# LOCAL RIEMANNIAN STOCHASTIC GRADIENT DESCENT: LEARNING AND UNCERTAINTY QUANTIFICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study statistical inference with infrequent communication for the full-participation Riemannian Federated Averaging Gradient Streams (RFedAGS) recursion with growing communication intervals, where client drift and geometric error must be controlled at the scale of all local observations. A moment-sensitive synchronized-state bound makes these errors negligible and yields a central limit theorem (CLT) for a computable fixed-anchor average in a regular local basin. The covariance separates the objective Hessian, client noise and communication schedule. With a fixed burn-in, admissible slowly varying intervals preserve first-order precision while communication frequency vanishes. Fresh pooled scores and exact or sample-based Hessians consistently estimate this covariance, giving feasible Wald regions. Gaussian and sphere experiments at matched total oracle budgets show that a prespecified half-tail attains near-nominal coverage with substantially lower communication cost than a fixed-interval baseline, at the cost of a larger uncertainty scale. A four-site principal-subspace application illustrates the construction.

## 1 INTRODUCTION

Distributed principal component analysis estimates a subspace from data held on different machines; Huang & Pan (2020) reduce communication through local Riemannian updates, controlling drift from the shared iterate. We study sampling uncertainty for manifold-valued estimates as the number of local updates between communications grows.

In Euclidean problems, local stochastic gradient descent (Local SGD), or federated averaging (FedAvg) under full participation, replaces frequent communication by multiple client updates (McMahan et al., 2017; Stich, 2019); heterogeneous client objectives then create local drift, motivating corrections such as control variates (Karimireddy et al., 2020). For inference, Li et al. (2022) establish limits and feasible uncertainty estimation for averages of synchronized Local SGD iterates, with a covariance that records the communication schedule and lets suitable intervals diverge without first-order variance inflation. This builds on iterate averaging in stochastic approximation (Robbins & Monro, 1951; Ruppert, 1988; Polyak & Juditsky, 1992) and complements trajectory-based covariance estimation and online bootstrap (Chen et al., 2020; Fang et al., 2018).

Riemannian stochastic gradient descent (RSGD) optimizes manifold-valued parameters (Bonnabel, 2013; Absil et al., 2008; Boumal, 2023), and averaged RSGD has a single-stream Polyak–Ruppert limit (Tripuraneni et al., 2018). Distributed subspace aggregation (Fan et al., 2019; Huang & Pan, 2020), surrogate optimization (Jordan et al., 2019; Saparbayeva et al., 2018), and federated variance reduction (Li & Ma, 2023) address related computational questions. RFedAGS transports and aggregates each client's local gradient stream (Huang et al., 2026); its theory covers optimization, heterogeneity and partial participation, whereas we study the full-participation recursion's *sampling uncertainty*, including covariance estimation after synchronization becomes infrequent.

Growing communication intervals couple geometric error with client drift. Each gradient belongs to its evaluation point's tangent space: RFedAGS transports it to the server, and the analysis then transports it to the target (Figure 1). This two-leg path and the retraction mapping a tangent step onto the manifold introduce errors that accumulate with local work; together with client drift and



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.

the boundary terms left by averaging, they must vanish on the scale of all local observations, which grows faster than the server-round count.

Our key step retains the variance reduction from within-round averaging when bounding the synchronized state: under a $(2+\delta)$ oracle moment the mean-square error is $O\{\eta_m + \gamma_m^{1+\min(\delta,1)}\}$, where $\eta_m$ is the local step and $\gamma_m = E_m\eta_m$ the effective step over $E_m$ local updates, finer than the coarse $O(\gamma_m)$ scale when intervals grow. This removes unnecessary restrictions on the averaging boundary terms: with a third moment, every power interval $E_m \asymp m^\beta$, $0 \le \beta < 1$, admits steps $\gamma_m \asymp m^{-\alpha}$ with $(1+\beta)/2 < \alpha < 1$ under the stated local geometric assumptions, and lower moments add one explicit schedule condition.

The limit is attained by a fixed-anchor tangent average with $O(d)$ averaging memory and no retraction-convexity requirement; a recursive geometric average has the same law under an additional local condition. For feasible inference, fresh client scores are pooled *before* their outer products, so heterogeneous mean gradients cancel at the target, and transported Hessians and scores then estimate the covariance in the reported estimate's tangent space; the additional storage and second-order communication costs of the full uncertainty calculation are counted explicitly.

At matched optimization-plus-inference budgets, Gaussian, sphere and rotation experiments show near-95% coverage with about two-thirds lower communication cost for a prespecified half-tail than for eight local steps per round, with larger uncertainty and estimation error, whereas insufficient removal of early iterates can cause severe undercoverage. Sections 2 and 3 develop the method and guarantees; Sections 4 and 5 give the comparisons and four-site subspace application.

## 2 METHODOLOGY

### 2.1 THE DISTRIBUTED TARGET AND GRADIENT STREAM

Let $(\mathcal{M}, g)$ be a connected, finite-dimensional Riemannian manifold of dimension $d$. Client $k \in \{1, \dots, K\}$ has a population objective $f_k : \mathcal{M} \to \mathbb{R}$ and weight $p_k > 0$, with $\sum_k p_k = 1$. The global objective is $f(x) = \sum_{k=1}^K p_k f_k(x)$ and the target is a local minimizer $x_\star \in \arg\min^{\mathrm{loc}}_{x\in\mathcal{M}} f(x)$, so that $f(x_\star) \le f(x)$ for all $x$ in some neighborhood of $x_\star$; clients may have different local minimizers, and stationarity at $x_\star$ concerns their weighted objective. For an intrinsic mean, $f_k$ is expected squared geodesic distance; for a principal subspace, a quadratic projection loss. In each case the gradient describes how the local objective changes along tangent directions and the Hessian its local sensitivity, which motivates an inference construction using gradient and Hessian messages in local tangent coordinates (Section A.1).

At a query point $x$, client $k$ observes a stochastic tangent vector $G_k(x, \xi) \in T_x\mathcal{M}$, conditionally unbiased for $\nabla f_k(x)$, with moment and covariance assumptions stated in Section C.1; here $T_x\mathcal{M}$ is the tangent space, $g$ gives its inner product and norm, and $d(x, y)$ is geodesic distance.

A retraction $R_x : T_x\mathcal{M} \to \mathcal{M}$, with local inverse $R_x^{-1}$, and a smooth transport $\mathcal{T}_x^y : T_x\mathcal{M} \to T_y\mathcal{M}$ implement these operations; on the local neighborhood used by the theory, $\mathcal{T}_x^y$ is an isometry, $\mathcal{T}_y^x = (\mathcal{T}_x^y)^{-1}$, and $\mathcal{T}_x^x = \mathrm{Id}$. The exponential map $\mathrm{Exp}_x$, with local inverse $\mathrm{Log}_x$, and Levi–Civita parallel transport are the canonical example; for an embedded manifold, locally smooth nearest-point projection gives the retraction $R_x(v) = \Pi_{\mathcal{M}}(x + v)$, which on the unit sphere is $(x+v)/\|x+v\|$ (Boumal, 2023, Section 5.12). Retraction–transport pairs used below satisfy the second-order compatibility condition (17).

At round $m$, every client starts from the synchronized point $x_m$, takes $E_m$ local steps with local step size $\eta_m$, transports its stochastic gradients to $T_{x_m}\mathcal{M}$, and sends one averaged tangent vector.

Initialize $x^k_{m,0} = x_m$ and write

$$x^k_{m,e+1} = R_{x^k_{m,e}}\left(-\eta_m G^k_{m,e+1}\right), \qquad G^k_{m,e+1} = G_k(x^k_{m,e}, \xi^k_{m,e+1}), \quad e = 0, \dots, E_m - 1, \tag{1}$$

$$U^k_{m+1} = \frac{1}{E_m} \sum_{e=0}^{E_m-1} \mathcal{T}^{x_m}_{x^k_{m,e}} G^k_{m,e+1}, \tag{2}$$

$$U_{m+1} = \sum_{k=1}^{K} p_k U^k_{m+1}, \qquad x_{m+1} = R_{x_m}(-\gamma_m U_{m+1}), \qquad \gamma_m = E_m \eta_m. \tag{3}$$

Equations (1)–(3) give a weighted full-participation form of RFedAGS (Huang et al., 2026); equal weights $p_k = 1/K$ recover its recursion after matching the step-size convention.

The statistical clock after $T$ communication rounds, $t_T = \sum_{m=1}^{T} E_m$, counts optimization oracle calls per participating client ($Kt_T$ calls in all under full participation); fresh calls for covariance estimation are counted separately below, and the averaged communication frequency is $\mathrm{ACF}_T = T/t_T$. Let $S^\star_k = \mathrm{Cov}\{G_k(x_\star, \xi)\}$ and $S = \sum_k p_k^2 S^\star_k$ be the client and pooled noise covariances at the target; with $t_T$ as the normalization, the averaging across clients is recorded in $S$.

Choose the interval $E_m$ and effective step $\gamma_m$ together and set $\eta_m = \gamma_m / E_m$; for power schedules $E_m \asymp m^\beta$ and $\gamma_m \asymp m^{-\alpha}$, local steps decay as $m^{-(\alpha+\beta)}$. Theorem C.9 gives admissible exponents for the available oracle moment, and Section 4.1 describes pilot selection of the step constants.

### 2.2 Two online geometric averages

Averaging reduces trajectory noise while retaining a manifold-valued output; let $b \geq 1$ activate averaging after $b-1$ discarded rounds and set $n_T = T - b + 1$. Both averages weight rounds equally, and a fixed burn-in leaves the first-order limit unchanged.

For the *fixed-anchor tangent average*, take the public anchor $a = x_b$ and maintain

$$\bar{v}_T = \frac{1}{n_T} \sum_{m=b}^{T} R_a^{-1}(x_m) \in T_a\mathcal{M}, \qquad \widehat{x}^{\mathrm{FA}}_T = R_a(\bar{v}_T). \tag{4}$$

This average is online, uses $O(d)$ server memory, and needs no convexity of a squared retraction distance. The anchor-free alternative recursively moves the current average toward each new point by a fraction $1/(m-b+1)$; its update and additional local condition are given with Theorem C.10.

**A proportional tail.** The basic theorems suppress only a fixed, asymptotically negligible burn-in; the finite-sample implementation also allows a prespecified fraction $\rho \in [0, 1)$: put $b_T = \lfloor \rho T \rfloor$, $n_{T,\rho} = T - b_T$, and form either geometric average from $x_{b_T+1}, \dots, x_T$, with total-clock variance factor

$$\nu_{T,\rho} = \frac{t_T}{n^2_{T,\rho}} \sum_{m=b_T+1}^{T} E_m^{-1}. \tag{5}$$

Discarding a positive fraction changes the first-order covariance, so this factor uses the retained rounds and their actual count; the limit and the geometric and plug-in conclusions are in Theorem C.11.

### 2.3 Federated covariance messages and output

At synchronized point $x_m$ client $k$ provides a fresh stochastic-gradient (*score*) message $G_k(x_m, \zeta^k_m)$ and either the exact Hessian $\nabla^2 f_k(x_m)$ or an unbiased self-adjoint Hessian message $B_k(x_m, \omega^k_m)$. The score and Hessian may use the same fresh observation; both are independent of the optimization calls conditional on the synchronized history. Write $u \otimes v$ for the outer-product operator $w \mapsto u\langle v, w\rangle$. For either final estimator $\widehat{x}_T \in \{\widehat{x}^{\mathrm{FA}}_T, \widetilde{x}_T\}$, pool the transported scores as

$\widehat{Z}_m = \sum_k p_k \mathcal{T}_{x_m}^{\widehat{x}_T} G_k(x_m, \zeta_m^k)$. With $\widehat{\nu}_T = (t_T/n_T^2) \sum_{m=b}^{T} E_m^{-1}$, set

$$\widehat{H}_T = \frac{1}{n_T} \sum_{m=b}^{T} \sum_{k=1}^{K} p_k \mathcal{T}_{x_m}^{\widehat{x}_T} B_k(x_m, \omega_m^k) \mathcal{T}_{\widehat{x}_T}^{x_m}, \tag{6}$$

$$\widehat{S}_T = \frac{1}{n_T} \sum_{m=b}^{T} \widehat{Z}_m \otimes \widehat{Z}_m, \tag{7}$$

$$\widehat{\Sigma}_T = \widehat{\nu}_T \widehat{H}_T^{-1} \widehat{S}_T \widehat{H}_T^{-1}. \tag{8}$$

The order of aggregation in (7) matters under client heterogeneity: the pooled gradient vanishes at $x_\star$ although individual client gradients may not, so taking the outer product after summing the weighted scores uses this cancellation, whereas weighting separate uncentered client outer products by $p_k^2$ would retain the individual mean-gradient terms even at the target. Consistency also accounts for transport to the estimated tangent space (Section J).

The operators in (6)–(8) act on $T_{\widehat{x}_T}\mathcal{M}$, with $\widehat{\Sigma}_T/t_T$ estimating the covariance of the local estimation error. The server retains the inference-round points and aggregated messages, transports them to the final tangent space, and forms the sandwich when $\widehat{H}_T \succ 0$ and the full Wald region (15) when also $\widehat{S}_T \succ 0$, otherwise reporting the failed conditioning check; under the plug-in assumptions and $S \succ 0$, these checks pass with probability tending to one.

For proportional-tail inference, the geometric average and all covariance sums use the same retained rounds and count. Algorithm 1 gives the complete procedure.

**Algorithm 1** Inferential RFedAGS under full participation

**Require:** Initial point $x_1$, weights $p_k$, intervals $E_m$, steps $\eta_m$, retraction $R$, transport $\mathcal{T}$, horizon $T$, first retained round $b$, confidence level $1 - \alpha_{\mathrm{W}}$.

Choose $T, b$ from the total budget as in Section 2.3; set $\gamma_m = E_m \eta_m$.
**for** $m = 1, \ldots, T$ **do**
  **if** $m = b$ **then**
    Set anchor $a = x_b$ and $\bar{v}_b = 0$.
  **if** $m > b$ **then**
    Update $\bar{v}_m = \bar{v}_{m-1} + \{R_a^{-1}(x_m) - \bar{v}_{m-1}\}/(m - b + 1)$.
  **for all** clients $k$ in parallel **do**
    Start at $x_m$; take $E_m$ steps (1); send stream (2).
    **if** $m \geq b$ **then**
      Send fresh score $G_k(x_m, \zeta_m^k)$ and Hessian $B_k(x_m, \omega_m^k)$.
  **if** $m \geq b$ **then**
    Retain $x_m$ and the weighted score and Hessian aggregates.
  Form $U_{m+1} = \sum_k p_k U_{m+1}^k$ and update $x_{m+1}$ by (3).
Set $\widehat{x}_T = R_a(\bar{v}_T)$; transport retained messages to its tangent space.
Form (6)–(7); if $\widehat{H}_T, \widehat{S}_T \succ 0$, compute (8) and return (15); otherwise return the estimate and report the failed check.
Recursive option: use (38) and output $\widetilde{x}_T$.

**Total inference cost.** The clock $t_T$ measures the work that generates the optimization trajectory; plug-in inference adds one fresh score per client in each of the $n_T$ retained rounds, giving $t_T + n_T$ gradient-oracle calls per client in Algorithm 1. A Hessian computed from that same fresh observation needs no additional draw, an independent stochastic Hessian oracle has its own cost, and exact Hessian evaluation has a computational cost even without a fresh observation.

For a prespecified proportional tail the total is $t_T + n_{T,\rho}$ calls per client: fix the interval schedule, step constants and $\rho$ before sampling, choose the largest complete-round horizon $T$ that fits the optimization-plus-score budget, and set $b = 1 + \lfloor \rho T \rfloor$ in Algorithm 1, so that anchor, average and fresh messages all start at the same round; the Wald statistic uses the optimization clock $t_T$ and the retained-round factor (5).

With float64 coordinates, an optimization vector and server downlink in every round, and a score vector and dense Hessian in each retained round, the communication cost is

$$8K\{2dT + (d + d^2)n_T\} \quad \text{bytes.} \tag{9}$$

This includes both uplinks and downlinks, but excludes network headers and latency. Retaining the points and aggregated messages for final transport uses $O(n_T d^2)$ server storage, so the point average is online and inexpensive, whereas full covariance estimation can be dominated by Hessian communication and storage in large dimension.

**Streaming principal subspaces.** For a concrete instance, let $\mathcal{M} = \mathrm{Gr}(r, D)$ be the space of $r$-dimensional subspaces of $\mathbb{R}^D$, with $1 \le r < D$ and tangent dimension $r(D-r)$. Client $k$ observes $h_k \in \mathbb{R}^D$ with second-moment matrix $\Sigma_k = \mathbb{E}(h_k h_k^\top)$. For an orthonormal representative $U \in \mathbb{R}^{D\times r}$, $U^\top U = I$, the objective and stochastic gradient are

$$f_k(U) = -\frac{1}{2}\mathrm{tr}(U^\top \Sigma_k U), \qquad G_k(U, h) = -(I - UU^\top)hh^\top U.$$

Centering the pooled distribution makes $\Sigma = \sum_k p_k \Sigma_k$ its covariance, even when client means differ. If $\lambda_r(\Sigma) > \lambda_{r+1}(\Sigma)$, the dominant subspace is locally identified and the Riemannian Hessian is positive definite in the quotient tangent space. These gradients and Hessians supply the messages for the covariance calculation above.

## 3 Theoretical Results

The argument follows the obstacles in Section 1: a one-round expansion separates sampling noise from local drift and geometric error; a sharper state bound controls their accumulation; averaging and fresh messages then give a computable limit law and Wald region. We first summarize Assumptions C.1–C.4. The analysis is local, with fixed client count and manifold dimension: the synchronized points, local iterates and reported averages remain in a common compact neighborhood $\mathcal{U}$ of a nondegenerate minimizer $x_\star$; the objective has positive-definite Hessian $H_\star = \nabla^2 f(x_\star)$ and a quadratic gradient linearization, and client gradients are locally Lipschitz after transport; retractions are smooth, their coordinate error obeys the second-order interface (17), and the two-point transport is isometric and invertible, exponential maps with local parallel transport being the canonical case. Oracle calls are conditionally unbiased with uniformly bounded $(2+\delta)$ moments, $\delta \in (0, 2]$; the centered noises are martingale differences, conditionally uncorrelated across clients, whose covariance fields satisfy the local Lipschitz bound (22) along the actual two-leg transport path. Write $\kappa = \min(\delta, 1)$. The schedule conditions control accumulated drift, the endpoints left by averaging and the martingale Lyapunov term; their full statements and verification routes are in Section C.1, and Theorem C.9 gives directly usable power schedules. Positive Hessians and locally smooth objectives are the familiar conditions behind iterate averaging (Polyak & Juditsky, 1992; Tripuraneni et al., 2018); the moment and schedule conditions have their Euclidean Local SGD counterpart in Li et al. (2022). On a compact normal neighborhood, bounded third derivatives give the objective expansion and smooth covariance fields give the transport bound; intrinsic means supported in a sufficiently small convex ball and principal subspaces with a positive eigengap are concrete examples (Section A.1), and the geometric interface and localization identify the basin in which these local conditions apply.

### 3.1 From local computation to a sampling law

Set $\Delta_m = R_{x_\star}^{-1}(x_m)$. Transporting each client gradient through the server gives the path $\mathcal{T}_{x_m}^{x_\star}\mathcal{T}_{x_{m,e}^k}^{x_m}$, without a transitivity assumption. Theorem C.7 yields

$$\Delta_{m+1} = (\mathrm{Id} - \gamma_m H_\star)\Delta_m - \gamma_m \epsilon_{m+1} + \gamma_m r_{m+1}.$$

Here $\epsilon_{m+1}$ averages the centered client noises, with weights $p_k/E_m$, after transport along this path, with leading covariance $S/E_m$, and $r_{m+1}$ collects client drift, objective nonlinearity and retraction error. Balancing this noise against contraction gives the following rate, where $q_m \asymp \eta_m + \gamma_m^{1+\kappa}$ is the nonincreasing comparison sequence in (23).

**Theorem 3.1** (Mean-square rate of the synchronized iterate)**.** *Under Assumptions C.1–C.4, the local neighborhood may be chosen small enough that*

$$\mathbb{E}\,\|\Delta_m\|^2 = O(q_m) = O\big(\eta_m + \gamma_m^{1+\kappa}\big), \qquad \kappa = \min(\delta, 1). \tag{10}$$

*In particular,* $\mathbb{E}\,\|\Delta_m\|^2 = O(\eta_m + \gamma_m^2)$ *if* $\delta \geq 1$*, and the weaker bound* $O(\gamma_m)$ *follows for every* $\delta > 0$ *because* $\eta_m \leq \gamma_m$ *and* $\gamma_m^{1+\kappa} \leq \gamma_m$ *eventually.*

The noise term $\gamma_m^2/E_m$ per round contracts at rate $\gamma_m$, giving $\eta_m$ in the state error: averaging $E_m$ local noises offsets part of the larger effective step, while the geometric remainder contributes $\gamma_m^{1+\kappa}$. The sharper rate (10) removes the Abel endpoints at the $\sqrt{t_T}$ scale, whereas the generic $O(\gamma_m)$ bound would unnecessarily restrict interval growth (proof in Section G).

For power schedules, the coarse scale gives endpoint order $T^{(\alpha+\beta-1)/2}$, requiring $\alpha + \beta < 1$, whereas the refined rate gives $T^{(2\alpha+\beta-1-r)/2}$ with $r = \min\{\alpha+\beta, \alpha(1+\kappa)\}$, allowing faster interval growth.

For $E_m = \max\{1, \lfloor c_E m^\beta \rfloor\}$ and $\gamma_m = c_\gamma m^{-\alpha}$ with $c_E, c_\gamma > 0$, $0 \leq \beta < 1$ and $1/2 < \alpha < 1$, the schedule conditions hold if $\alpha > (1+\beta)/2$ and $\min\{\alpha+\beta, \alpha(1+\kappa)\} > 2\alpha + \beta - 1$. When $\delta \geq 1$, only $(1+\beta)/2 < \alpha < 1$ is needed; when $0 < \delta < 1$, the second inequality matters only if $\beta > \alpha\delta$, in which case it becomes $\alpha(1-\delta) + \beta < 1$. The Lyapunov condition holds automatically (Theorem C.9).

The schedule factor is $\nu = \lim_{T\to\infty}(t_T/T^2)\sum_{m=1}^T E_m^{-1}$. Relative to the same weighted oracle and optimization clock, constant intervals retain first-order precision, while slowly varying intervals can also have $\nu = 1$ with $T/t_T \to 0$. For polynomial growth, $T/t_T \asymp T^{-\beta}$ and $\nu = (1-\beta^2)^{-1}$ quantify the communication–variance tradeoff (Theorem C.13).

Each synchronized point has equal weight while round $m$ combines $E_m$ observations, so the product of the arithmetic means of $E_m$ and $E_m^{-1}$ controls the noise. Constant and slowly varying intervals balance these weights asymptotically; for power intervals, later rounds receive the same weight as earlier, noisier rounds despite using more observations, leaving a variance factor above one, the same schedule factor as in Euclidean Local SGD (Li et al., 2022), and the geometric state bound shows when it survives transport and retraction errors. Square-root intervals have $\nu = 4/3$ with a fixed burn-in, logarithmic intervals have $\nu = 1$ and still give $T/t_T \to 0$, and single-stream averaged RSGD (Tripuraneni et al., 2018) corresponds to $K = E_m = 1$. These comparisons hold the optimization clock fixed; the fresh-score cost and proportional tail determine the finite-budget comparison below.

### 3.2 Geometric averages and feasible inference

For the target-chart average $\bar{\Delta}_T = T^{-1}\sum_{m=1}^T \Delta_m$, the representation and endpoint bounds yield $\sqrt{t_T}\bar{\Delta}_T = -H_\star^{-1}(\sqrt{t_T}/T)\sum_{m=1}^T \epsilon_{m+1} + o_{\mathbb{P}}(1)$, and Theorem C.8 gives its Gaussian limit with $\Sigma_\star = \nu H_\star^{-1} S H_\star^{-1}$. The fixed-anchor output makes this otherwise inaccessible target-chart average computable.

**Theorem 3.2** (Fixed-anchor average)**.** *Under Theorem C.8, the estimator in* (4) *satisfies*

$$\sqrt{t_T}\left\{R_{x_\star}^{-1}(\widehat{x}_T^{\mathrm{FA}}) - \bar{\Delta}_T\right\} \xrightarrow{\mathbb{P}} 0, \tag{11}$$

$$\sqrt{t_T}\, R_{x_\star}^{-1}(\widehat{x}_T^{\mathrm{FA}}) \xrightarrow{d} N(0, \Sigma_\star). \tag{12}$$

The fixed-anchor chart and its inverse have cancelling derivatives, so the anchor needs no prescribed convergence rate. The recursive average has the same limit under the local retraction-convexity and summability condition in Theorem C.5 (Theorem C.10). For a proportional tail, Theorem C.11 replaces $\nu$ by $\nu_\rho = (1-\rho^{1-\beta})/\{(1-\rho)^2(1-\beta^2)\}$ for polynomial intervals and by $1/(1-\rho)$ for constant intervals: excluding early points changes both initialization error and the first-order variance on the total optimization clock, so the finite estimator uses (5).

For feasible inference, the fresh messages are independent across clients and of the optimization calls conditional on the synchronized history; unbiased self-adjoint Hessian messages, exact Hessians included, have a uniformly bounded $(1+\tau)$ moment for some $\tau > 0$, and score and Hessian messages may share one fresh observation (Theorem C.6).

**Theorem 3.3** (Consistency of the federated sandwich estimator)**.** *Let* $\widehat{x}_T$ *be either* $\widehat{x}_T^{\mathrm{FA}}$ *under Theorem 3.2 or* $\widetilde{x}_T$ *under Theorem C.10. Under Assumptions C.1–C.4 and Assumption C.6, together*

*with Assumption C.5 only when $\widehat{x}_T = \widetilde{x}_T$, after transport to $T_{x_\star}\mathcal{M}$,*

$$\mathcal{T}_{\widehat{x}_T}^{x_\star} \widehat{H}_T \mathcal{T}_{x_\star}^{\widehat{x}_T} \xrightarrow{\mathbb{P}} H_\star, \qquad \mathcal{T}_{\widehat{x}_T}^{x_\star} \widehat{S}_T \mathcal{T}_{x_\star}^{\widehat{x}_T} \xrightarrow{\mathbb{P}} S, \tag{13}$$

$$\mathcal{T}_{\widehat{x}_T}^{x_\star} \widehat{\Sigma}_T \mathcal{T}_{x_\star}^{\widehat{x}_T} \xrightarrow{\mathbb{P}} \Sigma_\star. \tag{14}$$

For $S$ positive definite and confidence level $1-\alpha_{\mathrm{W}}$, $0 < \alpha_{\mathrm{W}} < 1$, let $\chi^2_{d,1-\alpha_{\mathrm{W}}}$ denote the $(1-\alpha_{\mathrm{W}})$ quantile of $\chi^2_d$ and define

$$\mathcal{C}_{T,1-\alpha_{\mathrm{W}}} = \left\{ y \in \mathcal{U} : t_T \left\langle R^{-1}_{\widehat{x}_T}(y), \widehat{\Sigma}_T^{-1} R^{-1}_{\widehat{x}_T}(y) \right\rangle \le \chi^2_{d,1-\alpha_{\mathrm{W}}} \right\}. \tag{15}$$

**Corollary 3.4** (Wald inference)**.** *Under Theorem 3.3, if $S \succ 0$, then*

$$\mathbb{P}\{x_\star \in \mathcal{C}_{T,1-\alpha_{\mathrm{W}}}\} \longrightarrow 1-\alpha_{\mathrm{W}}. \tag{16}$$

The quadratic form measures squared error in units of its estimated covariance, and its chi-square cutoff gives joint coverage for the manifold parameter; smooth functions of that parameter inherit uncertainty through the delta method (Theorem C.12). Proofs of plug-in consistency and coverage are in Section J.

## 4 Simulations

We test whether communication savings survive the cost of estimating uncertainty, and whether a schedule allowed by the limit theorem is calibrated at practical horizons. All configurations use eight equally weighted clients, the fixed-anchor estimator and feasible full-dimensional 95% Wald regions; coverage close to 95% indicates calibration, MSE is the mean squared geodesic distance to the target, the uncertainty size is $\{\mathrm{tr}(\widehat{\Sigma}_T/t_T)\}^{1/2}$ (useful only with adequate coverage), and communication cost is measured in bytes using (9).

### 4.1 Models and matched-budget comparisons

The Euclidean model has heterogeneous quadratic objectives with a common diagonal Hessian, eight opposite client shifts and independent Gaussian gradient noise in dimension five; the target is zero, exact Hessians and fresh scores estimate the covariance, and Theorem K.1 proves inference directly for this unmodified recursion. The curved comparison estimates the intrinsic mean on $S^2$: tangent Gaussian observations at the north pole are truncated to radius 0.90 radians and mapped by $\mathrm{Exp}$, opposite client shifts make the north pole the pooled target, and each fresh observation provides both $-\mathrm{Log}_x(Y)$ and the Hessian of its squared-distance loss, so this comparison includes the sampling variation and cost of second-order messages (distributions and initialization in Section B).

We compare $E_m = 1, 8$ with $E_m = \lfloor\sqrt{m}\rfloor$, using $\gamma_m = c_\gamma(m+50)^{-\alpha}$. Constants $(\alpha, c_\gamma, \rho)$ are $(0.65, 0.8, 0.1)$ for constant intervals and $(0.80, 1.6, 0.5)$ for the main square-root configuration, with at least 20 discarded rounds. Disjoint pilots selected original gains using last-iterate error and stability; the additional Gaussian gain and tail comparisons were specified before simulation. Each configuration takes the largest complete horizon with $t_T + T - \max\{20, \lfloor\rho T\rfloor\} \le B$, so inference work is inside the budget, and its anchor, average, fresh messages and factor (5) use the same retained rounds.

At this total budget, the Gaussian $E_m = 8$ baseline spends 14720 calls on optimization against 8623 for $E_m = 1$, because it requires fewer fresh inference scores; its smaller uncertainty size reflects these additional optimization calls, and at a matched optimization clock both constant intervals have $\nu = 1$.

The complete Gaussian and sphere configurations in Table 1 give calibrated uncertainty with substantial communication savings. Against $E_m = 8$, the Gaussian half-tail has the same 94.5% coverage with 1.35 rather than 4.36 MB, a 69% reduction, while its size increases from $3.41$ to $3.88$ in units of $10^{-3}$, about 14%; on the sphere, coverage is 95.4% versus 94.4%, communication cost falls from 1.107 to 0.378 MB (66%), size increases from 0.756 to 0.859 (14%), and mean geodesic error rises from $0.681$ to $0.760$ in units of $10^{-3}$. The savings thus come with greater dispersion and, on the

Table 1: Inference at total per-client oracle budget $B = 2^{14}$. Gaussian, sphere and $\mathrm{SO}(3)$ rows use 1000, 500 and 300 replications. MSE is mean squared geodesic error; time is the measured local computation per replication and, since implementations differ across models, compares schedules within each model only.

| Model | $E_m$ | $\rho$ | Coverage (%) | MSE $\times 10^6$ | Time (ms) |
|---|---|---|---|---|---|
| Gaussian | 1 | 0.1 | 94.6 | 20 | 3.49 |
| | 8 | 0.1 | 94.5 | 11.8 | 0.664 |
| | $\lfloor\sqrt{m}\rfloor$ | 0.5 | 94.5 | 14 | 0.252 |
| Sphere | 8 | 0.1 | 94.4 | 0.59 | 89.2 |
| | $\lfloor\sqrt{m}\rfloor$ | 0.5 | 95.4 | 0.727 | 80.8 |
| $\mathrm{SO}(3)$ | 8 | 0.1 | 94.3 | 0.553 | 172 |
| | $\lfloor\sqrt{m}\rfloor$ | 0.5 | 95.3 | 0.619 | 200 |

sphere, larger estimation error. At $B = 2^{16}$, Gaussian coverage remains 94.7% for the half-tail and 95.3% for $E_m = 8$ (costs and sizes in Table 3).

The $\mathrm{SO}(3)$ experiment estimates the identity rotation from eight bounded rotation-vector distributions with opposite means, using fresh sample Hessians and the same two schedule choices (Section B.3); its 67% communication saving accompanies 12% higher MSE and 16% longer computation, as the growing-interval schedule spends more of the total oracle budget on local optimization and fewer calls on inference messages.

### 4.2 What tail averaging changes

The direct tail comparison isolates the role of removing early iterates. On the sphere, changing $\rho$ from 0.1 to 0.5 at a fixed interval and step configuration raises coverage from 81.2% to 95.4% and reduces mean distance from $0.890$ to $0.760$ in units of $10^{-3}$; in the Gaussian design, the same ablation with unchanged $c_\gamma = 0.8$ raises coverage from 71.0% to 96.2%, and the half-tail with $c_\gamma = 1.6$ gives 94.5%. Tail removal also lowers communication cost (Section B.2): removing early iterates reduces initialization error, while the larger retained-round variance factor compensates for averaging fewer points, so the resulting regions are wider.

The sphere half-tail is stable to the two prespecified gain perturbations $c_\gamma = 1.2, 2.0$, which give 95.0% and 95.6% coverage, whereas a discarded fraction of 0.6 gives 98.2% coverage and a larger uncertainty size, the precision cost of discarding more rounds. Settings with Student $t_5$ noise, dimension 20, and symmetric Pareto noise of tail index 2.8 assess the moment and dimension assumptions; the Pareto law has no third moment, and with $\beta = 0.25$ it gives 94.6% coverage at per-client optimization budget $2^{14}$ in 500 replications, illustrating the lower-moment regime of the theory.

Initialization effects remain substantial at shorter horizons: at optimization budget $2^{14}$ with the original 10% tail, Gaussian $\beta = 0.75$ coverage is only 4.0% despite asymptotically admissible exponents, and at optimization budget $2^{12}$ the square-root sphere configuration has 0–3.8% coverage across radii 0.45–1.35, whereas synchronization after each step has 94.0–95.4%. These comparisons hold optimization work fixed and count fresh-score overhead separately; together with Table 1, they distinguish asymptotic admissibility from the finite-budget calibration achieved by the specified half-tail configurations.

## 5 Four-site Grassmann application

This application illustrates subspace inference when sites have unequal archive sizes but equal weight in the target distribution. The UCI Heart Disease archive Janosi et al. (1989), distributed under CC BY 4.0 and associated with the multi-site study of Detrano et al. (1989), contains processed records from Cleveland, Hungary, Switzerland, and Long Beach VA; after excluding a record missing one of the five selected fields and records with nonpositive resting blood pressure or maximum heart rate, the included counts are 303, 293, 117, and 140. The five variables are age, resting blood pressure,

maximum heart rate, exercise-induced ST depression, and the archive's numeric 0–4 disease diagnosis code, which enters as a numerical feature.

For raw feature vectors $h_{ki}$ at site $k$ with included count $n_k$, define $\mu = \frac{1}{4}\sum_{k=1}^{4} n_k^{-1}\sum_i h_{ki}$, $s_j^2 = \frac{1}{4}\sum_{k=1}^{4} n_k^{-1}\sum_i (h_{kij}-\mu_j)^2$ and $z_{kij} = (h_{kij}-\mu_j)/s_j$. The target is the top-two eigenspace of $\bar{C} = \frac{1}{4}\sum_k n_k^{-1}\sum_i z_{ki}z_{ki}^\top$, a point of $\mathrm{Gr}(2,5)$, the space of two-dimensional linear subspaces of $\mathbb{R}^5$. This covariance includes between-site mean variation because all sites use the same center, and each empirical site distribution has weight $1/4$, so the estimand is the exact equal-site finite-archive subspace, locally identified by the positive second-to-third eigenvalue gap of the finite-archive covariance. A deterministic warm start uses the first 20 complete records per site; the online oracle then samples with replacement from the four archived site distributions, with disjoint optimization and fresh-score streams.

Averaging starts after a fixed burn-in of 100 rounds ($b = 101$ in (4)) with $\gamma_m = c_\gamma(m+50)^{-0.65}$, giving 8092 retained rounds for $E_m = 1$ and 924 for $E_m = 8$ and finite variance factor $t_T(T-100)^{-2}\sum_{m=101}^{T} E_m^{-1}$. A disjoint pilot, blind to coverage, fixed $c_\gamma = 1.2$ for both schedules, and one prespecified trajectory per schedule was run to a per-site optimization clock of 8192. The Grassmann exponential map, logarithm, and parallel transport implement the geometric operations, and the tangent dimension is $2(5-2) = 6$. Fresh scores are sampled independently of the optimization stream, and the server computes the exact pooled Hessian from $\bar{C}$, algebraically equal to aggregating exact client Hessians; the application thus uses pooled sufficient statistics where the general protocol sends Hessian messages.

Table 2: Grassmann application at per-site optimization clock 8192. Geodesic error is the distance from the fixed-anchor estimate to the exact equal-site finite-archive target. The Wald statistic is the quadratic form in (15) evaluated at that target; its six-dimensional 95% threshold is 12.59. The tangent uncertainty size is $\{\mathrm{tr}(\widehat{\Sigma}_T/t_T)\}^{1/2}$.

| Schedule | Rounds | Geodesic error | Uncertainty size | Wald statistic |
|---|---|---|---|---|
| $E_m = 1$ | 8192 | 0.110 | 0.146 | 3.20 |
| $E_m = 8$ | 1024 | 0.208 | 0.155 | 5.69 |

Eight local updates per round reduce the number of synchronizations by a factor of eight, while the realized geodesic error is nearly twice as large (Table 2); the uncertainty size grows only from 0.146 to 0.155, so the realized error increase is larger than the change in estimated dispersion, and total gradient draws per site, including fresh scores, are 16284 and 9116, so synchronization counts alone do not measure the full inference cost. Both Wald statistics remain below the stated threshold, so the computed regions contain the exact subspace target in these two trajectories. Uncertainty is conditional on the empirical site distributions, describing variation from resampling records through the online oracle; the comparison is descriptive for one trajectory per schedule, and the replicated simulations assess coverage across repeated trajectories.

## 6 Concluding Remarks

The synchronized-state rate controls RFedAGS geometric drift and averaging endpoints on the scale of all local observations, and a fixed-anchor average attains a Gaussian law whose covariance is estimated by fresh pooled scores and Hessian messages. Slowly varying intervals preserve first-order precision with vanishing communication frequency, whereas power intervals incur an explicit variance factor. Gaussian, sphere and rotation experiments show that a prespecified half-tail can recover calibration with much lower communication cost, at the cost of larger uncertainty and estimation error.

Extending inference beyond full participation requires controlling client-selection noise together with local drift, and in larger dimensions approximate Hessian inverse actions could reduce dense communication and storage if their error preserves the sandwich limit; both extensions retain the need to identify and enter a regular local basin.

## 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 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

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

B. Afsari. Riemannian $L^p$ center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds—II. *The Annals of Statistics*, 33(3):1225–1259, 2005.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1987.

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5):304–310, 1989.

J. Fan, D. Wang, K. Wang, and Z. Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

L.-K. Huang and S. Pan. Communication-efficient distributed PCA by Riemannian optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4465–4474. PMLR, 2020.

Z. Huang, W. Huang, P. Jawanpuria, and B. Mishra. Riemannian federated learning via averaging gradient streams. In *International Conference on Learning Representations*, 2026. arXiv:2409.07223.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart Disease [Dataset]. UCI Machine Learning Repository, 1989.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020.

J. Li and S. Ma. Federated learning on Riemannian manifolds. *Applied Set-Valued Analysis and Optimization*, 5(2):213–232, 2023.

X. Li, J. Liang, X. Chang, and Z. Zhang. Statistical estimation and online inference via Local SGD. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1613–1661. PMLR, 2022.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. Efficient estimators from a slowly convergent Robbins–Monro process. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, 1988. `https://hdl.handle.net/1813/8664`.

B. Saparbayeva, M. M. Zhang, and L. Lin. Communication efficient parallel algorithms for optimization on manifolds. In *Advances in Neural Information Processing Systems 31*, 2018.

S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 650–687. PMLR, 2018.

# APPENDIX

Sections A and B give model examples and experimental protocols. Section C states the assumptions and additional limit results. Proofs follow in Sections D to J.

## A REPRESENTATIVE MODELS

### A.1 EXACT REDUCTIONS AND REPRESENTATIVE MODELS

**Proposition A.1** (Exact algorithmic reductions)**.** *The optimization recursion* (1)–(3) *has the following exact special cases.*

*(i) For $p_k = 1/K$, it is the full-participation RFedAGS gradient-stream recursion after a notational rescaling of the client stream.*

*(ii) If $\mathcal{M} = \mathbb{R}^d$, $R_x(v) = x + v$, and $\mathcal{T}_x^y = \mathrm{Id}$, then*

$$x_{m+1} = \sum_{k=1}^{K} p_k x_{m,E_m}^k = x_m - \eta_m \sum_{k=1}^{K} p_k \sum_{e=0}^{E_m-1} G_{m,e+1}^k,$$

*which is full-participation Local SGD/FedAvg.*

*(iii) If $K = 1$ and $E_m \equiv 1$, then $x_{m+1} = R_{x_m}(-\eta_m G_{m,1}^1)$; with* (38)*, this is the averaged RSGD recursion of Tripuraneni et al. (2018).*

**Intrinsic Fréchet means.** For client observations $Y_k \in \mathcal{M}$, take

$$f_k(x) = \frac{1}{2}\mathbb{E}d^2(x, Y_k), \qquad G_k(x, Y) = -\operatorname{Log}_x(Y)$$

inside a common normal neighborhood. Then

$$S = \sum_{k=1}^{K} p_k^2 \operatorname{Cov}\{\operatorname{Log}_{x_\star}(Y_k)\},$$

and $H_\star$ is the Hessian of the pooled Fréchet function. Existence, uniqueness, and regular intrinsic-mean limits under local support and Hessian conditions are classical Afsari (2011); Bhattacharya & Patrangenaru (2005); the new part here is the local-step federated trajectory and its schedule-indexed inference.

**Streaming principal subspaces.** The objective, gradient and eigengap condition for this model are given in Section 2.3; Section 5 applies the construction to a four-site principal-subspace problem.

### A.2 TRANSPORT THROUGH THE SERVER

The sphere makes the geometric distinction in the gradient stream visible. For nonantipodal unit vectors $x, y$ joined by their shortest geodesic and $v \in T_x S^2$, parallel transport is

$$\mathcal{T}_x^y v = v - \frac{\langle v, y\rangle}{1 + \langle x, y\rangle}(x + y).$$

Applying this map along the local-to-server and server-to-target legs can rotate a vector relative to direct local-to-target transport, as shown in Figure 1. Both routes preserve its norm. The covariance in (22) follows the two-leg route that occurs in the recursion.

## B EXPERIMENTAL PROTOCOLS AND SUPPORTING COMPARISONS

All simulations use fixed-anchor RFedAGS and $K = 8$ equally weighted clients. For independent replications $r = 1, \ldots, R$, write $D_r = d(\widehat{x}^{(r)}, x_\star)$. MSE is $R^{-1}\sum_r D_r^2$; coverage is the fraction of full-dimensional 95% Wald regions containing the target. The mean distance is $R^{-1}\sum_r D_r$, and uncertainty size is the mean of $\{\operatorname{tr}(\widehat{\Sigma}_T/t_T)\}^{1/2}$. All attempts count in coverage.

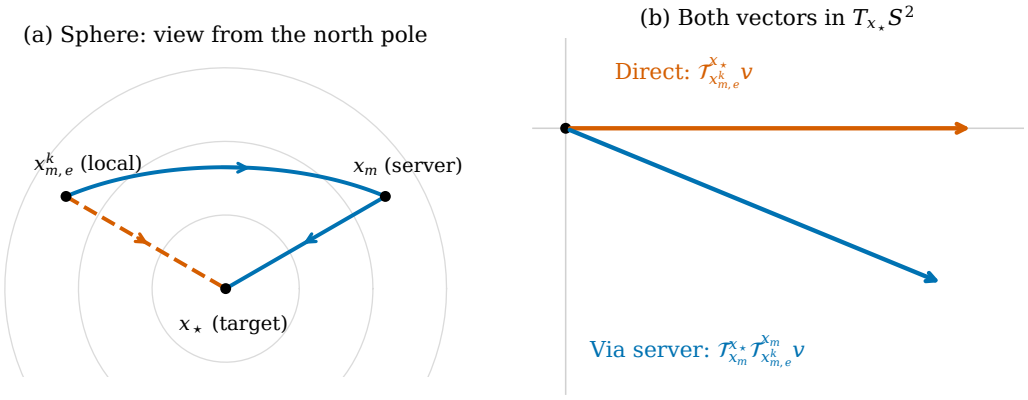



Figure 1: Parallel transport on $S^2$. Left: orthographic projection of shortest great-circle arcs in the northern hemisphere, with $x_\star$ at the north pole. The blue route passes through the server; the dashed orange route goes directly to the target. Right: transporting the same unit tangent vector $v$ along these routes gives different directions in the target tangent plane. Coordinates and transports are computed from the sphere formulas above.

### B.1 Models, initialization and step selection

The Euclidean model has heterogeneous quadratic client objectives

$$f_k(x) = \tfrac{1}{2}(x-a_k)^\top H(x-a_k), \qquad H = \mathrm{diag}(\lambda_1,\dots,\lambda_d), \quad \lambda_j = 1 + \frac{4(j-1)}{d-1},$$

with opposite client shifts satisfying $\sum_k p_k a_k = 0 = x_\star$ but nonzero individual gradients at $x_\star$. The oracle is $G_k(x,\xi) = H(x-a_k) + \xi$, with independent noise coordinates of variance $0.55 + 0.08(k-1) + 0.025(j-1)$. The main comparison uses Gaussian noise and $d = 5$, initialized with coordinates $0.55/\sqrt{d}$. Exact Hessians and independent fresh scores estimate the sandwich; Gaussian local updates are sampled from their exact round transition, preserving the law of the optimization trajectory.

The other Euclidean cells use variance-scaled Student $t_5$ or symmetric Pareto noise, in dimension 5 or 20. Before coordinate scaling, the Pareto variable is an independent fair sign times a Type-I Pareto magnitude with shape 2.8 and lower endpoint $\sqrt{(2.8-2)/2.8}$, giving unit variance. It has a finite $(2+\delta)$ moment for $0 < \delta < 0.8$ but no third moment.

For intrinsic means on $S^2$, observations are tangent Gaussians at the north pole, truncated by rejection sampling to a ball and mapped by Exp. The eight client means occur in opposite pairs along the two tangent axes, at distance 22% of the support radius from zero, with isotropic standard deviation 19% of that radius. The initial point is $\mathrm{Exp}_{x_\star}\{(0.42, 0.18)\}$ in an orthonormal tangent basis. The total-budget comparison uses radius 0.90 radians; the spread comparison uses 0.45, 0.90 and 1.35. Each fresh observation supplies the score $-\mathrm{Log}_x(Y)$ and the Hessian of $\frac{1}{2}d^2(x,Y)$. Dependence between these two messages is allowed by Theorem C.6.

Steps are $\gamma_m = c_\gamma(m+50)^{-\alpha}$ and $\eta_m = \gamma_m/E_m$. A disjoint pilot selected the constants using last-iterate error and stability, without using coverage. The exponents are $\alpha = 0.65$ for constant and logarithmic intervals, and $0.675, 0.80, 0.925$ for power exponents $\beta = 0.25, 0.50, 0.75$. The original Euclidean gains are $c_\gamma = 0.8$ except at $\beta = 0.75$, where $c_\gamma = 1.6$. The sphere pilot selected 0.8 for $E_m = 1$ and 1.6 for the square-root interval; the $E_m = 8$ comparison uses the same gain 0.8 as the synchronized baseline. These exponent pairs satisfy (37).

### B.2 Complete-budget protocol and tail comparisons

Fix the interval, gain and discarded fraction before the reported replications. At budget $B$, choose the largest integer $T$ for which $t_T + T - b_T \le B$, where $b_T = \max\{20, \lfloor \rho T \rfloor\}$. Average the pre-update points $x_{b_T+1}, \dots, x_T$, and use independent fresh scores at these same points. The anchor is $x_{b_T+1}$

Table 3: Gaussian $d = 5$ inference at total per-client budget $2^{16}$, 1000 replications per row. Constant intervals use $c_\gamma = 0.8$; the square-root half-tail uses 1.6. Coverage is reported as a percentage.

| $E_m$ | $\rho$ | Rounds | Calls | Coverage (%) | Size $\times 10^3$ | MB |
|---|---|---|---|---|---|---|
| 1 | 0.1 | 34492 | 65535 | 96.2 | 2.23 | 81.68 |
| 8 | 0.1 | 7363 | 65531 | 95.3 | 1.71 | 17.44 |
| $\lfloor\sqrt{m}\rfloor$ | 0.5 | 2129 | 65534 | 94.7 | 1.93 | 3.41 |

and the covariance uses the finite retained-round factor (5), with $b_T$ including the minimum discarded count. Exact Gaussian Hessians need no additional observation; on the sphere, the fresh score's observation also supplies the Hessian. Calls count optimization plus these fresh observations per client, and MB counts (9) across all clients.

The Gaussian comparison uses 1000 independent replications per configuration at budgets $2^{14}$ and $2^{16}$. Constant intervals use $(\alpha, c_\gamma) = (0.65, 0.8)$ and $\rho = 0.1$. Square-root intervals use $\alpha = 0.80$, with $(\rho, c_\gamma)$ equal to $(0.1, 0.8)$, $(0.5, 0.8)$ or $(0.5, 1.6)$. The unchanged-gain comparison isolates tail removal from the gain change. All Gaussian replications return a positive-definite covariance.

The communication cost formula separates the two configuration choices. At $B = 2^{14}$ and $\rho = 0.1$, changing $E_m = 8$ to $\lfloor\sqrt{m}\rfloor$ changes $(T, n_{T,\rho})$ from $(1840, 1656)$ to $(832, 749)$: communication cost falls from 4.357 to 1.971 MB for Gaussian data, and from 1.107 to 0.501 MB for sphere data. Keeping the square-root intervals and gain fixed, increasing $\rho$ to 0.5 yields $(T, n_{T,\rho}) = (844, 422)$ and costs of 1.350 and 0.378 MB, a further 31% and 24% reduction, respectively. The larger discarded fraction frees oracle calls for 12 additional optimization rounds while reducing the number of score and Hessian messages. At this horizon, the average local work per round is $t_T/T \approx 18.9$, compared with 8 for the constant interval. Thus the complete configurations combine more local work per round with fewer retained inference rounds. For fixed interval and tail choices, the message counts are independent of $\alpha$ and $c_\gamma$; those parameters affect the trajectory and its coverage.

At the larger budget, the half-tail gives 94.7% coverage with 3.41 MB versus 95.3% and 17.44 MB for $E_m = 8$. Its size is $1.93 \times 10^{-3}$ versus $1.71 \times 10^{-3}$: the communication saving persists, with a precision cost. For the special constant quadratic Hessian, a one-time cached transmission changes the communication cost to $8K\{d^2 + 2dT + dn_{T,\rho}\}$. At the smaller budget this gives 0.68 MB for either half-tail power configuration and 1.71 MB for $E_m = 8$. These counts do not measure network latency or Hessian computation time.

The sphere complete-budget comparison uses 500 independent replications for each of $E_m = 8, \rho = 0.1$ and square-root intervals with $\rho = 0.1, 0.5$, at radius 0.90 and $B = 2^{14}$. Their complete-round totals are 16376, 16359 and 16372 calls per client. All 1500 replications yield positive-definite Hessian and score covariance estimates. The two gain perturbations at $\rho = 0.5$, $c_\gamma = 1.2, 2.0$, have coverages 95.0% and 95.6%, sizes $0.857, 0.858$ in units of $10^{-3}$, and the same 0.378 MB communication cost. The tail-fraction perturbations use $c_\gamma = 1.6$: $\rho = 0.4$ gives 96.0% coverage, and $\rho = 0.6$ gives conservative 98.2% coverage. The latter's finite variance factor is 1.87 instead of the half-tail's 1.56. All 2000 perturbation runs pass the conditioning checks. These comparisons describe how the local precision tradeoff changes with the gain and tail fraction.

### B.3 Rotation means on $\mathrm{SO}(3)$

Identify a tangent vector with its body rotation vector in $\mathbb{R}^3$, with metric equal to its Euclidean inner product. Write $[v]_\times w = v \times w$. Client observations are $Y = \exp([\mu_k + \varepsilon]_\times)$, where $\varepsilon$ is uniform in the Euclidean ball of radius 0.3, and the eight means are the opposite pairs of $0.2e_1, 0.2e_2, 0.2e_3, 0.2(1,1,1)^\top/\sqrt{3}$. Equal weights and symmetry make the pooled score zero at the identity. All observations have rotation angle at most 0.5; the squared-distance objective has a positive Hessian in the local ball around the identity. The initial rotation vector is $(0.3, 0.2, -0.2)$. The exponential retraction and shortest-geodesic parallel transport implement RFedAGS. For $a = \mathrm{Log}(Q^{-1}R)$ in rotation-vector coordinates, parallel transport from $Q$ to $R$ acts on body vectors by $\exp(-[a]_\times/2)$. If $v = \mathrm{Log}(Q^{-1}Y)$, $\theta = \|v\|$ and $u = v/\theta$, the observation supplies score $-v$ and Hessian $hI + (1-h)uu^\top$, where $h = (\theta/2)\cot(\theta/2)$ and the limit at zero is $I$. The implementation

checks transport isometry and inverse consistency, exponential/logarithm inversion and the Hessian against a central difference of transported gradients.

Both configurations use $B = 16384$ and 300 independent replications, with separate optimization and fresh-observation generators. Constants $(\rho, \alpha, c_\gamma)$ are $(0.1, 0.65, 0.8)$ for $E_m = 8$ and $(0.5, 0.8, 1.6)$ for square-root intervals, fixed before simulation. The respective $(T, n_T, t_T)$ are $(1840, 1656, 14720)$ and $(844, 422, 15950)$; these counts also apply to the sphere comparison and corresponding Gaussian rows. The Gaussian $E_m = 1$ row has $(T, n_T, t_T) = (8623, 7761, 8623)$. All 600 rotation runs pass the Hessian and covariance checks; all realized local rotation angles stay below 0.414. The uncertainty sizes are $0.718$ and $0.815$ in units of $10^{-3}$. Squared errors, covariance matrices and Wald statistics are saved for every replication.

The times in Table 1 include local optimization and feasible inference. Communication costs for Gaussian $E_m = 1, 8$ and the half-tail are 20.42, 4.36 and 1.35 MB; for sphere $E_m = 8$ and the half-tail they are 1.107 and 0.378 MB, and for rotations 1.978 and 0.648 MB. Gaussian uncertainty sizes are 4.47, 3.41 and 3.88 in units of $10^{-3}$. Gaussian timings use vectorized exact round transitions, sphere timings use individual C++ trajectories, and rotation timings divide the NumPy batch time by its replication count. They measure computation on the recorded local implementations; communication cost is measured separately in bytes. Complete call totals are 16384 for Gaussian $E_m = 1$, 16376 for $E_m = 8$, and 16372 for the half-tail.

### B.4 Optimization-budget comparisons

The separate optimization-budget design has 30 settings with 500 independent replications each: 24 Euclidean and six sphere settings. The Euclidean budgets are $2^{12}, 2^{14}, 2^{16}$. Each trajectory completes the first round reaching its optimization budget; the normalization uses the actual $t_T$, and fresh covariance observations are additional. The discarded count is $\max\{20, \lfloor 0.1T \rfloor\}$. Schedules are $E_m = 1, 8, 1 + \lfloor \log_2(m+1) \rfloor$, and $\max\{1, \lfloor m^\beta \rfloor\}$ for $\beta = 0.25, 0.50, 0.75$. These comparisons characterize sensitivity to moments, dimension, spread and initialization; they are not equal-total-cost comparisons. Their principal limitations and the no-third-moment example are summarized in Section 4.2.

## C Detailed assumptions and intermediate results

### C.1 Assumptions, strength, and verification

All asymptotic statements are for fixed $K$ and fixed manifold dimension $d$. A finite burn-in is discarded and indices are relabeled, so the local analysis starts at $m = 1$.

**Assumption C.1** (Local geometric interface and localization)**.** There is a compact neighborhood $\mathcal{U} \subset \mathcal{M}$ containing $x_\star$ in its interior such that:

(G1) $R_x^{-1}(y)$ and $\mathcal{T}_x^y$ are defined for all $x, y \in \mathcal{U}$; the two-point inverse retraction $(x, y) \mapsto R_x^{-1}(y)$ is $C^2$, $R$ is $C^3$ near the zero section, $\mathcal{T}$ is $C^2$, and $\mathcal{T}_x^y$ is an isometry with inverse $\mathcal{T}_y^x$.

(G2) With $\Phi(v) = R_{x_\star}(v)$, there are $c_R, r_R > 0$ such that, for $x = \Phi(v) \in \mathcal{U}$, $R_x(w) \in \mathcal{U}$, and $\|w\| \le r_R$,

$$\left\| R_{x_\star}^{-1}\{R_x(w)\} - v - \mathcal{T}_x^{x_\star} w \right\| \le c_R\{\|v\|^2 \|w\| + \|w\|^2\}. \tag{17}$$

(G3) The theorem is stated for a localized trajectory: after a fixed finite burn-in, the synchronized iterates, all within-round local iterates, and the reported geometric averages remain in $\mathcal{U}$. The condition applies to the recursion (1)–(3) on this neighborhood.

**Assumption C.2** (Local objective regularity)**.** Each $f_k$ is $C^2$ on $\mathcal{U}$, $\nabla f(x_\star) = 0$, and the self-adjoint operator $H_\star = \nabla^2 f(x_\star)$ satisfies

$$\lambda_{\min}(H_\star) = \mu > 0. \tag{18}$$

There is $L < \infty$ such that, for $v = R_{x_\star}^{-1}(x)$,

$$\|\mathcal{T}_x^{x_\star} \nabla f(x) - H_\star v\| \le L \|v\|^2, \tag{19}$$

$$\left\| \mathcal{T}_y^x \nabla f_k(y) - \nabla f_k(x) \right\| \le L d(x, y), \qquad x, y \in \mathcal{U}, \quad k = 1, \dots, K. \tag{20}$$

For plug-in inference, the transported client Hessian fields satisfy

$$\max_{1\le k\le K} \left\| \mathcal{T}_x^{x_\star} \nabla^2 f_k(x) \mathcal{T}_{x_\star}^{x} - \nabla^2 f_k(x_\star) \right\|_{\mathrm{op}} \longrightarrow 0 \quad \text{as } x \to x_\star.$$

**Assumption C.3** (Stochastic oracle and composite covariance)**.** Let $\delta \in (0, 2]$. If $x \in \mathcal{U}$ is measurable with respect to the pre-query filtration $\mathcal{G}$, then the oracle satisfies, almost surely,

$$\mathbb{E}\{G_k(x, \xi) \mid \mathcal{G}\} = \nabla f_k(x), \qquad \mathbb{E}\{\|G_k(x, \xi)\|^{2+\delta} \mid \mathcal{G}\} \le C_G. \tag{21}$$

Writing $\varepsilon_k(x, \xi) = G_k(x, \xi) - \nabla f_k(x)$, assume that there is a deterministic covariance field $S_k(x)$ such that

$$\mathbb{E}\{\varepsilon_k(x, \xi) \otimes \varepsilon_k(x, \xi) \mid \mathcal{G}\} = S_k(x).$$

Define

$$S_k^\star = S_k(x_\star), \qquad S = \sum_{k=1}^{K} p_k^2 S_k^\star.$$

For $x, y \in \mathcal{U}$, let $A(y, x) = \mathcal{T}_y^{x_\star} \mathcal{T}_x^{y}$. There is $L_S < \infty$ such that

$$\|A(y, x) S_k(x) A(y, x)^* - S_k^\star\|_{\mathrm{op}} \le L_S\{d(y, x_\star) + d(x, y)\}. \tag{22}$$

Within a round, the oracle noises are martingale differences in local time and are conditionally uncorrelated across clients. Across rounds, the same conditional mean, moment, and covariance conditions hold relative to the global synchronized filtration. Full conditional independence across rounds is not required.

**Assumption C.4** (Steps, communication, and endpoint scaling)**.** Let $\gamma_m = E_m \eta_m$, $\kappa = \min(\delta, 1)$, and define the raw one-step scale

$$q_m^\circ = \eta_m + \gamma_m^{1+\kappa}.$$

There is an eventually positive nonincreasing comparison sequence $q_m$ and constants $0 < c_q \le C_q < \infty$ such that

$$c_q q_m^\circ \le q_m \le C_q q_m^\circ, \qquad \frac{q_m - q_{m+1}}{q_m} = o(\gamma_m). \tag{23}$$

The effective round steps satisfy

$$\gamma_m \downarrow 0, \qquad \sum_m \gamma_m = \infty, \qquad \sum_m \gamma_m^2 < \infty. \tag{24}$$

The envelope formulation permits integer-valued communication schedules whose raw scale has small discrete jumps. When $q_m^\circ$ itself satisfies the relative-change condition, one may simply take $q_m = q_m^\circ$. With $t_T = \sum_{m=1}^{T} E_m$ and $s_T = \sqrt{t_T}/T$, assume

$$s_T \longrightarrow 0, \tag{25}$$

$$\nu_T := \frac{t_T}{T^2} \sum_{m=1}^{T} E_m^{-1} \longrightarrow \nu \in (0, \infty),$$

$$s_T \sum_{m=1}^{T} \gamma_m \longrightarrow 0, \tag{26}$$

$$s_T \left\{ \frac{\sqrt{q_{T+1}}}{\gamma_T} + \sum_{m=2}^{T} \left( \frac{1}{\gamma_m} - \frac{1}{\gamma_{m-1}} \right) \sqrt{q_m} \right\} \longrightarrow 0, \tag{27}$$

$$s_T^{2+\delta} \sum_{m=1}^{T} E_m^{-(1+\delta/2)} \longrightarrow 0. \tag{28}$$

**Assumption C.5** (Recursive averaging only)**.** For $x \in \mathcal{U}$, the map $z \mapsto \left\| R_{x_\star}^{-1}(z) \right\|^2$ is convex along the local retraction segment used in (38). In addition, with $\bar{q}_m = m^{-1} \sum_{j=1}^{m} q_j$,

$$\frac{\sqrt{t_T}}{T} \sum_{m=1}^{T} \bar{q}_m \longrightarrow 0. \tag{29}$$

**Assumption C.6** (Fresh-score and Hessian messages)**.** Let $\mathcal{F}_m$ contain the history through the synchronized point $x_m$. At round $m$, conditional on $\mathcal{F}_m$, the variables $\zeta_m^k$ and $\omega_m^k$ are fresh, conditionally independent across clients, and independent of the optimization calls in that round. Across rounds, the centered Hessian messages and centered fresh-score outer products form martingale differences with respect to the filtration that reveals the inference messages. Conditional independence across rounds is a sufficient but not necessary construction. The self-adjoint Hessian message satisfies

$$\mathbb{E}\{B_k(x,\omega) \mid x\} = \nabla^2 f_k(x), \qquad \sup_{x\in\mathcal{U},k} \mathbb{E}\,\|B_k(x,\omega)\|_{\mathrm{op}}^{1+\tau} < \infty \tag{30}$$

for some $\tau > 0$. Exact Hessians are included by taking $B_k = \nabla^2 f_k$.

The assumptions are local and modular. The positive-definite Hessian in (18) gives regular identification and an invertible local linearization; nonsingularity of the full Gaussian covariance additionally requires $S \succ 0$. Without regular identification, a different normalization or a non-Gaussian limit may arise. The pairwise bound (20), rather than a bound only relative to $x_\star$, is what controls client drift. On a compact neighborhood it follows from a bounded client Hessian and a smooth transport. The composite covariance condition (22) is stated along the actual two-leg path used by the algorithm; it follows from a locally Lipschitz covariance field and smooth transport.

Assumption C.1(G3) separates entry into a regular basin from inference within it. Its path containment is required on the same small neighborhood used for the contraction in Theorem G.1. A warm start or a global-convergence argument can establish eventual entry, after which the local recursion must remain in the basin. A safeguard supplies a localization device only if its effect on the recursion is also controlled. In particular, convergence-in-probability and weak-limit statements transfer between safeguarded and original trajectories when they agree through round $T$ on events whose probabilities tend to one. The optimization analysis of Huang et al. (2026) addresses the separate convergence question under its own assumptions.

For the common-Hessian Gaussian model, Theorem K.1 derives the sampling law directly on $\mathbb{R}^d$ from the original recursion and any deterministic initial point.

The quadratic linearization in (19) follows, for example, from a locally Lipschitz Riemannian Hessian (or a bounded third covariant derivative) in the target chart; it controls the nonlinear objective remainder in the averaged recursion. The isometric inverse-transport condition is satisfied by Levi–Civita parallel transport along the chosen local geodesic. Common differentiated-retraction transports need not be isometric; using them requires replacing several covariance conjugations by explicit adjoints and is outside the current statement.

The $(2+\delta)$ moment in Assumption C.3 matches the Euclidean Local-SGD inference regime Li et al. (2022). A third moment ($\delta \geq 1$) is not needed for the abstract CLT, but it removes the additional heavy-tail restriction from the simple polynomial schedule corollary below. The Hessian-message assumption is needed only for feasible plug-in inference, not for estimation or the limit distribution with known $H_\star$ and $S$.

### C.2 One-round representation

Let

$$\Delta_m = R_{x_\star}^{-1}(x_m) \in T_{x_\star}\mathcal{M}, \qquad A_{m,e}^k = \mathcal{T}_{x_m}^{x_\star}\mathcal{T}_{x_{m,e}^k}^{x_m}.$$

The second definition records the actual path used by the algorithm.

**Theorem C.7** (Round-wise tangent representation)**.** *Under Assumptions C.1–C.3, after the finite burn-in there are $T_{x_\star}\mathcal{M}$-valued variables $\epsilon_{m+1}$ and $r_{m+1}$ such that*

$$\Delta_{m+1} = (\mathrm{Id} - \gamma_m H_\star)\Delta_m - \gamma_m \epsilon_{m+1} + \gamma_m r_{m+1}. \tag{31}$$

*Conditional on $\mathcal{F}_m$, the start-of-round filtration,*

$$\mathbb{E}(\epsilon_{m+1} \mid \mathcal{F}_m) = 0,$$
$$\mathbb{E}(\epsilon_{m+1} \otimes \epsilon_{m+1} \mid \mathcal{F}_m) = E_m^{-1} S + Q_m, \tag{32}$$
$$\mathbb{E}\left\|\epsilon_{m+1}\right\|^{2+\delta} \mid \mathcal{F}_m \le C E_m^{-(1+\delta/2)}, \tag{33}$$
$$\|Q_m\|_{\mathrm{op}} \le \frac{C}{E_m}\{\|\Delta_m\| + \gamma_m\}, \tag{34}$$
$$\mathbb{E}\left\|r_{m+1}\right\| \mid \mathcal{F}_m \le C\{\gamma_m + \|\Delta_m\|^2\}. \tag{35}$$

*Moreover, with $\kappa = \min(\delta, 1)$, the unscaled remainder $b_{m+1} = \gamma_m r_{m+1}$ satisfies*

$$\mathbb{E}\left\|b_{m+1}\right\|^2 \mid \mathcal{F}_m \le C\{\gamma_m^2 \|\Delta_m\|^4 + \gamma_m^{2+\kappa}\}. \tag{36}$$

The leading noise is the average of $KE_m$ transported client noises. Its covariance is $E_m^{-1}S$ because $K$ is fixed and the weights $p_k$ are absorbed into $S$. The term $Q_m$ contains covariance variation and the actual two-leg transport. The remainder contains client drift, the quadratic objective expansion, and the retraction coordinate error. The exponent $2 + \kappa$ in (36) records the improvement available from a third moment. The proof in Section F uses a truncated moment bound for the retraction error.

### C.3 Polyak–Ruppert limit

Define the inaccessible target-chart average

$$\bar{\Delta}_T = \frac{1}{T}\sum_{m=1}^{T} \Delta_m.$$

It is used only in the proof and to identify the first-order law. Summing the linear recursion and controlling its endpoints gives

$$\sqrt{t_T}\,\bar{\Delta}_T = -H_\star^{-1}\frac{\sqrt{t_T}}{T}\sum_{m=1}^{T}\epsilon_{m+1} + o_p(1).$$

Here (26) removes accumulated drift, (27) removes the Abel endpoint terms, and (28) gives the martingale Lyapunov condition. These are the three uses of the schedule restrictions in Section H.

**Theorem C.8** (Communication-indexed Polyak–Ruppert CLT)**.** *Under Assumptions C.1–C.4,*

$$\sqrt{t_T}\,\bar{\Delta}_T \xrightarrow{d} N(0, \Sigma_\star), \qquad \Sigma_\star = \nu H_\star^{-1} S H_\star^{-1}.$$

Each round receives weight $1/T$ in the average, while its leading noise covariance is $S/E_m$ by (32). Consequently the variance of the leading martingale sum is governed by

$$\nu_T = \left(\frac{1}{T}\sum_{m=1}^{T} E_m\right)\left(\frac{1}{T}\sum_{m=1}^{T} E_m^{-1}\right).$$

Thus $H_\star$ and $S$ describe local curvature and oracle noise, and $\nu$ records the effect of averaging equally across rounds with different numbers of local updates. Since Cauchy–Schwarz gives $(\sum E_m)(\sum E_m^{-1}) \ge T^2$, one has $\nu \ge 1$ whenever the limit exists. Thus $\nu = 1$ matches the first-order covariance of a fully synchronized or centralized averaged trajectory with the same weighted client oracle and optimization clock, whereas $\nu > 1$ quantifies the loss caused by an increasing communication interval.

### C.4 Polynomial schedules

**Corollary C.9** (Polynomial communication intervals)**.** *Suppose, for constants $c_E, c_\gamma > 0$,*

$$E_m = \max\{1, \lfloor c_E m^\beta \rfloor\}, \qquad \gamma_m = c_\gamma m^{-\alpha}, \qquad 0 \le \beta < 1, \qquad \frac{1}{2} < \alpha < 1.$$

*Let*

$$r = \min\{\alpha + \beta, \alpha(1+\kappa)\}, \qquad \kappa = \min(\delta, 1).$$

*Assumption C.4 holds whenever*

$$\alpha > \frac{1+\beta}{2}, \qquad r > 2\alpha + \beta - 1. \tag{37}$$

*The Lyapunov condition then holds automatically, and*

$$\nu = \frac{1}{1-\beta^2}.$$

*If $\delta \geq 1$, the second inequality in* (37) *is automatic under $\alpha < 1$ and $\beta < 1$; therefore every $0 \leq \beta < 1$ is covered by*

$$\frac{1+\beta}{2} < \alpha < 1.$$

*If $0 < \delta < 1$, the exact additional restriction is needed only when $\beta > \alpha\delta$; in that case it is equivalent to*

$$\alpha(1-\delta) + \beta < 1.$$

*Constant and slowly varying intervals have $\nu = 1$ under the analogous remainder conditions.*

### C.5 Recursive and proportional-tail guarantees

The *recursive retraction average* starts at $\widetilde{x}_b = x_b$ and updates

$$\widetilde{x}_{m+1} = R_{\widetilde{x}_m}\left(\frac{1}{m-b+2} R_{\widetilde{x}_m}^{-1}(x_{m+1})\right), \quad m \geq b. \tag{38}$$

This is the geometric streaming average of Tripuraneni et al. (2018). The additional condition for its sampling law is given with Theorem C.10.

**Theorem C.10** (Recursive retraction average)**.** *Under the assumptions of Theorem C.8 and Assumption C.5,*

$$\sqrt{t_T}\left\{R_{x_\star}^{-1}(\widetilde{x}_T) - \bar{\Delta}_T\right\} \xrightarrow{\mathbb{P}} 0, \tag{39}$$

*and hence*

$$\sqrt{t_T}\, R_{x_\star}^{-1}(\widetilde{x}_T) \xrightarrow{d} N(0, \Sigma_\star). \tag{40}$$

*For the power schedules in Theorem C.9,* (29) *follows from* (37)*.*

The recursive averaging map and its local Taylor structure are inherited from Tripuraneni et al. (2018). What changes here is the state rate and the statistical clock: the synchronized errors are of order $\{\eta_m + \gamma_m^{1+\kappa}\}^{1/2}$, while the limit is normalized by $\sqrt{t_T}$ rather than $\sqrt{T}$. Assumption C.5 is therefore stated separately instead of being bundled into the geometry needed for the RFedAGS trajectory. The proof is in Section I.2.

**Corollary C.11** (Proportional-tail inference)**.** *Fix $\rho \in [0, 1)$, let $b_T = \lfloor \rho T \rfloor$, and define*

$$\bar{\Delta}_{T,\rho} = \frac{1}{T - b_T} \sum_{m=b_T+1}^{T} \Delta_m.$$

*Put $n_{T,\rho} = T - b_T$ and $s_{T,\rho} = \sqrt{t_T}/n_{T,\rho}$. Suppose the assumptions of Theorem C.8 hold and, on the retained range,*

$$s_{T,\rho} \sum_{m=b_T+1}^{T} \gamma_m \longrightarrow 0,$$

$$s_{T,\rho}\left\{\frac{\sqrt{q_{b_T+1}}}{\gamma_{b_T+1}} + \frac{\sqrt{q_{T+1}}}{\gamma_T} + \sum_{m=b_T+2}^{T}\left(\frac{1}{\gamma_m} - \frac{1}{\gamma_{m-1}}\right)\sqrt{q_m}\right\} \longrightarrow 0,$$

$$s_{T,\rho}^{2+\delta} \sum_{m=b_T+1}^{T} E_m^{-(1+\delta/2)} \longrightarrow 0.$$

*If*

$$\nu_{T,\rho} = \frac{t_T}{(T - b_T)^2} \sum_{m=b_T+1}^{T} E_m^{-1} \longrightarrow \nu_\rho \in (0, \infty), \tag{41}$$

*then*

$$\sqrt{t_T}\, \bar{\Delta}_{T,\rho} \xrightarrow{d} N\big(0, \nu_\rho H_\star^{-1} S H_\star^{-1}\big)\,.$$

*The fixed-anchor tail average has the same limit; so does the recursive tail average under Assumption C.5 on the retained range. The conclusions of Theorems 3.3, 3.4 and C.12 hold for tail fresh-score and Hessian averages when $\widehat{\nu}_T$ is replaced by* (41)*.*

*For $E_m \sim c_E m^\beta$ with $0 \le \beta < 1$ and the power conditions in Theorem C.9,*

$$\nu_\rho = \frac{1 - \rho^{1-\beta}}{(1-\rho)^2(1-\beta^2)}. \tag{42}$$

*At $\rho = 0$ this reduces to $(1 - \beta^2)^{-1}$.*

A proportional tail removes the early server points, whose initialization error can persist at finite horizons. It also changes which round noises enter the average, so its covariance must use the retained range. With constant intervals, (42) gives $\nu_\rho = 1/(1-\rho)$: discarding a fraction $\rho$ increases the first-order variance on the total optimization clock by this factor. With growing intervals, later rounds average more local noise terms, and (42) accounts for both their smaller noise variance and the reduced number of retained rounds. This is why the finite implementation uses (5) when forming confidence regions. The proof is in Section I.3.

**Corollary C.12** (Smooth-functional inference)**.** *Under the corresponding estimator assumptions in Theorem 3.3, let $\psi : \mathcal{M} \to \mathbb{R}^r$ be $C^1$ near $x_\star$ and suppose $D\psi_{x_\star} \Sigma_\star D\psi_{x_\star}^*$ is nonsingular. Then*

$$\sqrt{t_T}\{\psi(\widehat{x}_T) - \psi(x_\star)\} \xrightarrow{d} N\big(0, D\psi_{x_\star} \Sigma_\star D\psi_{x_\star}^*\big)\,, \tag{43}$$

*and replacing the covariance by $D\psi_{\widehat{x}_T} \widehat{\Sigma}_T D\psi_{\widehat{x}_T}^*$ gives an asymptotically valid $r$-dimensional Wald region.*

### C.6 Schedule factors and communication frequency

**Proposition C.13.** *Recall the communication frequency $\mathrm{ACF}_T = T/t_T$.*

*(i) For every schedule, $\nu_T \ge 1$. If $E_m \equiv E_0$, then $\nu_T = 1$ and $\mathrm{ACF}_T = E_0^{-1}$.*

*(ii) If $E_m \sim c_E L(m)$, where $L$ is positive, eventually monotone, and slowly varying ($L(cx)/L(x) \to 1$ for every fixed $c > 0$) Bingham et al. (1987), then $\nu_T \to 1$ and $\mathrm{ACF}_T \sim \{c_E L(T)\}^{-1}$. Thus, at the level of the schedule factor, logarithmic or iterated-logarithmic intervals permit a vanishing communication frequency without first-order variance inflation; use in the CLT still requires the remainder conditions in Assumption C.4.*

*(iii) If $E_m \sim c_E m^\beta$ with $0 < \beta < 1$, then*

$$\nu_T \longrightarrow \frac{1}{1-\beta^2}, \qquad \mathrm{ACF}_T \sim \frac{1+\beta}{c_E} T^{-\beta}.$$

*(iv) The Local-SGD product condition*

$$\sup_{T \ge 1} \frac{1}{T^2} \left( \sum_{m=1}^{T} E_m^{1+\delta/2} \right) \left( \sum_{m=1}^{T} E_m^{-(1+\delta/2)} \right) < \infty \tag{44}$$

*implies the Lyapunov condition* (28)*. It does not by itself verify the manifold remainder or Abel-boundary conditions* (26)–(27)*.*

# D PROOFS FOR SCHEDULES AND ALGORITHMIC REDUCTIONS

*Proof of Theorem C.13.* Cauchy–Schwarz gives

$$\left(\sum_{m=1}^{T} E_m\right)\left(\sum_{m=1}^{T} E_m^{-1}\right) \geq T^2,$$

so $\nu_T \geq 1$. The constant-interval identities follow by direct substitution.

If $E_m \sim c_E L(m)$ with $L$ slowly varying and eventually monotone, integral comparison and Karamata's theorem (Bingham et al., 1987, Theorem 1.5.11) give

$$t_T \sim c_E T L(T), \qquad \sum_{m=1}^{T} E_m^{-1} \sim \frac{T}{c_E L(T)}.$$

Their product divided by $T^2$ converges to one, while $T/t_T \sim \{c_E L(T)\}^{-1}$. Taking $L(m) = \log^a(m+1)$ or $L(m) = \log^a\log(m+e)$ gives the stated examples.

If $E_m \sim c_E m^\beta$, integral comparison yields

$$t_T \sim \frac{c_E}{1+\beta} T^{1+\beta}, \qquad \sum_{m=1}^{T} E_m^{-1} \sim \frac{1}{c_E(1-\beta)} T^{1-\beta}.$$

The asserted formulas for $\nu_T$ and $\mathrm{ACF}_T$ follow.

Finally, Jensen's inequality gives

$$t_T^{1+\delta/2} = \left(\sum_{m=1}^{T} E_m\right)^{1+\delta/2} \leq T^{\delta/2} \sum_{m=1}^{T} E_m^{1+\delta/2}.$$

Therefore

$$\left(\frac{\sqrt{t_T}}{T}\right)^{2+\delta} \sum_{m=1}^{T} E_m^{-(1+\delta/2)} \leq T^{-\delta/2} \frac{1}{T^2} \left(\sum_{m=1}^{T} E_m^{1+\delta/2}\right)\left(\sum_{m=1}^{T} E_m^{-(1+\delta/2)}\right),$$

which tends to zero under (44). This argument concerns only the Lyapunov term and hence does not imply either of the separately stated remainder conditions. □

## D.1 POLYNOMIAL SCHEDULES

*Proof of Theorem C.9.* For $\beta > 0$, integer rounding does not affect the asymptotic equivalence $E_m \sim c_E m^\beta$; for $\beta = 0$, write the resulting constant interval as $E_0 = \max\{1, \lfloor c_E \rfloor\}$. In either case $\gamma_m = c_\gamma m^{-\alpha}$. Then

$$q_m^\circ = \eta_m + \gamma_m^{1+\kappa} \asymp m^{-r}, \qquad q_m = m^{-r}, \qquad r = \min\{\alpha+\beta, \alpha(1+\kappa)\}. \tag{45}$$

Because $1/2 < \alpha < 1$, the stochastic-approximation sums in (24) hold. Moreover, $(q_m - q_{m+1})/q_m = O(m^{-1}) = o(m^{-\alpha})$, which verifies (23). This smooth comparison envelope is the reason integer rounding in $E_m$ does not create a false pointwise regularity requirement on the raw scale $q_m^\circ$.

For $\beta > 0$, Karamata's sum theorem, or direct integral comparison, gives

$$t_T \sim \frac{c_E}{1+\beta} T^{1+\beta}, \qquad \sum_{m=1}^{T} E_m^{-1} \sim \frac{1}{c_E(1-\beta)} T^{1-\beta}.$$

For $\beta = 0$, $t_T = E_0 T$ and $\sum_{m=1}^{T} E_m^{-1} = T/E_0$. Consequently, in all cases $s_T \asymp T^{(\beta-1)/2} \to 0$ and

$$\nu_T = \frac{t_T}{T^2} \sum_{m=1}^{T} E_m^{-1} \longrightarrow \frac{1}{(1+\beta)(1-\beta)} = \frac{1}{1-\beta^2}.$$

Also,

$$s_T \sum_{m=1}^{T} \gamma_m = O\{T^{(1+\beta)/2-\alpha}\},$$

which tends to zero exactly under the first strict inequality in (37).

For the Abel boundary, monotonicity and the mean-value theorem give

$$\frac{1}{\gamma_m} - \frac{1}{\gamma_{m-1}} = O(m^{\alpha-1}).$$

Using (45), both the terminal term and, when divergent, the sum in (27) are of order at most $T^{\alpha-r/2}$, up to a logarithm at the borderline exponent. Hence their product with $s_T$ is

$$O\{T^{(\beta-1)/2+\alpha-r/2}\},$$

which vanishes under $r > 2\alpha + \beta - 1$. If the displayed sum is bounded, the weaker factor $s_T \to 0$ already suffices, so the same inequality is a valid uniform sufficient condition.

For the Lyapunov term, set $a = \beta(1 + \delta/2)$. If $a < 1$, then

$$s_T^{2+\delta} \sum_{m=1}^{T} E_m^{-(1+\delta/2)} = O\left(T^{(\beta-1)(2+\delta)/2} T^{1-\beta(2+\delta)/2}\right) = O(T^{-\delta/2}).$$

If $a = 1$, the same calculation has only an additional logarithm; if $a > 1$, the sum is bounded and the negative power $T^{-(1-\beta)(2+\delta)/2}$ remains. Thus (28) is automatic.

When $\delta \geq 1$, $\kappa = 1$ and $r = \min(\alpha + \beta, 2\alpha)$. If $\beta \leq \alpha$, the second inequality in (37) becomes $\alpha < 1$; if $\beta > \alpha$, it becomes $\beta < 1$. Both are already imposed. Hence $(1 + \beta)/2 < \alpha < 1$ covers every $\beta < 1$. When $0 < \delta < 1$, if $\beta \leq \alpha\delta$, then $r = \alpha + \beta$ and the second inequality again reduces to $\alpha < 1$. If $\beta > \alpha\delta$, then $r = \alpha(1 + \delta)$, and the inequality is equivalent to $\alpha(1 - \delta) + \beta < 1$.

It remains to verify (29). Since $r \geq \alpha > (1 + \beta)/2$, one has, according as $r < 1$, $r = 1$, or $r > 1$,

$$\bar{q}_m = O(m^{-r}), \qquad \bar{q}_m = O(m^{-1} \log m), \qquad \text{or} \qquad \bar{q}_m = O(m^{-1}).$$

Consequently, $s_T \sum_{m \leq T} \bar{q}_m \to 0$ in all three cases. This proves the final assertion for the recursive average. Constant and slowly varying $E_m$ give $\nu = 1$ by the same sum calculation, provided the abstract boundary and remainder conditions are retained. □

### D.2 Algorithmic reductions

*Proof of Theorem A.1.* For item (i), equal weights give the client average in RFedAGS. That algorithm transmits the sum of the transported local gradients and applies its server step to that sum. Equations (2)–(3) transmit the corresponding average $U_{m+1}^k$ and multiply it by $\gamma_m = E_m \eta_m$. The product is identical, so the synchronized trajectory is unchanged.

For item (ii), in Euclidean space every tangent space is $\mathbb{R}^d$, the transport is the identity, and $R_x(v) = x + v$. Summing the local recursion gives

$$x_{m,E_m}^k = x_m - \eta_m \sum_{e=0}^{E_m-1} G_{m,e+1}^k.$$

The server update is therefore

$$\begin{aligned} x_{m+1} &= x_m - E_m \eta_m \sum_{k=1}^{K} p_k \frac{1}{E_m} \sum_{e=0}^{E_m-1} G_{m,e+1}^k \\ &= \sum_{k=1}^{K} p_k x_{m,E_m}^k, \end{aligned}$$

which is full-participation Local SGD, equivalently the usual full-participation FedAvg update with common local step size.

For item (iii), if $K = 1$ and $E_m = 1$, then $\gamma_m = \eta_m$ and $U_{m+1} = G_{m,1}^1$. Thus $x_{m+1} = R_{x_m}\{-\eta_m G_{m,1}^1\}$, and (38) is exactly the streaming retraction average used for averaged RSGD. □

# E GEOMETRIC AND PROBABILISTIC PRELIMINARIES

All constants below may change from line to line and depend on the fixed neighborhood $\mathcal{U}$, the manifold dimension, $K$, and the constants in the assumptions, but not on $m$, $e$, or $T$. Finite burn-in terms are omitted. The symbol $\|\cdot\|$ denotes the Riemannian norm in the relevant tangent space and the associated operator norm.

## E.1 CONSEQUENCES OF THE GEOMETRIC INTERFACE

**Lemma E.1** (Metric and chart equivalence)**.** *Under Assumption C.1, after shrinking $\mathcal{U}$ if necessary, there are constants $0 < c < C < \infty$ such that, for $x, y \in \mathcal{U}$,*

$$c\, d(x,y) \le \left\|R_x^{-1}(y)\right\| \le C\, d(x,y), \qquad c\, d(x,x_\star) \le \left\|R_{x_\star}^{-1}(x)\right\| \le C\, d(x,x_\star).$$

*For every $q \in [1, 2+\delta]$ and every tangent vector $W$ for which $R_x(W) \in \mathcal{U}$,*

$$d\{x, R_x(W)\}^q \le C\, \|W\|^q\, \mathbf{1}\{\|W\| \le r_R\} + C\mathbf{1}\{\|W\| > r_R\}. \tag{46}$$

*Proof.* The first two inequalities follow from the inverse function theorem and compactness: the derivatives of each local retraction chart and its inverse are bounded above and bounded away from singularity. For (46), local Lipschitz continuity of $R_x$ gives $d\{x, R_x(W)\} \le C\, \|W\|$ on $\|W\| \le r_R$. On the complement, both points are in the compact set $\mathcal{U}$, so the distance is at most $\operatorname{diam}(\mathcal{U})$. Raising the two bounds to power $q$ proves the claim. □

**Lemma E.2** (Coordinate update with a low-moment envelope)**.** *Let $v = R_{x_\star}^{-1}(x)$ and*

$$\mathcal{R}(v,w) = R_{x_\star}^{-1}\{R_x(w)\} - v - \mathcal{T}_x^{x_\star} w.$$

*Under Assumption C.1,*

$$\|\mathcal{R}(v,w)\| \le C\{\|v\|^2\, \|w\| + \|w\|^2\}, \qquad \|w\| \le r_R, \tag{47}$$

$$\|\mathcal{R}(v,w)\| \le C(1 + \|w\|), \qquad R_x(w) \in \mathcal{U}. \tag{48}$$

*If $V$ satisfies $\mathbb{E}(\|V\|^{2+\delta} \mid \mathcal{F}) \le C$ and $w = -\gamma V$, then, for $\kappa = \min(\delta, 1)$,*

$$\mathbb{E}\{\|\mathcal{R}(v, -\gamma V)\| \mid \mathcal{F}\} \le C\{\gamma\, \|v\|^2 + \gamma^2\}, \tag{49}$$

$$\mathbb{E}\{\|\mathcal{R}(v, -\gamma V)\|^2 \mid \mathcal{F}\} \le C\{\gamma^2\, \|v\|^4 + \gamma^{2+\kappa}\}. \tag{50}$$

*Proof.* Equation (47) is exactly (17). For (48), the first two coordinate terms are bounded on the compact chart and $\mathcal{T}_x^{x_\star}$ is an isometry, so $\|\mathcal{R}(v,w)\| \le C + \|w\|$.

Let $A = \{\gamma\, \|V\| \le r_R\}$. On $A$, (47) gives

$$\|\mathcal{R}(v, -\gamma V)\| \le C\{\gamma\, \|v\|^2\, \|V\| + \gamma^2\, \|V\|^2\}.$$

Taking conditional expectations and using the uniform second moment proves the contribution $C(\gamma\, \|v\|^2 + \gamma^2)$. On $A^c$, (48) and Markov's inequality give

$$\mathbb{E}\{(1 + \gamma\, \|V\|)\mathbf{1}_{A^c} \mid \mathcal{F}\} \le C\gamma^{2+\delta}\mathbb{E}(\|V\|^{2+\delta} \mid \mathcal{F}),$$

which is $O(\gamma^{2+\delta})$ and hence no larger than $O(\gamma^2)$ eventually. This proves (49).

For the second moment, combining the local and global bounds on the events $A$ and $A^c$ gives the pointwise envelope

$$\|\mathcal{R}(v, -\gamma V)\|^2 \le C\gamma^2\, \|v\|^4\, \|V\|^2 + C \min\{\gamma^4\, \|V\|^4,\, 1 + \gamma^2\, \|V\|^2\}.$$

The first term has conditional expectation at most $C\gamma^2\, \|v\|^4$. If $0 < \delta < 1$, the second term is bounded pointwise by $C\gamma^{2+\delta}\, \|V\|^{2+\delta}$. If $\delta \ge 1$, use only the third moment and $\min(z^4, z^2 + 1) \le Cz^3$ for $z \ge 0$, obtaining $C\gamma^3\mathbb{E}(\|V\|^3 \mid \mathcal{F})$. These two cases equal $C\gamma^{2+\kappa}$ with $\kappa = \min(\delta, 1)$, proving (50). □

*Remark* E.3. For the exponential map with Levi–Civita parallel transport, (17) follows from the Jacobi-field expansion of the differential of the exponential map. For a general retraction and vector transport it is a checkable second-order compatibility condition. A second-order retraction alone does not imply it for an arbitrary transport; this is why the condition is stated on the pair $(R, \mathcal{T})$.

**Lemma E.4** (Inverse-error expansion). *If $x \to x_\star$ inside $\mathcal{U}$ and $\Delta = R_{x_\star}^{-1}(x)$, then*

$$\mathcal{T}_x^{x_\star} R_x^{-1}(x_\star) = -\Delta + O(\|\Delta\|^2). \tag{51}$$

*Proof.* Define $F(v) = \mathcal{T}^{x_\star}_{R_{x_\star}(v)} R^{-1}_{R_{x_\star}(v)}(x_\star)$. Smoothness gives $F(0) = 0$. Differentiating the identity $R_{R_{x_\star}(v)}\{R^{-1}_{R_{x_\star}(v)}(x_\star)\} = x_\star$ at $v = 0$ shows $DF(0) = -\mathrm{Id}$ because $DR_x(0) = \mathrm{Id}$ and $\mathcal{T}^{x_\star}_{x_\star} = \mathrm{Id}$. A second-order Taylor expansion on the compact chart yields $F(v) = -v + O(\|v\|^2)$. □

### E.2 Within-round filtrations and moment inequality

Let $\mathcal{F}_m$ contain the complete history through the synchronized point $x_m$. Inside round $m$, let $\mathcal{F}_{m,e}$ contain all client samples through local time $e$ and all local points $x^k_{m,j}$, $j \le e$. Then $x^k_{m,e}$ and $A^k_{m,e}$ are $\mathcal{F}_{m,e}$-measurable, while

$$\zeta^k_{m,e+1} = G_k(x^k_{m,e}, \xi^k_{m,e+1}) - \nabla f_k(x^k_{m,e})$$

has conditional mean zero given $\mathcal{F}_{m,e}$.

**Lemma E.5** (Finite-dimensional martingale moment bound). *Let $q \in [2, 4]$ and let $D_1, \dots, D_N$ be martingale differences in a fixed-dimensional Euclidean space with respect to a filtration $(\mathcal{G}_j)_{j=0}^N$. The standard Burkholder–Rosenthal inequality (Hall & Heyde, 1980, Theorem 2.12) gives*

$$\mathbb{E}\left\|\sum_{j=1}^N D_j\right\|^q \le C_q \left\{ \mathbb{E}\left(\sum_{j=1}^N \mathbb{E}(\|D_j\|^2 \mid \mathcal{G}_{j-1})\right)^{q/2} + \sum_{j=1}^N \mathbb{E}\,\|D_j\|^q \right\}.$$

*Proof.* Choose a fixed orthonormal basis. Apply the scalar Burkholder–Rosenthal inequality to every coordinate and use norm equivalence in fixed dimension. The resulting constant depends only on $q$ and $d$. □

## F Within-round drift and proof of the tangent representation

**Lemma F.1** (Within-round displacement). *Under Assumptions C.1 and C.3, for every $q \in [1, 2+\delta]$,*

$$\left\{\mathbb{E}\left(d(x^k_{m,e}, x_m)^q \mid \mathcal{F}_m\right)\right\}^{1/q} \le Ce\eta_m, \qquad e = 0, \dots, E_m.$$

*Consequently,*

$$\frac{1}{E_m}\sum_{e=0}^{E_m-1} \left\{\mathbb{E}\left(d(x^k_{m,e}, x_m)^q \mid \mathcal{F}_m\right)\right\}^{1/q} \le C\gamma_m. \tag{52}$$

*Proof.* For one local step, apply (46) with $W = -\eta_m G^k_{m,j+1}$. On the event $\eta_m \left\|G^k_{m,j+1}\right\| \le r_R$, the $q$th power of the displacement is at most $C\eta_m^q \left\|G^k_{m,j+1}\right\|^q$. On the complement it is bounded by a constant, while Markov's inequality gives

$$\mathbb{P}\{\eta_m \left\|G^k_{m,j+1}\right\| > r_R \mid \mathcal{F}_{m,j}\} \le C\eta_m^q \mathbb{E}(\left\|G^k_{m,j+1}\right\|^q \mid \mathcal{F}_{m,j}).$$

The conditional oracle moment is uniformly bounded, so

$$\left\{\mathbb{E}\left(d(x^k_{m,j+1}, x^k_{m,j})^q \mid \mathcal{F}_{m,j}\right)\right\}^{1/q} \le C\eta_m.$$

By the triangle inequality for $d$ and conditional Minkowski,

$$\left\{\mathbb{E}(d(x^k_{m,e}, x_m)^q \mid \mathcal{F}_m)\right\}^{1/q} \le \sum_{j=0}^{e-1} \left\{\mathbb{E}(d(x^k_{m,j+1}, x^k_{m,j})^q \mid \mathcal{F}_m)\right\}^{1/q} \le Ce\eta_m.$$

Averaging $e = 0, \dots, E_m - 1$ and using $E_m\eta_m = \gamma_m$ proves (52). □

Recall the centered local noise

$$\zeta_{m,e+1}^k = G_{m,e+1}^k - \nabla f_k(x_{m,e}^k)$$

and the two-leg map $A_{m,e}^k = \mathcal{T}_{x_m}^{x_\star} \mathcal{T}_{x_{m,e}^k}^{x_m}$. Set

$$\epsilon_{m+1} = \frac{1}{E_m} \sum_{k=1}^K p_k \sum_{e=0}^{E_m-1} A_{m,e}^k \zeta_{m,e+1}^k,$$

$$d_{m+1} = \frac{1}{E_m} \sum_{k=1}^K p_k \sum_{e=0}^{E_m-1} \left\{ A_{m,e}^k \nabla f_k(x_{m,e}^k) - \mathcal{T}_{x_m}^{x_\star} \nabla f_k(x_m) \right\}.$$

Then, exactly,

$$\mathcal{T}_{x_m}^{x_\star} U_{m+1} = \mathcal{T}_{x_m}^{x_\star} \nabla f(x_m) + \epsilon_{m+1} + d_{m+1}. \tag{53}$$

**Lemma F.2** (Noise and client-drift bounds)**.** *Under Assumptions C.1–C.3,*

$$\mathbb{E}(\epsilon_{m+1} \mid \mathcal{F}_m) = 0, \tag{54}$$

$$\mathbb{E} \left\| \epsilon_{m+1} \right\|^{2+\delta} \mid \mathcal{F}_m \le C E_m^{-(1+\delta/2)}, \tag{55}$$

$$\mathbb{E} \left\| d_{m+1} \right\|^q \mid \mathcal{F}_m \le C \gamma_m^q, \qquad 1 \le q \le 2+\delta. \tag{56}$$

*Furthermore,*

$$\mathbb{E}(\epsilon_{m+1} \otimes \epsilon_{m+1} \mid \mathcal{F}_m) = E_m^{-1} S + Q_m, \qquad \left\| Q_m \right\|_{\mathrm{op}} \le \frac{C}{E_m} \{ \left\| \Delta_m \right\| + \gamma_m \}.$$

*Proof.* For local time $e$, group the client contributions as

$$D_{m,e+1} = \sum_{k=1}^K p_k A_{m,e}^k \zeta_{m,e+1}^k.$$

The vector $D_{m,e+1}$ is measurable after the time-$e+1$ oracle calls and has conditional mean zero given $\mathcal{F}_{m,e}$. Thus $D_{m,1}, \dots, D_{m,E_m}$, rather than an arbitrary ordering of all client-level calls, form the required martingale-difference sequence. This grouping avoids any unstated conditional-independence assumption across clients. Iterated conditioning proves (54).

Take $q = 2+\delta$. Isometry, the fixed number of clients, conditional Minkowski, and the oracle moment bound give $\mathbb{E}(\left\| D_{m,e+1} \right\|^q \mid \mathcal{F}_{m,e}) \le C$. Therefore Lemma E.5 gives

$$\begin{aligned} \mathbb{E} \left\| \epsilon_{m+1} \right\|^q \mid \mathcal{F}_m &\le \frac{C}{E_m^q} \left\{ E_m^{q/2} + E_m \right\} \\ &\le C E_m^{-q/2} = C E_m^{-(1+\delta/2)}. \end{aligned}$$

This is (55).

For the drift, isometry and (20) imply

$$\left\| A_{m,e}^k \nabla f_k(x_{m,e}^k) - \mathcal{T}_{x_m}^{x_\star} \nabla f_k(x_m) \right\| \le L d(x_{m,e}^k, x_m).$$

Conditional Minkowski and (52) therefore yield $(\mathbb{E} \left\| d_{m+1} \right\|^q \mid \mathcal{F}_m)^{1/q} \le C \gamma_m$, proving (56).

For the covariance, cross-products from distinct local times have zero conditional expectation, and the assumed conditional uncorrelatedness removes cross-client terms within a local time. Hence

$$\mathbb{E}(\epsilon_{m+1} \otimes \epsilon_{m+1} \mid \mathcal{F}_m) = \frac{1}{E_m^2} \sum_{k=1}^K p_k^2 \sum_{e=0}^{E_m-1} \mathbb{E} \left\{ A_{m,e}^k S_k(x_{m,e}^k) (A_{m,e}^k)^* \mid \mathcal{F}_m \right\}.$$

Subtract $E_m^{-1} S$. By (22), the norm of each summand difference is at most

$$C \{ d(x_m, x_\star) + d(x_{m,e}^k, x_m) \}.$$

Metric–chart equivalence, Theorem E.1, bounds the first distance by $C \left\| \Delta_m \right\|$. Summing the second and applying (52) gives

$$\left\| Q_m \right\|_{\mathrm{op}} \le \frac{C}{E_m} \{ \left\| \Delta_m \right\| + \gamma_m \}.$$

This completes the proof. □

*Proof of Theorem C.7.* By (19), the transported objective gradient has remainder $\rho_m^f$ defined by

$$\mathcal{T}_{x_m}^{x_\star}\nabla f(x_m) = H_\star\Delta_m + \rho_m^f, \qquad \left\|\rho_m^f\right\| \le L\left\|\Delta_m\right\|^2. \tag{57}$$

The server update and Theorem E.2 give the exact coordinate identity

$$\begin{aligned}\Delta_{m+1} &= \Delta_m - \gamma_m\mathcal{T}_{x_m}^{x_\star}U_{m+1} + \mathcal{R}(\Delta_m, -\gamma_m U_{m+1})\\ &= (\mathrm{Id} - \gamma_m H_\star)\Delta_m - \gamma_m\epsilon_{m+1} + b_{m+1},\end{aligned}$$

where, using (53) and (57),

$$b_{m+1} = -\gamma_m d_{m+1} - \gamma_m\rho_m^f + \mathcal{R}(\Delta_m, -\gamma_m U_{m+1}).$$

Set $r_{m+1} = b_{m+1}/\gamma_m$. This proves (31).

It remains to bound the remainder. From (53), the compactness of $\mathcal{U}$, (55), and (56),

$$\mathbb{E}(\|U_{m+1}\|^{2+\delta} \mid \mathcal{F}_m) \le C.$$

Indeed, the transported population gradient is bounded on $\mathcal{U}$, the noise average has a uniformly bounded $(2+\delta)$ moment, and $\gamma_m$ is bounded. Thus (49), (56), and (57) imply

$$\begin{aligned}\mathbb{E}(\|b_{m+1}\| \mid \mathcal{F}_m) &\le C\gamma_m^2 + C\gamma_m\left\|\Delta_m\right\|^2 + C\{\gamma_m\left\|\Delta_m\right\|^2 + \gamma_m^2\}\\ &\le C\gamma_m\{\gamma_m + \left\|\Delta_m\right\|^2\}.\end{aligned}$$

Dividing by $\gamma_m$ proves (35).

For the second moment, use $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, (56) with $q=2$, and (50):

$$\begin{aligned}\mathbb{E}(\|b_{m+1}\|^2 \mid \mathcal{F}_m) &\le C\gamma_m^2\mathbb{E}(\|d_{m+1}\|^2 \mid \mathcal{F}_m) + C\gamma_m^2\left\|\Delta_m\right\|^4\\ &\quad + C\{\gamma_m^2\left\|\Delta_m\right\|^4 + \gamma_m^{2+\kappa}\}\\ &\le C\{\gamma_m^2\left\|\Delta_m\right\|^4 + \gamma_m^{2+\kappa}\},\end{aligned}$$

where the smaller term $\gamma_m^4$ is absorbed. This is (36). The noise mean, moment, covariance, and $Q_m$ bounds are Theorem F.2. The theorem follows. □

## G MEAN-SQUARE RATE

**Lemma G.1** (One-step contraction)**.** *Under Assumptions C.1–C.4 and the conclusions of Theorem C.7, the neighborhood $\mathcal{U}$ may be shrunk around $x_\star$ so that, for all sufficiently large $m$,*

$$\mathbb{E}(\|\Delta_{m+1}\|^2 \mid \mathcal{F}_m) \le (1 - c\gamma_m)\left\|\Delta_m\right\|^2 + C\gamma_m q_m \tag{58}$$

*for constants $c, C > 0$.*

*Proof.* Write $A_m = (\mathrm{Id} - \gamma_m H_\star)\Delta_m$ and $b_{m+1} = \gamma_m r_{m+1}$. From (31),

$$\Delta_{m+1} = A_m - \gamma_m\epsilon_{m+1} + b_{m+1}.$$

Because $H_\star \succeq \mu\mathrm{Id}$ and is bounded,

$$\left\|A_m\right\|^2 \le (1 - 2\mu\gamma_m + C\gamma_m^2)\left\|\Delta_m\right\|^2. \tag{59}$$

The cross term between $A_m$ and $\epsilon_{m+1}$ has conditional expectation zero. Also, (32) and (34) imply, after shrinking $\mathcal{U}$ and taking $m$ large,

$$\mathbb{E}(\|\epsilon_{m+1}\|^2 \mid \mathcal{F}_m) \le C/E_m.$$

Thus the pure noise contribution is at most $C\gamma_m^2/E_m = C\gamma_m\eta_m$.

For the cross term with $b_{m+1}$, (35) gives

$$\begin{aligned}2\left|\mathbb{E}\{\langle A_m, b_{m+1}\rangle \mid \mathcal{F}_m\}\right| &\le C\left\|\Delta_m\right\|\gamma_m\{\gamma_m + \left\|\Delta_m\right\|^2\}\\ &\le \frac{\mu}{4}\gamma_m\left\|\Delta_m\right\|^2 + C\gamma_m^3 + C\gamma_m\left\|\Delta_m\right\|^3.\end{aligned}$$

The last term is bounded by $(\mu/4)\gamma_m \|\Delta_m\|^2$ if the chart radius is chosen small enough. For the noise–remainder cross term, Young's inequality gives

$$2\gamma_m \mathbb{E}\{\|\epsilon_{m+1}\| \, \|b_{m+1}\| \mid \mathcal{F}_m\} \le \frac{1}{2}\gamma_m^2 \mathbb{E}(\|\epsilon_{m+1}\|^2 \mid \mathcal{F}_m) + 2\mathbb{E}(\|b_{m+1}\|^2 \mid \mathcal{F}_m). \tag{60}$$

By (36), the last term is at most $C\gamma_m^2 \|\Delta_m\|^4 + C\gamma_m^{2+\kappa}$. The first part is absorbed into the contraction after shrinking $\mathcal{U}$ and using $\gamma_m \to 0$. Combining (59)–(60) yields

$$\mathbb{E}(\|\Delta_{m+1}\|^2 \mid \mathcal{F}_m) \le (1 - c\gamma_m) \|\Delta_m\|^2 + C\{\gamma_m \eta_m + \gamma_m^3 + \gamma_m^{2+\kappa}\}.$$

Since $\kappa \le 1$, $\gamma_m^3 \le \gamma_m^{2+\kappa}$ eventually, and the additive terms equal $C\gamma_m q_m^\circ \le C'\gamma_m q_m$ by (23). □

**Lemma G.2** (Deterministic comparison)**.** *Suppose $a_m \ge 0$, $q_m > 0$ and $\gamma_m > 0$ with $\gamma_m \to 0$ satisfy, for constants $c, C > 0$,*

$$a_{m+1} \le (1 - c\gamma_m)a_m + C\gamma_m q_m, \qquad q_{m+1} \le q_m, \qquad \frac{q_m - q_{m+1}}{q_m} = o(\gamma_m).$$

*Then $a_m = O(q_m)$.*

*Proof.* Let $z_m = a_m/q_m$. Since $q_m/q_{m+1} = 1 + o(\gamma_m)$, for all sufficiently large $m$, $q_m/q_{m+1} \le 1 + (c/4)\gamma_m$. Hence

$$\begin{aligned} z_{m+1} &\le \frac{q_m}{q_{m+1}}\{(1 - c\gamma_m)z_m + C\gamma_m\} \\ &\le (1 - c\gamma_m/2)z_m + C'\gamma_m. \end{aligned}$$

Choose $m_0$ so that this inequality holds and $0 \le 1 - c\gamma_m/2 \le 1$ for all $m \ge m_0$. Set $M = \max\{z_{m_0}, 2C'/c\}$. If $z_m \le M$, then

$$z_{m+1} \le (1 - c\gamma_m/2)M + C'\gamma_m \le M.$$

Induction gives $z_m \le M$ for every $m \ge m_0$, hence $a_m = O(q_m)$. □

*Proof of Theorem 3.1.* Take expectations in (58) and set $a_m = \mathbb{E}\|\Delta_m\|^2$. Assumption C.4 and Theorem G.2 give $a_m = O(q_m)$. By (23), this is $O(q_m^\circ) = O(\eta_m + \gamma_m^{1+\kappa})$. If $\delta \ge 1$, then $\kappa = 1$ and $q_m^\circ = \eta_m + \gamma_m^2$. Finally, because $E_m \ge 1$, $\eta_m = \gamma_m/E_m \le \gamma_m$, and because $\gamma_m \to 0$, $\gamma_m^{1+\kappa} \le \gamma_m$. Hence $q_m \le C\gamma_m$ eventually by (23). □

# H Proof of the Polyak–Ruppert limit

## H.1 Abel decomposition

Rearrange (31) as

$$H_\star \Delta_m = \frac{\Delta_m - \Delta_{m+1}}{\gamma_m} - \epsilon_{m+1} + r_{m+1}.$$

Averaging $m = 1, \dots, T$ gives

$$H_\star \bar{\Delta}_T = A_T - M_T + R_T, \tag{61}$$

where

$$\begin{aligned} A_T &= \frac{1}{T}\sum_{m=1}^{T} \frac{\Delta_m - \Delta_{m+1}}{\gamma_m}, \\ M_T &= \frac{1}{T}\sum_{m=1}^{T} \epsilon_{m+1}, \\ R_T &= \frac{1}{T}\sum_{m=1}^{T} r_{m+1}. \end{aligned}$$

Summation by parts yields the exact identity

$$TA_T = \frac{\Delta_1}{\gamma_1} - \frac{\Delta_{T+1}}{\gamma_T} + \sum_{m=2}^{T} \left(\frac{1}{\gamma_m} - \frac{1}{\gamma_{m-1}}\right) \Delta_m. \tag{62}$$

**Lemma H.1** (Boundary and remainder are negligible)**.** *Under the assumptions of Theorem C.8,*

$$\sqrt{t_T}A_T \xrightarrow{\mathbb{P}} 0, \qquad \sqrt{t_T}R_T \xrightarrow{\mathbb{P}} 0.$$

*Proof.* By Theorem 3.1, $\mathbb{E}\,\|\Delta_m\| \le C\sqrt{q_m}$. Multiply (62) by $s_T = \sqrt{t_T}/T$ and take expectations. The fixed initial term is $O(s_T)$, which vanishes by (25). The terminal and sum terms are bounded by a constant times the expression in (27). Thus $\mathbb{E}\left\|\sqrt{t_T}A_T\right\| \to 0$, proving the first convergence by Markov's inequality.

For the remainder, (35) and Theorem 3.1 give

$$\mathbb{E}\,\|r_{m+1}\| \le C(\gamma_m + q_m) \le C'\gamma_m$$

eventually. Therefore

$$\mathbb{E}\left\|\sqrt{t_T}R_T\right\| \le Cs_T \sum_{m=1}^{T} \gamma_m \longrightarrow 0$$

by (26). Markov's inequality proves the second statement. □

### H.2 Martingale central limit theorem

**Lemma H.2** (Conditional covariance convergence)**.** *Under the assumptions of Theorem C.8,*

$$\frac{t_T}{T^2}\sum_{m=1}^{T} \mathbb{E}(\epsilon_{m+1}\otimes\epsilon_{m+1} \mid \mathcal{F}_m) \xrightarrow{\mathbb{P}} \nu S.$$

*Proof.* By (32), the leading term equals

$$\frac{t_T}{T^2}\sum_{m=1}^{T} E_m^{-1}S = \nu_T S \longrightarrow \nu S.$$

For the perturbation, (34) and Cauchy–Schwarz give

$$\mathbb{E}\left\|\frac{t_T}{T^2}\sum_{m=1}^{T} Q_m\right\|_{\mathrm{op}} \le C\frac{t_T}{T^2}\sum_{m=1}^{T} E_m^{-1}\{\sqrt{q_m}+\gamma_m\}.$$

The total weight $(t_T/T^2)\sum_m E_m^{-1} = \nu_T$ is bounded. For any fixed $M$, the contribution of $m \le M$ is at most $C_M t_T/T^2 = C_M s_T^2 \to 0$. For $m > M$, the weighted average is bounded by $C\nu_T \sup_{m>M}(\sqrt{q_m}+\gamma_m)$, which tends to zero as $M \to \infty$. Hence the perturbation converges to zero in $L^1$ and therefore in probability. □

**Lemma H.3** (Martingale array CLT)**.** *Under the assumptions of Theorem C.8,*

$$\frac{\sqrt{t_T}}{T}\sum_{m=1}^{T} \epsilon_{m+1} \xrightarrow{d} N(0,\nu S). \tag{63}$$

*Proof.* For a fixed $u \in T_{x_\star}\mathcal{M}$, apply the scalar martingale central limit theorem to

$$X_{T,m} = \frac{\sqrt{t_T}}{T}\langle u, \epsilon_{m+1}\rangle .$$

By Theorem H.2, the sum of conditional variances converges in probability to $\nu\langle u, Su\rangle$. Moreover, (33) gives the conditional Lyapunov bound

$$\sum_{m=1}^{T}\mathbb{E}(|X_{T,m}|^{2+\delta} \mid \mathcal{F}_m) \le C\,\|u\|^{2+\delta}\left(\frac{\sqrt{t_T}}{T}\right)^{2+\delta}\sum_{m=1}^{T} E_m^{-(1+\delta/2)} \longrightarrow 0$$

by (28). Thus the scalar limit is normal. The Cramér–Wold device gives (63). This is the standard martingale-array argument of (Hall & Heyde, 1980, Corollary 3.1). □

*Proof of Theorem C.8.* Multiply (61) by $\sqrt{t_T}$. By Theorem H.1, the boundary and remainder terms are $o_p(1)$. By Theorem H.3,

$$\sqrt{t_T} H_\star \bar{\Delta}_T = -\frac{\sqrt{t_T}}{T} \sum_{m=1}^{T} \epsilon_{m+1} + o_p(1) \xrightarrow{d} N(0, \nu S).$$

Since $H_\star$ is invertible and self-adjoint, multiplication by $H_\star^{-1}$ and Slutsky's theorem yield

$$\sqrt{t_T} \bar{\Delta}_T \xrightarrow{d} N(0, \nu H_\star^{-1} S H_\star^{-1}).$$

□

# I Proofs for the computable geometric averages

## I.1 Fixed-anchor average

**Lemma I.1** (Fixed-chart cancellation)**.** *Let $a \in \mathcal{U}$ be fixed or be a random post-burn-in anchor taking values in a compact subset of $\mathcal{U}$, and define $\psi_a(v) = R_a^{-1}\{R_{x_\star}(v)\}$. There is an invertible linear map $J_a : T_{x_\star}\mathcal{M} \to T_a\mathcal{M}$ and a constant $C$ such that*

$$\psi_a(v) = v_\star + J_a v + \rho_a(v), \qquad \|\rho_a(v)\| \le C \|v\|^2, \tag{64}$$

$$R_{x_\star}^{-1}\{R_a(v_\star + z)\} = J_a^{-1} z + \widetilde{\rho}_a(z), \qquad \|\widetilde{\rho}_a(z)\| \le C \|z\|^2, \tag{65}$$

*where $v_\star = R_a^{-1}(x_\star)$. The constants are uniform for $a$ in a compact subset of $\mathcal{U}$.*

*Proof.* For each realized anchor, the map $\psi_a$ is a smooth local diffeomorphism because it is a composition of two retraction charts. Set $J_a = D\psi_a(0)$. The inverse function theorem makes $J_a$ invertible and identifies the derivative of the inverse map as $J_a^{-1}$. The two bounds are ordinary second-order Taylor expansions on compact coordinate neighborhoods. □

*Proof of Theorem 3.2.* A fixed burn-in changes neither $T$ nor $t_T$ to first order, so write the average from $m = 1$ to $T$. Apply (64) to every $\Delta_m$:

$$\bar{v}_T - v_\star = J_a \bar{\Delta}_T + \bar{\rho}_T, \qquad \|\bar{\rho}_T\| \le \frac{C}{T} \sum_{m=1}^{T} \|\Delta_m\|^2.$$

Applying (65) to $\widehat{x}_T^{\mathrm{FA}} = R_a(\bar{v}_T)$ gives

$$R_{x_\star}^{-1}(\widehat{x}_T^{\mathrm{FA}}) = \bar{\Delta}_T + J_a^{-1} \bar{\rho}_T + O\{\left\|J_a \bar{\Delta}_T + \bar{\rho}_T\right\|^2\}. \tag{66}$$

By Theorem 3.1 and $q_m \le C\gamma_m$,

$$\mathbb{E}\{\sqrt{t_T}\,\|\bar{\rho}_T\|\} \le C \frac{\sqrt{t_T}}{T} \sum_{m=1}^{T} q_m \le C \frac{\sqrt{t_T}}{T} \sum_{m=1}^{T} \gamma_m \longrightarrow 0.$$

Hence $\sqrt{t_T}\bar{\rho}_T = o_p(1)$. By Theorem C.8, $\bar{\Delta}_T = O_p(t_T^{-1/2})$, so $\sqrt{t_T}\left\|\bar{\Delta}_T\right\|^2 = O_p(t_T^{-1/2}) = o_p(1)$. The same holds for the square of $\bar{\rho}_T$ and the cross-product. Multiplying (66) by $\sqrt{t_T}$ proves (11); (12) follows from Slutsky's theorem. □

## I.2 Recursive retraction average

Let $\widetilde{\Delta}_m = R_{x_\star}^{-1}(\widetilde{x}_m)$.

**Lemma I.2** (Second moment of the recursive average)**.** *Under Assumption C.5,*

$$\mathbb{E}\left\|\widetilde{\Delta}_m\right\|^2 \le \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\,\|\Delta_j\|^2 = O(\bar{q}_m).$$

*Proof.* The update $\widetilde{x}_{m+1}$ lies at retraction fraction $1/(m+1)$ from $\widetilde{x}_m$ to $x_{m+1}$. Convexity in Assumption C.5 gives

$$\left\|\widetilde{\Delta}_{m+1}\right\|^2 \le \frac{m}{m+1}\left\|\widetilde{\Delta}_m\right\|^2 + \frac{1}{m+1}\left\|\Delta_{m+1}\right\|^2.$$

Induction yields $\left\|\widetilde{\Delta}_m\right\|^2 \le m^{-1}\sum_{j=1}^m \left\|\Delta_j\right\|^2$. Taking expectations and using Theorem 3.1 proves the claim. This is the local averaging argument underlying the recursive RSGD average in Tripuraneni et al. (2018). □

**Lemma I.3** (Recursive-average Taylor recursion)**.** *There are errors $e_{m+1}$ such that*

$$\widetilde{\Delta}_{m+1} = \widetilde{\Delta}_m + \frac{1}{m+1}(\Delta_{m+1} - \widetilde{\Delta}_m) + \frac{1}{m+1}e_{m+1}, \tag{67}$$

*and*

$$\mathbb{E}\left\|e_{m+1}\right\| \le C(q_{m+1} + \bar{q}_m). \tag{68}$$

*Proof.* For fixed $\widetilde{x}_m$, define the smooth map on target coordinates

$$F_m(z) = R_{x_\star}^{-1}\left[R_{\widetilde{x}_m}\left\{\frac{1}{m+1}R_{\widetilde{x}_m}^{-1}(R_{x_\star}(z))\right\}\right].$$

At $z = \widetilde{\Delta}_m$, the inverse retraction equals zero, so $F_m(\widetilde{\Delta}_m) = \widetilde{\Delta}_m$. The chain rule and $DR_x(0) = \mathrm{Id}$ show $DF_m(\widetilde{\Delta}_m) = (m+1)^{-1}\mathrm{Id}$. Differentiating once more, every second-derivative term contains at least one factor $(m+1)^{-1}$; compactness gives

$$\sup_z \left\|D^2F_m(z)\right\|_{\mathrm{op}} \le C/(m+1)$$

on the relevant segment. Taylor's theorem therefore yields

$$\left\|F_m(\Delta_{m+1}) - \widetilde{\Delta}_m - (m+1)^{-1}(\Delta_{m+1} - \widetilde{\Delta}_m)\right\| \le \frac{C}{m+1}\left\|\Delta_{m+1} - \widetilde{\Delta}_m\right\|^2.$$

This is (67) with $\left\|e_{m+1}\right\| \le C\left\|\Delta_{m+1} - \widetilde{\Delta}_m\right\|^2$. Taking expectations, using $(a+b)^2 \le 2a^2 + 2b^2$, Theorem 3.1, and Theorem I.2 proves (68). □

*Proof of Theorem C.10.* Multiply (67) by $m+1$:

$$(m+1)\widetilde{\Delta}_{m+1} = m\widetilde{\Delta}_m + \Delta_{m+1} + e_{m+1}.$$

Iterating gives, up to an immaterial initial term,

$$\widetilde{\Delta}_T - \bar{\Delta}_T = \frac{1}{T}\sum_{m=1}^T e_m.$$

Hence, by (68),

$$\begin{aligned}\mathbb{E}\left\|\sqrt{t_T}(\widetilde{\Delta}_T - \bar{\Delta}_T)\right\| &\le C\frac{\sqrt{t_T}}{T}\sum_{m=1}^T (q_m + \bar{q}_m) \\ &\le C\frac{\sqrt{t_T}}{T}\sum_{m=1}^T \gamma_m + C\frac{\sqrt{t_T}}{T}\sum_{m=1}^T \bar{q}_m \longrightarrow 0\end{aligned}$$

by (26) and (29). Markov's inequality proves (39); (40) follows from Theorem C.8 and Slutsky's theorem. The final assertion for power schedules is proved in Section D.1. □

### I.3 PROPORTIONAL-TAIL AVERAGES

*Proof of Theorem C.11.* Apply the Abel decomposition and martingale-array argument in the proof of Theorem C.8 to $m = b_T + 1, \dots, T$. The stated tail conditions make both endpoint terms and the retained remainders negligible; the leading covariance is exactly (41). The fixed-anchor cancellation is pathwise for one common, possibly data-dependent anchor held fixed within each realized average. The recursive and plug-in conclusions follow by the same tail Cesàro arguments. Finally, integral comparison gives

$$\sum_{m=b_T+1}^{T} E_m^{-1} \sim \frac{T^{1-\beta}(1-\rho^{1-\beta})}{c_E(1-\beta)}, \qquad t_T \sim \frac{c_E T^{1+\beta}}{1+\beta},$$

which yields (42). The strict inequalities in (37) also control the lower Abel endpoint because $b_T \asymp T$ when $\rho > 0$; $\rho = 0$ is the original result. □

## J PROOFS FOR FEDERATED PLUG-IN INFERENCE

For this section, a fixed burn-in is again suppressed. All random vectors and operators are identified with Euclidean coordinates through an arbitrary orthonormal basis of $T_{x_\star}\mathcal{M}$; the conclusions do not depend on that basis.

**Lemma J.1** (Martingale Cesàro law)**.** *Let $D_m$ be a martingale-difference sequence in a fixed finite-dimensional normed space. If there is $p \in (1, 2]$ such that $\sup_m \mathbb{E}\,\|D_m\|^p < \infty$, then the martingale moment bound (Hall & Heyde, 1980, Theorem 2.10) yields*

$$\frac{1}{T}\sum_{m=1}^{T} D_m \longrightarrow 0 \quad \textit{in } L^p \textit{ and hence in probability}.$$

*Proof.* Choose a fixed coordinate basis. For each coordinate, the martingale von Bahr–Esseen inequality, or equivalently the $p \le 2$ case of the Burkholder inequality, gives

$$\mathbb{E}\left|\sum_{m=1}^{T} D_{m,j}\right|^p \le C_p \sum_{m=1}^{T} \mathbb{E}\,|D_{m,j}|^p \le CT.$$

Norm equivalence and the finite number of coordinates therefore imply $\mathbb{E}\left\|T^{-1}\sum_{m=1}^{T} D_m\right\|^p \le CT^{1-p} \to 0$. □

**Lemma J.2** (Stability of a transported final anchor)**.** *Under Assumption C.1, there is $C < \infty$ such that, for all $x, y \in \mathcal{U}$,*

$$\left\|\mathcal{T}_y^{x_\star}\mathcal{T}_x^y - \mathcal{T}_x^{x_\star}\right\|_{\mathrm{op}} \le C d(y, x_\star). \tag{69}$$

*Consequently, for a tangent vector $v \in T_x\mathcal{M}$ and a linear operator $B : T_x\mathcal{M} \to T_x\mathcal{M}$,*

$$\left\|\mathcal{T}_y^{x_\star}\mathcal{T}_x^y v - \mathcal{T}_x^{x_\star} v\right\| \le C d(y, x_\star)\,\|v\|, \tag{70}$$

$$\left\|\mathcal{T}_y^{x_\star}\mathcal{T}_x^y B \mathcal{T}_y^x \mathcal{T}_{x_\star}^y - \mathcal{T}_x^{x_\star} B \mathcal{T}_{x_\star}^x\right\|_{\mathrm{op}} \le C d(y, x_\star)\,\|B\|_{\mathrm{op}}. \tag{71}$$

*Proof.* For fixed $x$, the two maps in (69) agree at $y = x_\star$ because $\mathcal{T}_{x_\star}^{x_\star} = \mathrm{Id}$. Smoothness of the two-point transport and compactness of $\mathcal{U}$ give a derivative bound that is uniform in $x$. The mean-value theorem on a finite collection of coordinate neighborhoods proves (69). The vector bound follows by application to $v$. For the operator bound, insert and subtract $\mathcal{T}_x^{x_\star} B \mathcal{T}_y^x \mathcal{T}_{x_\star}^y$ and use isometry, the vector-map bound, and the analogous bound for the inverses. Since $\mathcal{T}_y^x = (\mathcal{T}_x^y)^{-1}$, both perturbations are $O\{d(y, x_\star)\}$. □

**Lemma J.3** (Direct-anchor Hessian and score laws)**.** *Define operators and vectors in $T_{x_\star}\mathcal{M}$ by*

$$H_m^\star = \sum_{k=1}^{K} p_k \mathcal{T}_{x_m}^{x_\star} B_k(x_m, \omega_m^k) \mathcal{T}_{x_\star}^{x_m},$$

$$Z_m^\star = \sum_{k=1}^{K} p_k \mathcal{T}_{x_m}^{x_\star} G_k(x_m, \zeta_m^k).$$

*Under Assumptions C.1–C.4 and Assumption C.6,*

$$\frac{1}{T}\sum_{m=1}^{T} H_m^\star \xrightarrow{\mathbb{P}} H_\star, \tag{72}$$

$$\frac{1}{T}\sum_{m=1}^{T} Z_m^\star \otimes Z_m^\star \xrightarrow{\mathbb{P}} S. \tag{73}$$

*Proof.* **Step 1: Hessian-message fluctuation.** Let

$$\mathcal{H}_m = \mathbb{E}(H_m^\star \mid \mathcal{F}_m) = \sum_{k=1}^{K} p_k \mathcal{T}_{x_m}^{x_\star} \nabla^2 f_k(x_m) \mathcal{T}_{x_\star}^{x_m}.$$

The difference $H_m^\star - \mathcal{H}_m$ is a martingale difference with respect to the filtration that reveals the fresh Hessian messages round by round. Put $p_H = \min(1+\tau, 2) > 1$. Isometry, fixed $K$, and (30) imply $\sup_m \mathbb{E}\,\|H_m^\star - \mathcal{H}_m\|_{\mathrm{op}}^{p_H} < \infty$. Thus Theorem J.1 gives

$$\frac{1}{T}\sum_{m=1}^{T} (H_m^\star - \mathcal{H}_m) \xrightarrow{\mathbb{P}} 0. \tag{74}$$

**Step 2: deterministic Hessian target.** By the continuity clause in Assumption C.2, $\mathcal{H}_m \xrightarrow{\mathbb{P}} H_\star$ because $x_m \xrightarrow{\mathbb{P}} x_\star$ by Theorem 3.1. The fields are bounded on the compact neighborhood, so the convergence also holds in $L^1$ after a standard uniform-integrability argument. Cesàro averaging yields $T^{-1}\sum_m \mathcal{H}_m \xrightarrow{\mathbb{P}} H_\star$. Together with (74), this proves (72).

**Step 3: conditional second moment of the fresh score.** Let

$$h_m = \mathcal{T}_{x_m}^{x_\star} \nabla f(x_m).$$

Conditional independence across clients and the zero mean of each fresh oracle noise give

$$\mathbb{E}(Z_m^\star \otimes Z_m^\star \mid \mathcal{F}_m) = h_m \otimes h_m + \sum_{k=1}^{K} p_k^2 \mathcal{T}_{x_m}^{x_\star} S_k(x_m) \mathcal{T}_{x_\star}^{x_m}. \tag{75}$$

The gradient expansion (19) and isometry imply $\|h_m\| \le C(\|\Delta_m\| + \|\Delta_m\|^2)$ in the local chart. Since the chart is bounded and Theorem 3.1 gives $\mathbb{E}\,\|\Delta_m\|^2 \to 0$,

$$\frac{1}{T}\sum_{m=1}^{T} \|h_m\|^2 \xrightarrow{\mathbb{P}} 0. \tag{76}$$

In (22), take $y = x_\star$ and $x = x_m$. Then $A(x_\star, x_m) = \mathcal{T}_{x_m}^{x_\star}$, so

$$\left\|\mathcal{T}_{x_m}^{x_\star} S_k(x_m) \mathcal{T}_{x_\star}^{x_m} - S_k^\star\right\|_{\mathrm{op}} \le L_S d(x_m, x_\star) \xrightarrow{\mathbb{P}} 0.$$

Boundedness and Cesàro averaging therefore show that the second term in (75), averaged over $m$, converges in probability to $S$. Together with (76),

$$\frac{1}{T}\sum_{m=1}^{T} \mathbb{E}(Z_m^\star \otimes Z_m^\star \mid \mathcal{F}_m) \xrightarrow{\mathbb{P}} S. \tag{77}$$

**Step 4: outer-product fluctuation.** Set

$$D_m = Z_m^\star \otimes Z_m^\star - \mathbb{E}(Z_m^\star \otimes Z_m^\star \mid \mathcal{F}_m).$$

This is a martingale difference in the space of self-adjoint operators. With $p_S = 1 + \delta/2 \in (1, 2]$, Jensen's inequality, fixed $K$, isometry, and (21) imply

$$\sup_m \mathbb{E}\,\|D_m\|_{\mathrm{op}}^{p_S} \le C \sup_m \mathbb{E}\,\|Z_m^\star\|^{2+\delta} < \infty.$$

By Theorem J.1, $T^{-1}\sum_m D_m \xrightarrow{\mathbb{P}} 0$. Combining this with (77) proves (73). □

*Proof of Theorem 3.3.* Write $\widehat{x} = \widehat{x}_T$ and transport the reported quantities to the target tangent space:

$$\bar{H}_T = \mathcal{T}_{\widehat{x}}^{x_\star} \widehat{H}_T \mathcal{T}_{x_\star}^{\widehat{x}},$$
$$\bar{S}_T = \mathcal{T}_{\widehat{x}}^{x_\star} \widehat{S}_T \mathcal{T}_{x_\star}^{\widehat{x}}.$$

Both geometric averages are consistent by their central limit theorems, so $d(\widehat{x}, x_\star) = o_p(1)$.

For the Hessian estimator, (71) gives

$$\left\| \bar{H}_T - \frac{1}{T} \sum_{m=1}^{T} H_m^\star \right\|_{\mathrm{op}} \le C d(\widehat{x}, x_\star) \frac{1}{T} \sum_{m=1}^{T} \sum_{k=1}^{K} p_k \left\| B_k(x_m, \omega_m^k) \right\|_{\mathrm{op}}.$$

The average on the right is $O_p(1)$ by Markov's inequality and (30); hence the difference is $o_p(1)$. The first assertion (13) follows from (72).

For the score estimator, define

$$\bar{Z}_m = \mathcal{T}_{\widehat{x}}^{x_\star} \widehat{Z}_m = \sum_{k=1}^{K} p_k \mathcal{T}_{\widehat{x}}^{x_\star} \mathcal{T}_{x_m}^{\widehat{x}} G_k(x_m, \zeta_m^k).$$

By (70),

$$\left\| \bar{Z}_m - Z_m^\star \right\| \le C d(\widehat{x}, x_\star) \sum_{k=1}^{K} p_k \left\| G_k(x_m, \zeta_m^k) \right\|.$$

Using $\|u \otimes u - v \otimes v\|_{\mathrm{op}} \le (\|u\| + \|v\|) \|u - v\|$, we obtain

$$\left\| \bar{S}_T - \frac{1}{T} \sum_{m=1}^{T} Z_m^\star \otimes Z_m^\star \right\|_{\mathrm{op}} \le C d(\widehat{x}, x_\star) \frac{1}{T} \sum_{m=1}^{T} \left\{ \sum_{k=1}^{K} p_k \left\| G_k(x_m, \zeta_m^k) \right\| \right\}^2.$$

The average is $O_p(1)$ because the score messages have uniformly bounded second moments. The right side is therefore $o_p(1)$, and (13) follows from (73).

Finally, for fixed $b$, $n_T/T \to 1$ and $(t_T/T^2) \sum_{m=1}^{b-1} E_m^{-1} \to 0$ by (25). Thus $\widehat{\nu}_T - \nu_T \to 0$ and $\widehat{\nu}_T \to \nu$. Moreover, (13) with $H_\star \succ 0$ implies that $\widehat{H}_T$ is invertible with probability tending to one. Isometric conjugation commutes with inversion. Hence the continuous mapping theorem gives

$$\mathcal{T}_{\widehat{x}}^{x_\star} \widehat{\Sigma}_T \mathcal{T}_{x_\star}^{\widehat{x}} = \widehat{\nu}_T \bar{H}_T^{-1} \bar{S}_T \bar{H}_T^{-1} \xrightarrow{\mathbb{P}} \nu H_\star^{-1} S H_\star^{-1} = \Sigma_\star,$$

which proves (14). □

*Proof of Theorems 3.4 and C.12.* Let $\widehat{\Delta}_T = R_{x_\star}^{-1}(\widehat{x}_T)$. By Theorems 3.2 and C.10, as applicable,

$$\sqrt{t_T} \widehat{\Delta}_T \xrightarrow{d} N(0, \Sigma_\star). \tag{78}$$

The inverse-error expansion (51) gives

$$\mathcal{T}_{\widehat{x}_T}^{x_\star} R_{\widehat{x}_T}^{-1}(x_\star) = -\widehat{\Delta}_T + O_p(\left\| \widehat{\Delta}_T \right\|^2) = -\widehat{\Delta}_T + o_p(t_T^{-1/2}). \tag{79}$$

If $S \succ 0$, then $\Sigma_\star \succ 0$. Combining (14), (78), and (79) with Slutsky's theorem yields

$$t_T \left\langle R_{\widehat{x}_T}^{-1}(x_\star), \widehat{\Sigma}_T^{-1} R_{\widehat{x}_T}^{-1}(x_\star) \right\rangle \xrightarrow{d} \chi_d^2.$$

This is precisely the statistic defining the event $\{x_\star \in \mathcal{C}_{T,1-\alpha_{\mathrm{W}}}\}$, proving (16).

For a $C^1$ map $\psi$, Taylor expansion in the target chart gives

$$\psi(\widehat{x}_T) - \psi(x_\star) = D\psi_{x_\star} \widehat{\Delta}_T + o_p(\left\| \widehat{\Delta}_T \right\|).$$

Equation (78) and the delta method prove (43). Smoothness of $D\psi$ and (14) imply consistency of $D\psi_{\widehat{x}_T} \widehat{\Sigma}_T D\psi_{\widehat{x}_T}^*$ after the canonical local identification of tangent spaces. Slutsky's theorem then gives the stated functional Wald region. □

# K INFERENCE FOR THE GAUSSIAN QUADRATIC RECURSION

The common Hessian in the Gaussian experiment makes the pooled round transition linear: the client shifts cancel exactly. This gives a direct connection between the implemented recursion and its inference law on the whole Euclidean space.

**Proposition K.1** (Gaussian quadratic model). *Let $f_k(x) = \frac{1}{2}(x-a_k)^\top H(x-a_k)$ on $\mathbb{R}^d$, with a common $H \succ 0$, target $x_\star = \sum_k p_k a_k$, and oracle $G_k(x,\xi) = H(x-a_k)+\xi$. The noises are independent across clients and oracle calls, with client $k$ having law $N(0,S_k)$; let $S = \sum_k p_k^2 S_k \succ 0$. Run the Euclidean recursion* (1)–(3) *from any deterministic $x_1$, with*

$$E_m = \max\{1, \lfloor c_E m^\beta \rfloor\}, \qquad \gamma_m = c_\gamma (m+m_0)^{-\alpha}, \qquad \eta_m = \gamma_m / E_m,$$

*where $c_E, c_\gamma > 0$, $m_0 \geq 0$, $0 \leq \beta < 1$ and $(1+\beta)/2 < \alpha < 1$. Then*

$$\mathbb{E}\,\|x_m - x_\star\|^2 = O(\eta_m).$$

*Fix $\rho \in [0,1)$ and an integer $b_0 \geq 0$, let $b_T = \max\{b_0, \lfloor \rho T \rfloor\}$, and average the points $x_{b_T+1}, \dots, x_T$. Both Euclidean averaging procedures equal this arithmetic average, denoted by $\widehat{x}_T$. Using independent fresh scores at these same points and exact Hessian messages in* (6)–(8)*, with the retained-round variance factor, gives*

$$\sqrt{t_T}(\widehat{x}_T - x_\star) \xrightarrow{d} N(0, \nu_\rho H^{-1} S H^{-1}),$$
$$\widehat{H}_T = H, \qquad \widehat{S}_T \xrightarrow{\mathbb{P}} S, \qquad \widehat{\Sigma}_T \xrightarrow{\mathbb{P}} \nu_\rho H^{-1} S H^{-1},$$

*where $\nu_\rho$ is given by* (42)*. The Wald regions* (15) *consequently have limiting coverage $1-\alpha_{\mathrm{W}}$ under the original, unmodified recursion.*

The proposition covers the Gaussian configurations with constant intervals and $(\alpha, c_\gamma) = (0.65, 0.8)$, and square-root intervals with $\alpha = 0.80$ and $c_\gamma = 0.8$ or $1.6$. It allows the finite initial transient generated by the specified starting point and step offset. Its proof uses the exact linear transition, rather than Assumption C.1(G3).

*Proof.* Write $\Delta_m = x_m - x_\star$ and $B_m = I - \eta_m H$, and denote the noise in client $k$'s local step $e$ by $\xi^k_{m,e}$. Unrolling the local updates gives

$$x^k_{m,E_m} = a_k + B_m^{E_m}(x_m - a_k) - \eta_m \sum_{e=0}^{E_m-1} B_m^{E_m-1-e} \xi^k_{m,e}.$$

In Euclidean space the server update equals $\sum_k p_k x^k_{m,E_m}$. Since $\sum_k p_k(a_k - x_\star) = 0$, the exact centered round recursion is

$$\Delta_{m+1} = A_m \Delta_m - \gamma_m \epsilon_{m+1}, \qquad A_m = B_m^{E_m}, \tag{80}$$
$$\epsilon_{m+1} = \frac{1}{E_m} \sum_{e=0}^{E_m-1} B_m^{E_m-1-e} \sum_k p_k \xi^k_{m,e}.$$

The round noises are independent centered Gaussian vectors, independent of the preceding trajectory, with covariance

$$\mathbb{E}(\epsilon_{m+1}\epsilon_{m+1}^\top) = \frac{1}{E_m^2} \sum_{j=0}^{E_m-1} B_m^j S B_m^j = \frac{S}{E_m} + O_{\mathrm{op}}(\gamma_m / E_m).$$

Here $O_{\mathrm{op}}$ bounds the operator norm. For all sufficiently large $m$, $0 \prec B_m \preceq I$, and $\left\|B_m^j - I\right\|_{\mathrm{op}} \leq j\eta_m \|H\|_{\mathrm{op}} \leq C\gamma_m$ for $j < E_m$, which proves the covariance expansion without a commutation assumption on $H$ and $S$. Also, if $\mu = \lambda_{\min}(H)$,

$$\|A_m\|_{\mathrm{op}} \leq e^{-\mu\gamma_m}, \qquad \|A_m - I + \gamma_m H\|_{\mathrm{op}} \leq C\gamma_m^2.$$

The second bound follows by expanding $(1-\eta_m\lambda)^{E_m}$ on each eigenspace of $H$. Independence in (80) therefore gives

$$\mathbb{E}\,\|\Delta_{m+1}\|^2 \leq (1 - c\gamma_m)\mathbb{E}\,\|\Delta_m\|^2 + C\gamma_m^2 / E_m$$

eventually. All earlier rounds have finite Gaussian moments. With $q_m = m^{-(\alpha+\beta)} \asymp \eta_m$, we have $q_m/q_{m+1} = 1 + O(m^{-1}) = 1 + o(\gamma_m)$; the deterministic comparison in Theorem G.2 proves the asserted global state bound.

Put $n_{T,\rho} = T - b_T$ and $s_{T,\rho} = \sqrt{t_T}/n_{T,\rho}$. Equation (80) can be written in the form (31) with $r_{m+1} = \gamma_m^{-1}(A_m - I + \gamma_m H)\Delta_m$. Thus $\mathbb{E}\,\|r_{m+1}\| \le C\gamma_m\sqrt{\eta_m} \le C'\gamma_m$ eventually. Since $s_{T,\rho} = O(T^{(\beta-1)/2})$,

$$s_{T,\rho} \sum_{m=b_T+1}^{T} \mathbb{E}\,\|r_{m+1}\| \le O\Big(T^{(1+\beta)/2-\alpha}\Big) + O(s_{T,\rho}) \longrightarrow 0.$$

Summation by parts over the retained range bounds the expected norm of the other error by

$$\begin{aligned} & s_{T,\rho}\,\mathbb{E}\left\| \sum_{m=b_T+1}^{T} \frac{\Delta_m - \Delta_{m+1}}{\gamma_m} \right\| \\ & \le C s_{T,\rho} \left\{ \frac{\sqrt{\eta_{b_T+1}}}{\gamma_{b_T+1}} + \frac{\sqrt{\eta_{T+1}}}{\gamma_T} + \sum_{m=b_T+2}^{T} m^{\alpha-1}\sqrt{\eta_m} \right\} \\ & = O\Big(T^{(\alpha-1)/2}\Big) + O(s_{T,\rho}) \longrightarrow 0. \end{aligned}$$

The bounds also cover $\rho = 0$, when the initial boundary is fixed. Consequently

$$\sqrt{t_T}(\widehat{x}_T - x_\star) = -H^{-1} s_{T,\rho} \sum_{m=b_T+1}^{T} \epsilon_{m+1} + o_p(1).$$

The sum is exactly Gaussian. Its covariance converges to $\nu_\rho S$ because

$$s_{T,\rho}^2 \sum_{m=b_T+1}^{T} E_m^{-1} \longrightarrow \nu_\rho, \qquad s_{T,\rho}^2 \sum_{m=b_T+1}^{T} \gamma_m E_m^{-1} \longrightarrow 0.$$

The first limit follows from the power sums in (42); replacing $\lfloor \rho T \rfloor$ by its maximum with $b_0$ changes finitely many terms when $\rho = 0$. For the second, split off finitely many initial terms and use $\gamma_m \to 0$ on the rest. This proves the CLT and also shows why finite early rounds disappear.

For the fresh messages, pooling first gives $\widehat{Z}_m = H\Delta_m + \zeta_m$, where the $\zeta_m$ are independent $N(0, S)$ vectors independent of the optimization trajectory. The state bound implies $n_{T,\rho}^{-1} \sum_{m=b_T+1}^{T} \|\Delta_m\|^2 \xrightarrow{\mathbb{P}} 0$ by expectation and Cesàro averaging. The Gaussian law of large numbers gives $n_{T,\rho}^{-1} \sum_{m=b_T+1}^{T} \zeta_m \zeta_m^\top \xrightarrow{\mathbb{P}} S$; Cauchy–Schwarz makes the cross terms vanish. Expanding (7) yields $\widehat{S}_T \xrightarrow{\mathbb{P}} S$. Exact Hessians give $\widehat{H}_T = H$, and the retained-round factor converges to $\nu_\rho$. The covariance limit follows. Since $H, S \succ 0$, the conditioning checks pass with probability tending to one, and Slutsky's theorem gives the stated Wald coverage. □