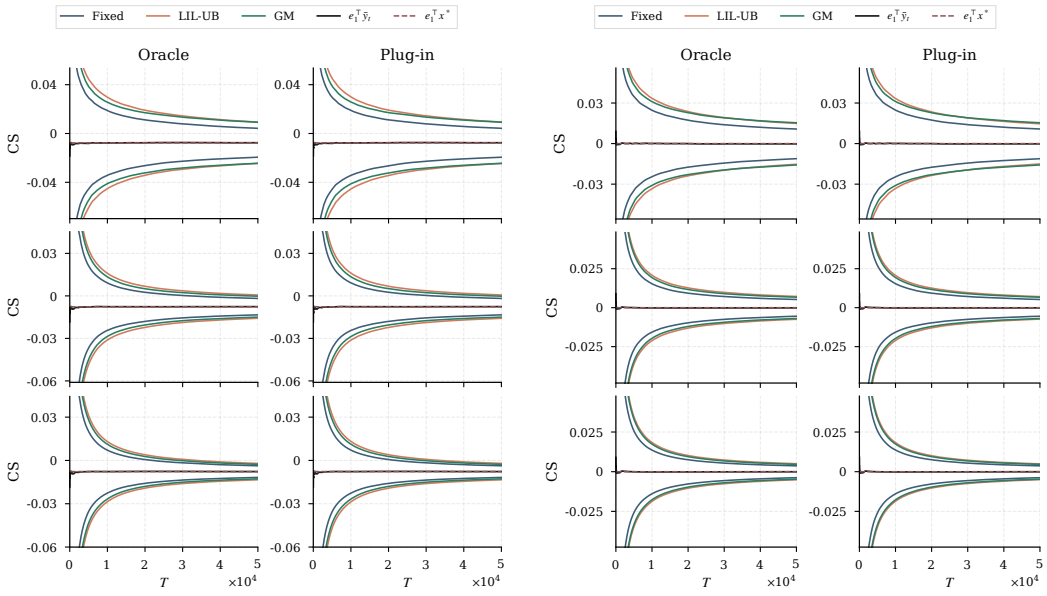
Appendix B.5 and Table B.3 report the without-replacement experiment.

B ADDITIONAL EXPERIMENTS AND DISCUSSIONS

In this section, we report additional experimental results that complement Section 3. Unless otherwise specified, the data-generating models, initialization, step sizes, and warm-up procedure follow Section 3.1. For the additional simulations, pathwise failure is the percentage of replications with at least one noncoverage between the stated monitoring start and the terminal release. Average marginal half-widths are averaged over coordinates and replications. Brackets in the tables give 95% Wilson Monte Carlo intervals for the reported rates.

B.1 OTHER RESULTS IN SECTION 3

Figure B.1 shows Monte Carlo average first-coordinate trajectories for the linear and logistic models. The averages approach the corresponding target values as the bands contract over communication rounds. The average LIL-UB and GM bands contain the target lines, and the oracle and plug-in trajectories are visually close.



(a) Linear regression.

(b) Logistic regression.

Figure B.1: Monte Carlo average first-coordinate trajectories. Each panel shows the Monte Carlo average of $e_1^\top \bar{y}_t$, the corresponding average truth line $e_1^\top x^\star$, and average marginal bands for the fixed-time Gaussian benchmark, LIL-UB, and GM. In each subfigure, Rows 1–3 correspond to the C_1, C_5, and *Log* communication schedules, respectively, and the two columns compare oracle and plug-in covariance.

Figures B.2–B.4 report the logistic and Poisson simulation results. As in the main experiments, LIL-UB and GM have lower empirical time-uniform failure rates than the narrower fixed-time Gaussian benchmark, and the plug-in covariance closely tracks the oracle covariance.

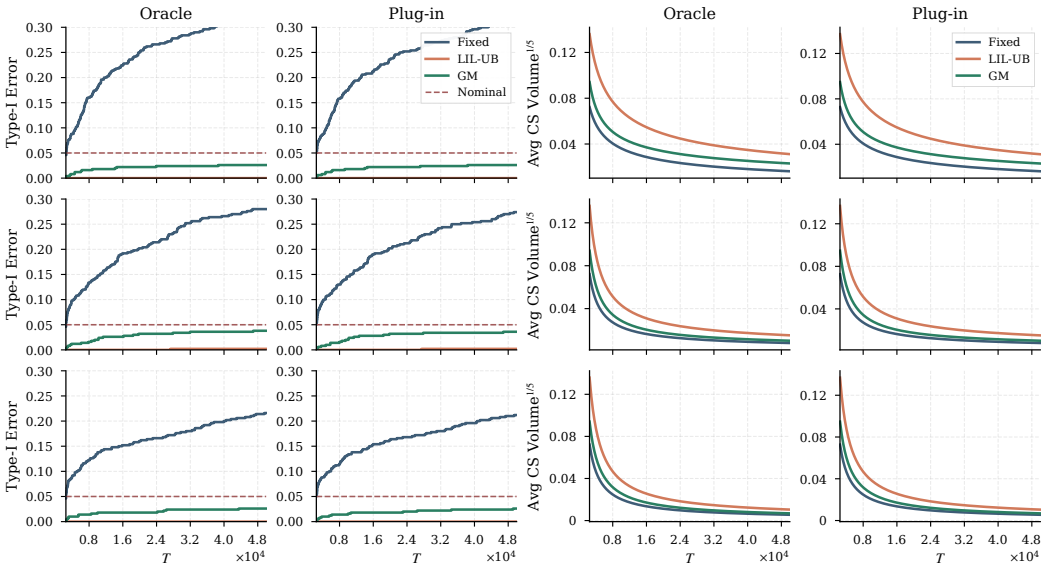


Figure B.2: Logistic regression: time-uniform type-I error and average confidence-set volume. Rows 1–3 correspond to the C_1, C_5, and *Log* communication schedules, respectively. The first two columns report time-uniform type-I error under oracle and plug-in covariance, and the last two columns report the corresponding average volume of the full $d = 5$ confidence sets. The dashed horizontal line marks the nominal 0.05 level.

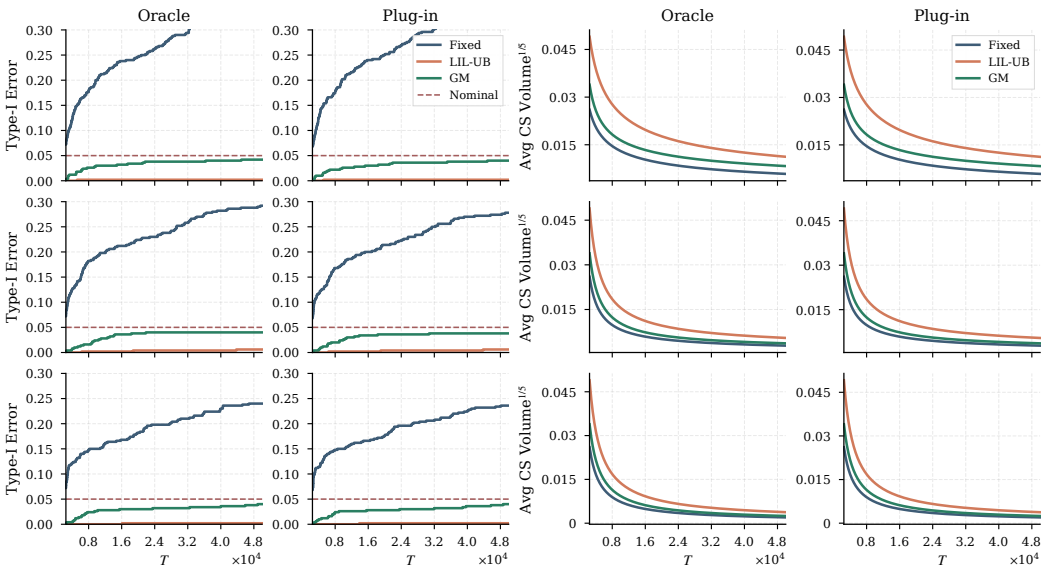


Figure B.3: Poisson regression: time-uniform type-I error and average confidence-set volume. Rows 1–3 correspond to the C_1, C_5, and *Log* communication schedules, respectively. The first two columns report time-uniform type-I error under oracle and plug-in covariance, and the last two columns report the corresponding average volume of the full $d = 5$ confidence sets. The dashed horizontal line marks the nominal 0.05 level.

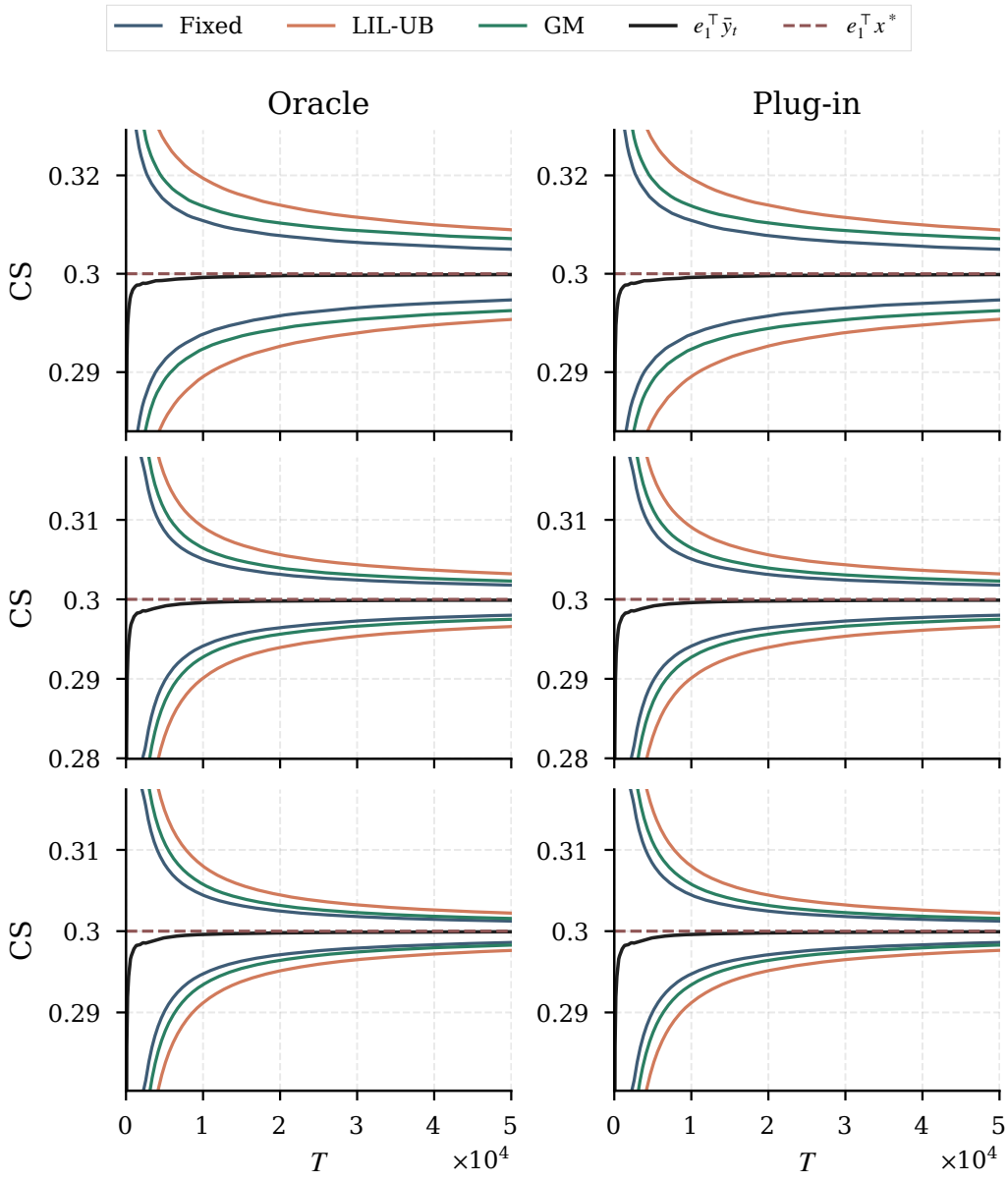


Figure B.4: Poisson regression: Monte Carlo average first-coordinate trajectory. Each panel shows the Monte Carlo average of $e_1^\top \bar{y}_t$, the corresponding average truth line $e_1^\top x^\star$, and average marginal bands for the fixed-time Gaussian benchmark, LIL-UB, and GM. Rows 1–3 correspond to the C_1, C_5, and *Log* communication schedules, respectively, and the two columns compare oracle and plug-in covariance.

B.2 REAL DATA

We use the 2022 Behavioral Risk Factor Surveillance System (BRFSS) public-use data[1] (Hacker et al., 2024), a state-based survey of U.S. adults on health status and risk behaviors. The raw data contain 445,132 records. The response is self-reported diagnosed diabetes, and the six covariates are standardized age, standardized BMI, sex, current smoking status, physical inactivity, and fair or poor self-rated health. We use the geographic partition to construct a federated-style experiment and inspect coefficient uncertainty across communication rounds.

After standardizing continuous variables and removing records with missing or invalid values, we retain $N = 358{,}927$ records. To use the dataset's geographic structure, we partition records by the respondent's state or participating jurisdiction, as identified by the BRFSS state identifier. Clients with fewer than 3000 usable records are removed, leaving $K = 50$ clients. Client sizes range from 3214 to 21,471, and we set $p_k = n_k/N$.

We fit a logistic regression model with loss $\ell(x; \xi_i) = \log(1 + \exp(a_i^\top x)) - b_i a_i^\top x$, where $\xi_i = (a_i, b_i)$, and $a_i \in \mathbb{R}^7$ contains an intercept and the six covariates described above. We construct plug-in time-uniform confidence sequences for the full coefficient vector and report their coordinate projections for the six covariate coefficients. Both the LIL-UB radius (24) and the GM radius (22) with $n_0 = 1$ are used. We compare two communication schedules. The first uses $E_m \equiv 1$ and $T = 3214$ releases, while the second uses $E_m \equiv 5$ and $T = 642$ releases; hence the two schedules have nearly the same t_T values. We initialize at zero and use $\gamma_m = 16(m + 10)^{-0.6}$ with local step size $\eta_m = \gamma_m / E_m$. For reference, we also compute the centralized estimator on the same retained records and use it as a benchmark. The results are summarized in Figures B.5 and B.6.

The LIL-UB and GM bands contain the centralized benchmark along the displayed communication path; the fixed-time Gaussian bands are narrower. Estimates for age, BMI, and fair or poor self-rated health remain positive and separated from zero across the monitored rounds, indicating stable positive associations with diagnosed diabetes in the fitted logistic model. The two schedules, $E_m = 1$ and $E_m = 5$, produce comparable coefficient trajectories and band widths.

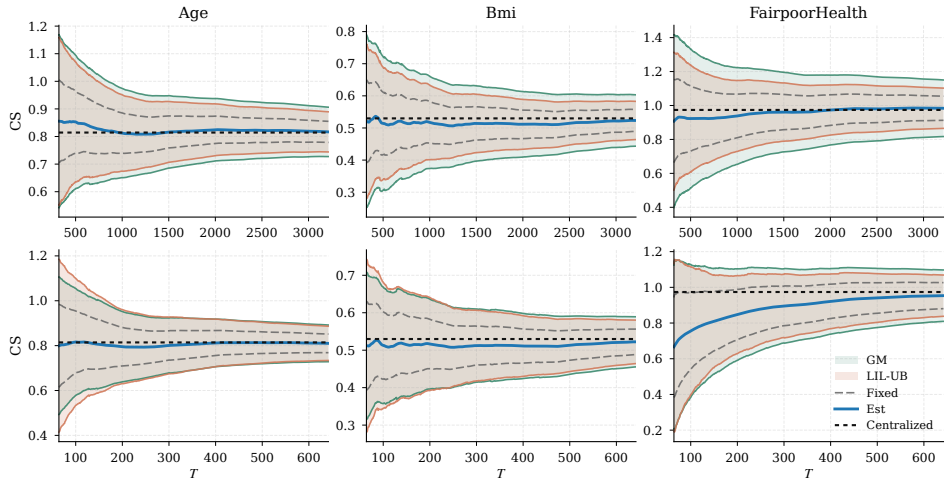


Figure B.5: BRFSS coefficient trajectories for age, BMI, and fair or poor self-rated health. Each panel shows $\bar{y}_T$ (Est), the centralized benchmark, and marginal bands for the fixed-time Gaussian benchmark, LIL-UB, and GM. The upper and lower rows correspond to the $E_m \equiv 1$ and $E_m \equiv 5$ communication schedules, respectively.

Figure B.6 reports the remaining BRFSS coefficient trajectories for sex, current smoking status, and physical inactivity, complementing Figure B.5. Under both communication schedules, the confidence sequences track the centralized benchmark, and the results are comparable between $E_m \equiv 1$ and $E_m \equiv 5$.

[1]Data are available at `https://www.cdc.gov/brfss/annual_data/annual_2022.html`.

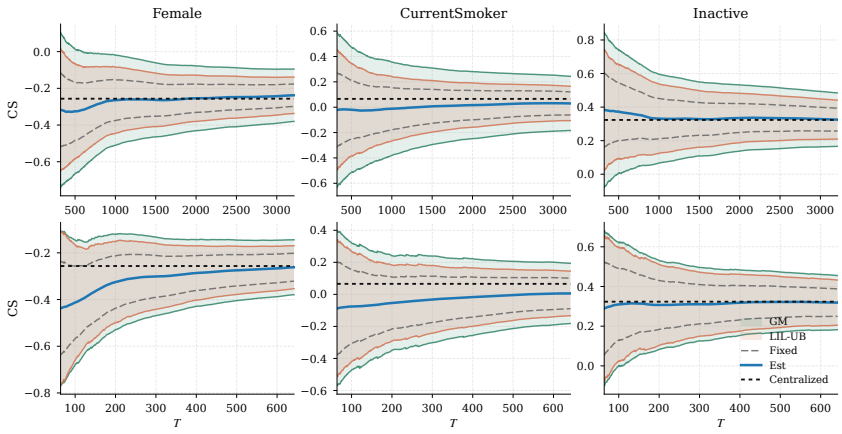


Figure B.6: BRFSS coefficient trajectories for sex, current smoking status, and physical inactivity. Each panel shows $\bar{y}_T$ (Est), the centralized benchmark, and marginal bands for the fixed-time Gaussian benchmark, LIL-UB, and GM. The upper and lower rows correspond to the $E_m \equiv 1$ and $E_m \equiv 5$ communication schedules, respectively.

B.3 Sensitivity to the GM design point and monitoring start

We examine how the GM design point and monitoring start affect coverage and interval width. We use the Linear C_5 setting with $K = 10$, $d = 5$, $T = 50{,}000$, 500 replications, and plug-in covariance. The design points are $n_0 \in \{0.5v_{\mathrm{mid}}, v_{\mathrm{mid}}, 2v_{\mathrm{mid}}, 20{,}000\}$, where $v_{\mathrm{mid}} = \sqrt{v_{0.05T} v_T} \approx 5{,}477$, and are evaluated on common Monte Carlo paths. We consider monitoring starts at 5%, 10%, and 20% of the run and report average marginal half-widths at v_{mid} and the terminal release.

Panel A. Pathwise failure (%)

Design point	n_0	5% start	10% start	20% start
$0.5v_{\mathrm{mid}}$	2,739	1.6 [0.8, 3.1]	1.0 [0.4, 2.3]	1.0 [0.4, 2.3]
v_{mid}	5,477	2.0 [1.1, 3.6]	1.8 [0.9, 3.4]	1.8 [0.9, 3.4]
$2v_{\mathrm{mid}}$	10,955	1.4 [0.7, 2.9]	1.2 [0.6, 2.6]	1.2 [0.6, 2.6]
Main setting	20,000	0.6 [0.2, 1.7]	0.6 [0.2, 1.7]	0.6 [0.2, 1.7]

Panel B. Average half-widths and the GM multiplier minimum

Design point	n_0	Half-width ($\times 10^{-3}$)		$\widehat{n}_{\min}/n_0$
		At v_{mid}	At v_T	
$0.5v_{\mathrm{mid}}$	2,739	14.366	7.809	1.000
v_{mid}	5,477	14.070	7.426	1.000
$2v_{\mathrm{mid}}$	10,955	14.480	7.239	1.000
Main setting	20,000	15.705	7.346	0.600

Table B.1: GM design-point and monitoring-start sensitivity in the Linear C_5 setting ($K = 10$, $d = 5$, $T = 50{,}000$, 500 replications, plug-in covariance). Half-widths are reported in units of 10^{-3}. The quantity $\widehat{n}_{\min}$ is the observed variance-clock value minimizing the GM multiplier.

Table B.1 shows that the main effect of n_0 is to redistribute precision over the monitoring window. The log-midpoint choice gives the smallest half-width at v_{mid}, whereas larger choices favor later releases. For the three design points within the observed variance-clock range, $\widehat{n}_{\min}/n_0 = 1$ to the reported precision, consistent with Corollary 3. Thus n_0 can be chosen near the variance clock of a target release, or at the geometric midpoint of a pre-specified monitoring window.

Pathwise failure from the 5% start ranges from 0.6% to 2.0%. Moving the start to 10% or 20% changes these rates by at most 0.6 percentage points. Later starts evaluate nested, shorter monitoring windows.

B.4 Client count and heterogeneity

We investigate the effects of client count and heterogeneity under two observation budgets. We use the Linear C_5 setting with $d = 5$, equal client weights, oracle covariance, and 300 replications. The local targets are

$$x_k^\star = x^\star + hu_k, \qquad K^{-1}\sum_{k=1}^{K} u_k = 0, \qquad K^{-1}\sum_{k=1}^{K} \|u_k\|^2 = 1,$$

with $K \in \{5, 10, 50\}$ and $h \in \{0, 1\}$. This construction gives $\zeta_\star = h$ and $\mathrm{tr}(\Sigma) = (5 + 6h^2)/K$. We set $x^\star = (0, 1/4, 1/2, 3/4, 1)^\top$. In each replication, we center independent $\mathcal{N}(0, I_5)$ vectors across clients and rescale them by a common factor to satisfy the normalization above. The resulting directions u_k are held fixed throughout the run. We use the GM boundary with $\alpha = 0.05$ and the exact normalization in (20). For each combination of budget and K, we fix $n_0 = \sqrt{v_{T_{\mathrm{start}}} v_T}$ in advance, where $T_{\mathrm{start}} = \lceil 0.10T \rceil + 1$ is the first monitored round.

For the fixed per-client budget, we use $T = 20{,}000$ releases, so the total observation count increases with K. For the fixed total-system budget, we use 0.96 million observations and set $T = 40{,}000$, $20{,}000$, and $4{,}000$ for $K = 5$, 10, and 50, respectively. Pathwise failure is evaluated after the 10% monitoring start; half-width and RMSE are evaluated at the terminal release.

h	K	T	N_{sys} ($\times 10^6$)	Half-width ($\times 10^{-3}$)	RMSE ($\times 10^{-3}$)	$\mathrm{tr}(\Sigma)$	Failure (%)
Panel A. Fixed per-client budget							
0	5	20,000	0.48	6.552	3.442	1.000	5.7 [3.6, 8.9]
0	10	20,000	0.96	4.633	2.575	0.500	4.0 [2.3, 6.9]
0	50	20,000	4.80	2.072	1.145	0.100	3.0 [1.6, 5.6]
1	5	20,000	0.48	9.714	5.238	2.200	5.0 [3.1, 8.1]
1	10	20,000	0.96	6.870	3.784	1.100	4.0 [2.3, 6.9]
1	50	20,000	4.80	3.073	1.659	0.220	3.7 [2.1, 6.4]
Panel B. Fixed total-system budget							
0	5	40,000	0.96	4.633	2.506	1.000	4.0 [2.3, 6.9]
0	10	20,000	0.96	4.633	2.460	0.500	4.7 [2.8, 7.7]
0	50	4,000	0.96	4.633	2.666	0.100	11.0 [7.9, 15.0]
1	5	40,000	0.96	6.869	3.666	2.200	2.7 [1.4, 5.2]
1	10	20,000	0.96	6.870	3.744	1.100	8.0 [5.4, 11.6]
1	50	4,000	0.96	6.871	3.912	0.220	7.3 [4.9, 10.9]

Table B.2: Client count and heterogeneity in the Linear C_5 setting ($d = 5$, oracle covariance, 300 replications). N_{sys} is the total observation count across clients. Failure denotes at least one noncoverage after the 10% monitoring start. Half-width and RMSE are measured at the terminal release.

The results are reported in Table B.2. Under the fixed per-client budget, increasing K from 5 to 50 reduces the homogeneous half-width from 6.552×10^{-3} to 2.072×10^{-3}, approximately a factor of $\sqrt{10}$, and lowers RMSE from 3.442×10^{-3} to 1.145×10^{-3}. The heterogeneous setting shows the same $K^{-1/2}$ pattern. Under the fixed total-system budget, however, half-widths remain nearly unchanged across K: about 4.633×10^{-3} for $h = 0$ and 6.87×10^{-3} for $h = 1$.

Increasing h from 0 to 1 multiplies $\mathrm{tr}(\Sigma)$ by $11/5 = 2.2$ and the half-width by approximately $\sqrt{2.2} = 1.48$. Thus both the total observation budget and the client-specific covariance determine precision. The fixed-total $K = 50, h = 0$ setting has 11.0% pathwise failure with 4,000 releases, indicating that the number of communication rounds also affects coverage at a fixed observation budget.

B.5 PARTIAL CLIENT PARTICIPATION

We further examine partial client participation in the heterogeneous Linear C_5 setting with $K = 10$, $d = 5$, $T = 50{,}000$, and 500 paired replications. At each round, $m = qK$ clients are sampled uniformly without replacement, where $q \in \{1, 0.8, 0.6, 0.5\}$. The participation levels share the same local targets, data streams, initialization, and client permutations. Oracle covariance components are used to isolate the effect of client sampling. Selected clients start from the current synchronized parameter, and the server averages their updated parameters with weights $1/(qK)$. The local updates follow Section 3.1, and both GM calibrations use the fixed design point $n_0 = 20{,}000$ and $\alpha = 0.05$.

For $\delta_k = x^\star - x_k^\star$, define

$$A_q = \frac{1}{mK}\sum_{k=1}^{K}\left\{(1+\|\delta_k\|_2^2)I_d + \delta_k\delta_k^\top\right\} = \frac{A_1}{q}, \qquad B_q = \frac{1-m/K}{m(K-1)}\sum_{k=1}^{K}\delta_k\delta_k^\top.$$

The leading terminal covariance is $T^{-2}(v_T A_q + TB_q)$: A_q accounts for data noise and B_q for client-selection noise. We compare the unchanged full-participation calibration with a q-adjusted GM calibration using $A_q = A_1/q$. For terminal fixed-time inference, we also use the participation-aware oracle covariance $T^{-2}(v_T A_q + TB_q)$.

Panel A. Empirical noncoverage (%)

q	Active	GM pathwise failure		Terminal fixed-time noncoverage	
		Unchanged	q-adjusted	Unchanged	PP oracle
1.0	10	1.2 [0.6, 2.6]	1.2 [0.6, 2.6]	5.4 [3.7, 7.7]	5.4 [3.7, 7.7]
0.8	8	6.8 [4.9, 9.4]	1.8 [0.9, 3.4]	13.0 [10.3, 16.2]	2.4 [1.4, 4.1]
0.6	6	29.6 [25.8, 33.7]	2.8 [1.7, 4.6]	33.0 [29.0, 37.2]	3.4 [2.1, 5.4]
0.5	5	51.2 [46.8, 55.6]	4.6 [3.1, 6.8]	49.8 [45.4, 54.2]	4.6 [3.1, 6.8]

Panel B. Estimation error and client-selection variance

q	Active	RMSE ($\times 10^{-3}$)	Selection-variance share (%)
1.0	10	3.935	0.0
0.8	8	4.585	11.5
0.6	6	5.598	20.6
0.5	5	6.348	24.5

Table B.3: Partial client participation in the heterogeneous Linear C_5 setting ($K = 10$, $d = 5$, $T = 50{,}000$, 500 paired replications). Pathwise failure is measured after the 5% monitoring start. "Unchanged" retains the full-participation calibration; q-adjusted uses A_1/q; PP oracle uses the participation-aware covariance for terminal fixed-time inference. The selection-variance share is the percentage of the leading variance attributable to TB_q.

Table B.3 shows that, as q decreases from 1 to 0.5, RMSE increases from 3.935×10^{-3} to 6.348×10^{-3}. The effect on coverage is larger: retaining the full-participation calibration raises GM pathwise failure from 1.2% to 51.2% and terminal noncoverage from 5.4% to 49.8%. For the partial-participation settings, the q adjustment reduces pathwise failure to 1.8%–4.6%, while the terminal participation-aware oracle gives 2.4%–4.6% noncoverage.

At $q = 0.5$, client-selection noise contributes 24.5% of the leading variance. These results show the importance of accounting for client sampling in covariance calibration. The q-adjusted comparison is empirical, and a time-uniform extension requires the selection covariance described in Appendix A.6.

B.6 DIMENSION AND COVARIANCE ESTIMATION

We assess covariance-estimation accuracy and computational cost as dimension increases. We use the Logistic C_5 setting with $K = 10$, $T = 50{,}000$, $n_0 = 20{,}000$, and 500 replications. The five-dimensional target is extended to $d = 20$ and $d = 30$ by zero-padding, so the signal norm remains

fixed. Oracle and sample-reuse plug-in covariances are evaluated on the same training paths. The plug-in estimator follows the sample-reuse scheme in Section 3.1. At each round, we compute the average client Hessian and the outer product of the average client gradient. Their averages over all rounds, including warm-up, form the two components of the sandwich estimator. Proposition 1 establishes consistency for the fresh-sample construction.

Panel A. Empirical noncoverage (%)

d	GM pathwise failure		LIL-UB failure	Terminal GM noncoverage	
	Oracle	Plug-in	Oracle	Oracle	Plug-in
5	1.6	1.6	0.0	0.2	0.2
20	3.0	2.8	0.0	0.6	0.6
30	4.4	4.0	0.6	1.4	1.4

Panel B. Estimation accuracy and computational cost

d	RMSE ratio	GM half-width ratio	Covariance error (%)	Cost ratio
5	0.993	1.001	1.13	$1.82\times$
20	1.071	1.003	2.24	$5.17\times$
30	1.094	1.004	2.74	$8.19\times$

Table B.4: Dimension sensitivity and covariance estimation in the Logistic C_5 setting. Pathwise failure is measured after the 5% monitoring start; plug-in denotes the sample-reuse estimator. RMSE ratio is empirical RMSE divided by its first-order approximation. The half-width ratio is plug-in/oracle at the terminal release. Covariance error is the median relative Frobenius error. Cost ratio compares covariance estimation with training only on the same machine.

The results in Table B.4 show that the plug-in covariance closely tracks the oracle covariance. GM pathwise failure rates differ by at most 0.4 percentage points, and terminal noncoverage rates are identical. Plug-in half-widths differ from the oracle by 0.1%–0.4%, with median relative covariance errors of 1.13%–2.74%. The computational cost nevertheless increases with dimension: covariance estimation cost rises from $1.82\times$ the training-only cost at $d = 5$ to $8.19\times$ at $d = 30$.

The oracle results also show the effect of dimension on coverage. From $d = 5$ to $d = 30$, GM pathwise failure increases from 1.6% to 4.4%, and terminal GM noncoverage from 0.2% to 1.4%. At $d = 30$, the oracle GM failure rate is $22/500$, with a 95% Wilson interval of 2.9%–6.6%. LIL-UB is more conservative, with pathwise failure no larger than 0.6% over these settings.

B.7 Communication schedules and exact normalization

We examine how schedule normalization affects the resulting confidence regions. We use the Linear setting with $K = 10$, $d = 5$, $T_{\max} = 50{,}000$, oracle covariance, and 500 replications. In addition to C_1, C_5, and Log, we consider an irregular schedule, AltBlock, that alternates $E_m = 1$ and $E_m = 9$ in paired growing blocks. After a 2,500-round warm-up with $E_m = 1$, the jth pair contains $5j^2$ rounds with $E_m = 1$ followed by $5j^2$ rounds with $E_m = 9$, for $j = 1, 2, \ldots$. The final block is truncated at $T_{\max}$.

For each schedule, we set $T_{\text{start}} = \lceil 0.05 T_{\max} \rceil + 1 = 2{,}501$ and choose

$$n_0 = \sqrt{v_{T_{\text{start}}} v_{T_{\max}}} \tag{45}$$

in advance. This design point is held fixed across normalization variants for that schedule. We compare the exact factor $\nu_T = t_T v_T / T^2$ with a fixed reference ν_{ref} and with $0.8\nu_{\text{ref}}$ and $1.2\nu_{\text{ref}}$. The reference is 1 for C_1, C_5, and Log, and $25/9$ for AltBlock. The exact radius scale is $\sqrt{\nu_T / t_T} = \sqrt{v_T}/T$, so replacing ν_T by ν_{ref} multiplies the half-width by $\sqrt{\nu_{\text{ref}}/\nu_T}$.

Panel A. Schedule factors and terminal half-width

Schedule	ν_{ref}	Relative factor error (%)		Half-width ratio
		Maximum	Terminal	
C_1	1.000	0.0	0.0	1.000
C_5	1.000	80.0	15.2	0.932
Log	1.000	262.0	62.2	0.785
AltBlock	2.778	64.0	0.4	1.002

Panel B. Pathwise failure (%)

Schedule	Exact ν_T	ν_{ref}	$0.8\nu_{\mathrm{ref}}$	$1.2\nu_{\mathrm{ref}}$
C_1	4.2 [2.8, 6.3]	4.2 [2.8, 6.3]	10.2 [7.8, 13.2]	1.4 [0.7, 2.9]
C_5	2.6 [1.5, 4.4]	22.0 [18.6, 25.8]	36.2 [32.1, 40.5]	10.8 [8.4, 13.8]
Log	2.6 [1.5, 4.4]	57.8 [53.4, 62.1]	74.4 [70.4, 78.0]	46.4 [42.1, 50.8]
AltBlock	2.0 [1.1, 3.6]	1.2 [0.6, 2.6]	6.8 [4.9, 9.4]	1.0 [0.4, 2.3]

Table B.5: Schedule-normalization sensitivity in the Linear setting ($K = 10$, $d = 5$, $T_{\max} = 50{,}000$, oracle covariance, 500 replications). Each schedule uses the design point in (45). Relative factor error is $|\nu_T/\nu_{\mathrm{ref}} - 1|$, summarized by its maximum over the monitored window and terminal value. The half-width ratio compares the reference and exact terminal calibrations. Pathwise failure is evaluated after the 5% monitoring start. The displayed 2.778 is the rounded AltBlock reference $25/9$.

Table B.5 shows that replacing the exact normalization has a larger effect on coverage than changing n_0 over the grid in Table B.1. For C_1, $\nu_T = 1$ exactly and the two calibrations coincide. For C_5, using the reference value makes the terminal half-width 6.8% smaller and raises pathwise failure from 2.6% to 22.0%. For Log, the half-width is 21.5% smaller and failure increases from 2.6% to 57.8%. The $0.8\nu_{\mathrm{ref}}$ and $1.2\nu_{\mathrm{ref}}$ comparisons show the same pattern: smaller radius scales give narrower regions and more failures.

For AltBlock, the schedule-factor discrepancy is larger earlier in the monitored window, but its terminal value is within 0.4% of $25/9$. The exact and reference-based terminal half-widths are therefore nearly identical, and their failure rates are close. These results favor retaining ν_T, which is available directly from the communication schedule.