DISTRIBUTED INFERENCE FOR MANIFOLD M-ESTIMATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop one-round confidence regions for manifold M-estimators from distributed data. An independent initial estimate places all workers in a common chart, where reported scores, Hessians and score second moments determine an estimate and its covariance. Under fixed-dimensional regularity, we separate the nonlinear error from the additional error caused by using only the master worker's Hessian. This decomposition yields explicit accuracy and machine-growth conditions for centralized first-order inference, with confidence regions invariant to orthonormal frame choice. For principal component analysis, the same summaries recover the exact centralized subspace, reducing the requirement to consistency. Spherical and subspace experiments connect coverage to accuracy and Hessian estimation, and quantify the communication trade-off between one-round recentring and two-round covariance estimation. A Dry Bean application illustrates uncertainty for feature contributions to a principal subspace.

1 INTRODUCTION

A mean direction, a rotation and a principal subspace are statistical parameters constrained to manifolds. When their observations are distributed across machines, fitting the parameter is only part of the task: confidence regions must also describe its sampling uncertainty. For a principal subspace, for example, uncertainty in its projection onto a feature axis determines how precisely that feature's contribution is estimated. An approximation adequate for fitting can be inadequate for inference if its error is comparable to the standard error of the combined sample.

Distributed Euclidean estimation makes this distinction explicit. With nonlinear losses, averaging local estimates can retain bias that becomes consequential as machine count grows (Zhang et al., 2013; Rosenblatt and Nadler, 2016). Gradient-corrected objectives use local optimization with pooled score information: approximate Newton and iterative surrogate methods exchange updates (Shamir et al., 2014; Fan et al., 2023), while communication-efficient surrogate likelihood supports both low-dimensional estimation and confidence intervals (Jordan et al., 2019). On manifolds, Saparbayeva et al. (2018) extend distributed surrogate optimization toward the centralized empirical minimizer. The inferential question is how one report from each worker can deliver both a sufficiently accurate center and its sampling covariance in compatible coordinates.

Centralized manifold statistics supplies the local ingredients: intrinsic means and tangent covariances (Pennec, 2006; Bhattacharya and Patrangenaru, 2005), asymptotic laws for geodesically convex M-estimators (Brunel, 2023), empirical likelihood (Kurisu and Otsu, 2026) and higher-order confidence procedures (Huang and Zhang, 2026). Streaming Riemannian estimation (Tripuraneni et al., 2018) and distributed principal component analysis (PCA) (Huang and Pan, 2020) address data access and communication. For eigenspace aggregation, Fan et al. (2019) show that adding machines at fixed local sample size reduces sampling variation while eigenspace bias can persist. Connecting these developments requires tracking the distributed approximation error on the scale of the centralized sampling law.

The geometric obstacle is to align scores and covariances while tracking the target's coordinate in a data-dependent chart. Our independent initial estimate gives every worker a common anchor. Conditional on it, worker observations remain independent and the target coordinate is fixed. Reports of scores, score second moments and, optionally, Hessians then determine the estimate and covariance. An expansion separates sampling fluctuation from the nonlinear initial-estimate remainder and the

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.

additional Hessian-estimation error of a master-worker surrogate. Adding workers decreases the sampling error, so these approximation terms must decrease sufficiently fast. One reporting round is counted after initial-estimate construction and anchor broadcast; comparisons include their costs.

With ρ-Hölder Hessians and a root-n initial estimate from a sample of size n, $m = o(n^\rho)$ workers suffice for centralized first-order inference. Sending the pooled Hessian removes the additional master-Hessian error, and the mapped regions are exactly invariant to orthonormal frame choice. PCA admits a stronger algebraic reduction: the Newton summaries recover the exact worker-sample subspace, so recentring eliminates the one-step localization error and relaxes the initial-estimate requirement to consistency. Spherical experiments test the general nonlinear construction; Grassmann experiments distinguish localization, Hessian estimation and covariance costs; a Dry Bean analysis turns subspace uncertainty into feature-level intervals.

Section 2 develops the common-chart estimators and confidence regions. Section 3 gives their approximation and coverage guarantees; Sections 4 and 5 study their numerical behavior and the Dry Bean application. Proofs and implementation details follow in the appendices.

2 METHODOLOGY

2.1 TARGET AND COMMON COORDINATES

Let $\mathcal{M}$ be a smooth d-dimensional Riemannian manifold without boundary, with fixed d, and let $\ell : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$ be a measurable loss for a parameter $\theta \in \mathcal{M}$ and an observation Z in a measurable sample space $\mathcal{Z}$. The target $\theta_\star$ is an isolated regular local minimizer of $F(\theta) = \mathbb{E}\{\ell(\theta; Z)\}$. Worker j holds $\mathcal{D}_j = \{Z_{j,i} : i = 1, \dots, n\}$, $j = 1, \dots, m$. An independent initial sample $\mathcal{D}_0$ contains n_0 observations from the same distribution. All observations are i.i.d.; write $N = mn$ for the worker sample size and $T = N + n_0$ for the total. The initial estimate locates the chart, while the workers supply the inferential information. Relative to using all T observations, the first-order variance factor is $T/N = 1 + n_0/N$, which tends to one when $n_0 = o(N)$ (Remark A.1).

The common chart uses a smooth local orthonormal frame $E_p : \mathbb{R}^d \to T_p\mathcal{M}$ and a smooth retraction Ret_p from a neighborhood of zero in $T_p\mathcal{M}$ into $\mathcal{M}$. A retraction agrees to first order with geodesic motion. For $v \in \mathbb{R}^d$ near zero, set $\mathcal{R}_p(v) = \mathrm{Ret}_p\{E_p v\}$. The chart loss, score, and Hessian are

$$\phi_{p,z}(v) = \ell\{\mathcal{R}_p(v); z\}, \qquad s_p(v, z) = \nabla_v \phi_{p,z}(v), \qquad H_p(v, z) = \nabla_v^2 \phi_{p,z}(v).$$

Their population counterparts are

$$F_p(v) = \mathbb{E}\{\phi_{p,Z}(v)\}, \qquad b_p(v) = \mathbb{E}\{s_p(v, Z)\}, \qquad A_p(v) = \mathbb{E}\{H_p(v, Z)\}.$$

For p sufficiently close to $\theta_\star$, the target coordinate is $v_\star(p) = \mathcal{R}_p^{-1}(\theta_\star)$. Thus $\mathcal{R}_p\{v_\star(p)\} = \theta_\star$, whereas $\mathcal{R}_p(0) = p$. Let $p_0 = p_0(\mathcal{D}_0)$ denote the initial anchor. An estimator with coordinate $\widetilde{v}$ has inferential error $\widetilde{v} - v_\star(p_0)$ in its chart. Conditionally on the initial estimate, this target coordinate is fixed and the worker scores remain independent. The observable origin supplies all worker summaries; the inverse chart enters the analysis and defines membership in the mapped confidence set.

Changing the orthonormal frame by $E'_p = E_p U_p$ with orthogonal U_p rotates scores and estimates by U_p^T and covariances by conjugation. When the required inverses exist and the surrogate uses the unique local minimizer in the chart, the mapped estimate, reporting decision, and confidence region remain unchanged (Proposition E.8).

2.2 LOCAL REGULARITY

Retractions and smooth local frames follow Absil et al. (2008). The following conditions justify differentiation under the expectation and control the shared chart. At the stationary target, the chart Hessian equals the intrinsic Hessian in the chosen frame. Write $d(p, q)$ for geodesic (shortest-path) distance between $p, q \in \mathcal{M}$ and $B_r(0)$ for the open ball in $\mathbb{R}^d$ of radius r centered at zero.

Assumption 2.1 (Local geometry)**.** There exist a neighborhood V of $\theta_\star$, a radius $r_0 > 0$, and constants $0 < c_R \le C_R < \infty$ such that the following hold.

(G1) The map $(p, v) \mapsto \mathcal{R}_p(v)$ is smooth on an open set containing $V \times \overline{B}_{r_0}(0)$, $\mathcal{R}_p(0) = p$, and $D\mathcal{R}_p(0) = E_p$.

(G2) For every $p \in V$, $\mathcal{R}_p$ is a diffeomorphism from $B_{r_0}(0)$ onto its image and

$$c_R \|u - v\| \le d(\mathcal{R}_p(u), \mathcal{R}_p(v)) \le C_R \|u - v\|$$

for all $u, v \in B_{r_0}(0)$.

(G3) The point $\theta_\star$ belongs to the image of every such chart, and $v_\star(p) \in B_{r_0/4}(0)$ for every $p \in V$.

Assumption 2.2 (Loss regularity)**.** For some $\rho \in (0, 1]$ and $\tau_s, \tau_H \in (0, 1]$, $v \mapsto \phi_{p,z}(v)$ is twice continuously differentiable on an open neighborhood of $\overline{B}_{r_0}(0)$ for every $p \in V$ and almost every z. There are nonnegative measurable envelopes B, M, L, C_s, C_H such that

$$\|s_p(0, z)\| \le B(z), \tag{2.1}$$

$$\sup_{v \in B_{r_0}(0)} \|H_p(v, z)\|_{\mathrm{op}} \le M(z), \tag{2.2}$$

$$\|H_p(u, z) - H_p(v, z)\|_{\mathrm{op}} \le L(z) \|u - v\|^\rho, \tag{2.3}$$

$$\|s_p(0, z) - s_q(0, z)\| \le C_s(z) d(p, q)^{\tau_s},$$

$$\|H_p(0, z) - H_q(0, z)\|_{\mathrm{op}} \le C_H(z) d(p, q)^{\tau_H} \tag{2.4}$$

for all $p, q \in V$, $u, v \in B_{r_0}(0)$ and almost every $z \in \mathcal{Z}$. Moreover,

$$\mathbb{E}\{B^2 + M^2 + C_s^2 + C_H^2\} < \infty, \qquad \mathbb{E}L < \infty, \qquad \mathbb{E}\,|\ell(\theta_\star; Z)| < \infty. \tag{2.5}$$

Assumption 2.3 (Regular target)**.** The target satisfies $b_{\theta_\star}(0) = 0, A_\star := A_{\theta_\star}(0) \succ 0$. Define $Q_\star = \mathbb{E}\{s_{\theta_\star}(0, Z) s_{\theta_\star}(0, Z)^\mathsf{T}\}, \Sigma_\star = A_\star^{-1} Q_\star A_\star^{-1}$. Full-dimensional Wald inference additionally assumes $Q_\star \succ 0$.

Assumption 2.4 (Independent initial estimate)**.** The initial anchor $p_0 = p_0(\mathcal{D}_0)$ is measurable, independent of the worker blocks, and, for a deterministic sequence $a_{n_0} \downarrow 0$,

$$d(p_0, \theta_\star) = O_p(a_{n_0}). \tag{2.6}$$

Positive definiteness of $A_\star$ makes $\theta_\star$ a strict local minimizer. Smooth changes of anchor preserve this minimizer while changing its coordinates. Condition (G3) holds on a sufficiently small anchor neighborhood: $v_\star(\theta_\star) = 0$ and continuity of the local inverse allows V to be shrunk until $\|v_\star(p)\| < r_0/4$ throughout V. A local empirical fit in a fixed admissible reference chart supplies a root-n_0 initial estimate (Propositions E.1 and E.2). The score second moment controls sampling fluctuation, the Hessian second moment controls the master Hessian, and $\mathbb{E}L < \infty$ controls the Taylor remainder.

2.3 Worker summaries and one-round estimators

Conditionally on p_0, worker j forms $f_j(v) = n^{-1} \sum_i \phi_{p_0, Z_{j,i}}(v)$ and sends its score $g_j = \nabla f_j(0)$ and second moment $S_j = n^{-1} \sum_i s_{p_0}(0, Z_{j,i}) s_{p_0}(0, Z_{j,i})^\mathsf{T}$. The pooled-Hessian procedure also sends $H_j = \nabla^2 f_j(0)$. The server aggregates

$$g_N = \frac{1}{m} \sum_{j=1}^m g_j, \qquad A_N = \frac{1}{m} \sum_{j=1}^m H_j, \qquad S_N = \frac{1}{m} \sum_{j=1}^m S_j, \qquad Q_N = S_N - g_N g_N^\mathsf{T}. \tag{2.7}$$

The matrix Q_N is precisely the empirical covariance, with divisor N, of all worker scores in the common chart. Pooling the uncentered moments before subtracting $g_N g_N^\mathsf{T}$ preserves variation between worker mean scores.

The centralized benchmark minimizes $f_N(v) = m^{-1} \sum_j f_j(v)$ over the regular chart:

$$\widehat{v}_N = \operatorname*{arg\,min}_{v \in B_{r_0}(0)} f_N(v), \qquad \widehat{\theta}_N = \mathcal{R}_{p_0}(\widehat{v}_N). \tag{2.8}$$

It uses the same N worker observations. Statistical error measures its distance to the population target; one-round approximation error measures the difference between a distributed estimate and this centralized fit.

Aggregated Newton estimator. When every worker transmits its Hessian, $\widetilde{v}_N^{\mathrm{N}} = -A_N^{-1} g_N, \widetilde{\theta}_N^{\mathrm{N}} = \mathcal{R}_{p_0}(\widetilde{v}_N^{\mathrm{N}})$. The corresponding covariance estimator is

$$\widehat{\Sigma}_N^{\mathrm{A}} = A_N^{-1} Q_N A_N^{-1}. \tag{2.9}$$

It estimates the root-N covariance $\Sigma_\star$; the coordinate covariance is $\widehat{\Sigma}_N^{\mathrm{A}}/N$.

Master-worker surrogate. Worker 1 is colocated with the coordinator and retains its objective. After aggregation it minimizes

$$\widetilde{f}_N(v) = f_1(v) - \langle g_1 - g_N, v \rangle . \tag{2.10}$$

The correction makes $\nabla \widetilde{f}_N(0) = g_N$ while leaving $\nabla^2 \widetilde{f}_N(v) = \nabla^2 f_1(v)$. The master therefore follows the pooled score using its own objective's Hessian. Its local minimizer and manifold image are

$$\widetilde{v}_N^{\mathrm{S}} = \operatorname*{arg\,min}_{v \in B_{r_0}(0)} \widetilde{f}_N(v), \qquad \widetilde{\theta}_N^{\mathrm{S}} = \mathcal{R}_{p_0}(\widetilde{v}_N^{\mathrm{S}}).$$

Using the master Hessian gives

$$\widehat{\Sigma}_N^{\mathrm{H}} = H_1^{-1} Q_N H_1^{-1}. \tag{2.11}$$

Newton with $\widehat{\Sigma}_N^{\mathrm{A}}$ is the pooled-Hessian procedure; each worker sends $d + d(d+1)$ scalars for g_j, S_j, H_j. Surrogate with $\widehat{\Sigma}_N^{\mathrm{H}}$ sends $d + d(d+1)/2$ scalars per nonmaster worker and is useful when communicating Hessians is costly. Its extra approximation term is controlled by the master sample size. Pairing the surrogate with $\widehat{\Sigma}_N^{\mathrm{A}}$ uses the pooled Hessian for uncertainty but retains the same surrogate point. We use Newton+A and surrogate+H as the two primary profiles in the experiments.

2.4 Confidence regions and reporting

Let $\chi^2_{d,1-\alpha}$ be the $(1-\alpha)$-quantile of χ^2_d. For a point $\widetilde{v}$ and positive-definite $\widehat{\Sigma}$, report the mapped Wald region

$$\mathcal{C}_{1-\alpha} = \left\{ \mathcal{R}_{p_0}(v) : N(v - \widetilde{v})^{\mathsf{T}} \widehat{\Sigma}^{-1} (v - \widetilde{v}) \le \chi^2_{d,1-\alpha} \right\}$$

when the entire ellipsoid lies in the chart. A sufficient computable check is

$$\|\widetilde{v}\| + \left\{ \frac{\chi^2_{d,1-\alpha} \lambda_{\max}(\widehat{\Sigma})}{N} \right\}^{1/2} < r_0. \tag{2.12}$$

A point is reportable when its Hessian inversion or local solve succeeds and it lies in the declared chart. A Wald region additionally requires positive-definite covariance and (2.12); otherwise the point is retained and the confidence set is empty. No-report events count as noncoverage. Algorithm 1 gives the complete protocol. The experiments use chart radii 1.5 for the spherical exponential map and 2 for Grassmann graph coordinates; the theory uses a sufficiently small neighborhood satisfying the assumptions.

2.5 Examples

Intrinsic Fréchet means. For $\ell(\theta; z) = d^2(\theta, z)/2$, the target minimizes expected squared geodesic distance. The cut locus of z consists of points where shortest geodesics from z cease to give a smooth squared-distance function. If candidate parameters and observations range over compact sets that stay away from these singular pairs, the loss derivatives are bounded. A positive-definite population Hessian at the target then gives the required local minimum. A sufficiently small strongly convex normal ball is a standard setting satisfying these conditions (Afsari, 2011; Bhattacharya and Patrangenaru, 2003; 2005; Bhattacharya and Lin, 2017). Centralized mean limit theory can also allow cut-locus observations under different conditions (Hotz et al., 2024).

Principal subspaces on a Grassmann manifold. Let $\mathcal{M} = \mathrm{Gr}(D, k)$ be the manifold of k-dimensional subspaces of $\mathbb{R}^D$ (Edelman et al., 1998). For $X \in \mathbb{R}^D$, use

$$\ell(U; X) = -\frac{1}{2} \left\| U^\mathsf{T} X \right\|^2, \tag{2.13}$$

where $U \in \mathbb{R}^{D\times k}$ is an orthonormal basis for the parameter subspace. The matrix U belongs to the Stiefel manifold; its span belongs to $\mathrm{Gr}(D, k)$. Since $\ell(UO; X) = \ell(U; X)$ for every orthogonal $k \times k$ matrix O, the loss depends only on that span (Edelman et al., 1998). Write $\lambda_1 \geq \cdots \geq \lambda_D$ for the eigenvalues of $\Sigma_X = \mathbb{E}(XX^\mathsf{T})$. An eigengap $\Delta = \lambda_k - \lambda_{k+1} > 0$ makes the leading subspace an isolated regular minimizer. In an eigenbasis, identify the $d = k(D - k)$ coordinates with $\mathrm{vec}\, K$, $K \in \mathbb{R}^{(D-k)\times k}$. The population Hessian acts in matrix form as

$$(A_\star K)_{ba} = (\lambda_a - \lambda_{k+b}) K_{ba}. \tag{2.14}$$

Writing $X_\parallel$ and $X_\perp$ for coordinates along the leading eigenspace and its orthogonal complement, the score matrix is the negative cross block $-X_\perp X_\parallel^\mathsf{T}$. Therefore $\mathbb{E}\|X\|^4 < \infty$ verifies the score second moment and the local derivative envelopes, while the positive eigengap makes the Hessian positive definite. These calculations are detailed in Section F.

3 Theoretical Results

Let $n_0, n, N \to \infty$ with fixed d and write $\delta_N = a_{n_0} + N^{-1/2}$, where $a_{n_0} \downarrow 0$ is the deterministic initial-estimate rate in (2.6). Conditional on the initial estimate, all workers use a fixed coordinate system. We first derive the centralized estimator's sampling law, then bound the error of each distributed approximation in these coordinates. Averaging over the initial estimate gives the unconditional conclusions. Proofs appear in Sections G to I.

Under Assumptions 2.1 to 2.4, the centralized minimizer is locally unique with probability tending to one and satisfies

$$\sqrt{N}\{\widehat{v}_N - v_\star(p_0)\} = -A_\star^{-1} N^{-1/2} \sum_{j,i} s_{\theta_\star}(0, Z_{j,i}) + o_p(1).$$

This is the regular M-estimation fluctuation with covariance $\Sigma_\star$ (Theorem E.6; van der Vaart, 1998, Chapter 5). Any consistent initial estimate suffices for this centralized law; a one-round correction additionally needs its anchor remainder to be small.

3.1 One-round approximation

The next theorem compares both computable procedures with the centralized fit $\widehat{v}_N$. Aggregating Hessians removes the discrepancy between the master and pooled Hessians; the initial-estimate error controls the remaining Taylor remainder.

Theorem 3.1. *Under Assumptions 2.1 to 2.4, the following hold with probability tending to one.*

(a) The aggregated Newton estimator is well defined and

$$\left\| \widetilde{v}_N^{\mathrm{N}} - \widehat{v}_N \right\| = O_p(\delta_N^{1+\rho}), \qquad d(\widetilde{\theta}_N^{\mathrm{N}}, \widehat{\theta}_N) = O_p(\delta_N^{1+\rho}). \tag{3.1}$$

(b) The master-worker surrogate has the expansion

$$\widetilde{v}_N^{\mathrm{S}} = -H_1^{-1} g_N + R_N^{\mathrm{S}}, \qquad \left\| R_N^{\mathrm{S}} \right\| = O_p(\delta_N^{1+\rho}), \tag{3.2}$$

and

$$\begin{aligned} \left\| \widetilde{v}_N^{\mathrm{S}} - \widehat{v}_N \right\| &= O_p\{\delta_N^{1+\rho} + n^{-1/2}\delta_N\}, \\ d(\widetilde{\theta}_N^{\mathrm{S}}, \widehat{\theta}_N) &= O_p\{\delta_N^{1+\rho} + n^{-1/2}\delta_N\}. \end{aligned} \tag{3.3}$$

(c) The first-order difference between the two one-round procedures is

$$\widetilde{v}_N^{\mathrm{S}} - \widetilde{v}_N^{\mathrm{N}} = (A_N^{-1} - H_1^{-1}) g_N + O_p(\delta_N^{1+\rho}). \tag{3.4}$$

The term $\delta_N^{1+\rho}$ is the error left by linearizing the pooled score at the initial estimate. Here a_{n_0} measures initial-estimate displacement and $N^{-1/2}$ measures sampling fluctuation; the Hessian smoothness exponent ρ determines how their sum propagates through one Newton step. For the surrogate, $H_1 - A_N = O_p(n^{-1/2})$ additionally multiplies a score of order δ_N. Sending pooled Hessians removes this Hessian-estimation term; retaining the master objective saves communication at a cost determined by the local sample size.

Corollary 3.2. *Under the assumptions of Theorem 3.1, the Newton estimator is root-N equivalent to the centralized estimator $\widehat{v}_N$ if*

$$\sqrt{N}\,\delta_N^{1+\rho} \longrightarrow 0. \tag{3.5}$$

The surrogate is root-N equivalent if, in addition,

$$\sqrt{N}\,n^{-1/2}\delta_N \longrightarrow 0. \tag{3.6}$$

That is, under the corresponding condition,

$$\sqrt{N}\{\widetilde{v}_N^{\bullet} - \widehat{v}_N\} = o_p(1), \qquad \sqrt{N}\,d(\widetilde{\theta}_N^{\bullet}, \widehat{\theta}_N) = o_p(1), \qquad \bullet \in \{\mathrm{N}, \mathrm{S}\}.$$

If $n_0 = n$ and the initial estimate is root-n, then $\delta_N = O(n^{-1/2})$ and $m = o(n^{\rho})$ implies both conditions. In particular, Lipschitz Hessians ($\rho = 1$) give $m = o(n)$.

For a root-n_0 initial estimate, $N = o(n_0^{1+\rho})$ ensures (3.5); adding $m = o(n_0)$ ensures (3.6). Thus $n_0 = T^{\gamma}$ with $1/(1+\rho) < \gamma < 1$ gives Newton equivalence and $T/N \to 1$; surrogate equivalence holds if $m = o(T^{\gamma})$. With Lipschitz Hessians, the initial-sample size can lie between the $\sqrt{T}$ and T scales.

In Euclidean coordinates, the surrogate reduces to communication-efficient surrogate likelihood (Jordan et al., 2019); Proposition E.9 gives the exact objective identity. The random common chart extends this comparison to manifold parameters while preserving a single population target.

The exponent $1 + \rho$ is attained by a one-dimensional population Newton example (Proposition E.7); quadratic losses eliminate this remainder exactly.

3.2 Asymptotic inference

Write $r_N^{\mathrm{N}} = \delta_N^{1+\rho}$ and $r_N^{\mathrm{S}} = \delta_N^{1+\rho} + n^{-1/2}\delta_N$ for the two approximation rates.

Theorem 3.3 (One-round inference)**.** *Suppose Assumptions 2.1 to 2.4 hold.*

(a) If $\sqrt{N} r_N^{\bullet} \to 0$ for $\bullet \in \{\mathrm{N}, \mathrm{S}\}$, then

$$\sqrt{N}\{\widetilde{v}_N^{\bullet} - v_\star(p_0)\} = -A_\star^{-1}\frac{1}{\sqrt{N}}\sum_{j=1}^{m}\sum_{i=1}^{n} s_{\theta_\star}(0, Z_{j,i}) + o_p(1), \tag{3.7}$$

and hence

$$\sqrt{N}\{\widetilde{v}_N^{\bullet} - v_\star(p_0)\} \Longrightarrow N(0, \Sigma_\star). \tag{3.8}$$

(b) Irrespective of the root-N equivalence condition,

$$A_N \xrightarrow{p} A_\star, \qquad H_1 \xrightarrow{p} A_\star, \qquad Q_N \xrightarrow{p} Q_\star, \tag{3.9}$$

and therefore

$$\widehat{\Sigma}_N^{\mathrm{A}} \xrightarrow{p} \Sigma_\star, \qquad \widehat{\Sigma}_N^{\mathrm{H}} \xrightarrow{p} \Sigma_\star. \tag{3.10}$$

(c) If $Q_\star \succ 0$ and $\widehat{\Sigma}_N$ is either consistent estimator in (3.10)*, then, under the rate condition for $\bullet$,*

$$N\{\widetilde{v}_N^{\bullet} - v_\star(p_0)\}^{\mathsf{T}}\widehat{\Sigma}_N^{-1}\{\widetilde{v}_N^{\bullet} - v_\star(p_0)\} \Longrightarrow \chi_d^2. \tag{3.11}$$

Consequently, the confidence set returned in Algorithm 1 satisfies

$$\mathbb{P}\{\theta_\star \in \mathcal{C}_{1-\alpha}^{\bullet}\} \longrightarrow 1 - \alpha. \tag{3.12}$$

Both covariance profiles consistently estimate the centralized limiting covariance using only the common-anchor summaries. Wald calibration also requires the distributed approximation error to vanish at the root-N scale: consistency of the covariance alone does not center a confidence region at the target.

When $Q_\star$ is singular, the linear limit still holds and Wald inference can be restricted to identified linear functionals.

For a smooth feature $h : \mathcal{M} \to \mathbb{R}^q$, its Jacobian $\widehat{J}_N = D(h \circ \mathcal{R}_{p_0})(\widetilde{v}_N^\bullet)$ gives estimated covariance $\widehat{J}_N \widehat{\Sigma}_N \widehat{J}_N^\mathsf{T}/N$ without additional worker summaries. Corollary E.5 states the delta-method limit; the application uses diagonal entries of a subspace projection. Euclidean, one-worker and quadratic reductions are given in Proposition E.9.

3.3 Exact PCA recovery and covariance choice

PCA has an algebraic structure beyond the general one-step approximation. Use the graph chart $U(K) = (U_0 + V_0 K)(\mathrm{I}_k + K^\mathsf{T} K)^{-1/2}$, where U_0 and V_0 are orthonormal bases for the anchor and its complement. Here $K \in \mathbb{R}^{q \times k}$ with complement dimension $q = D - k$. Write $y = U_0^\mathsf{T} X$ and $z = V_0^\mathsf{T} X$. Let $\widehat{\Sigma}_X$ be the worker second-moment matrix in this basis and $\widehat{\Sigma}_{zz} = N^{-1} \sum_{j,i} z_{j,i} z_{j,i}^\mathsf{T}$. Let G_N be the $q \times k$ matrix form of g_N, and $(A_N)_{ab}$ the $q \times q$ Hessian block for graph columns a, b. Define $B_{ab} = \mathrm{tr}\{(A_N)_{ab}\}/q$ and $C = k^{-1} \sum_a \{B_{aa} \mathrm{I}_q - (A_N)_{aa}\}$. Then

$$\begin{pmatrix} B & -G_N^\mathsf{T} \\ -G_N & C \end{pmatrix} = \widehat{\Sigma}_X - \frac{\mathrm{tr}(\widehat{\Sigma}_{zz})}{q} \mathrm{I}_D. \tag{3.13}$$

Thus (3.13) recovers its leading eigenspace exactly whenever the empirical eigengap is positive (Proposition F.1).

The scalar shift leaves eigenvectors unchanged. An orthonormal representative $\widehat{U}$ of the recovered subspace has initial-chart coordinate $K = (V_0^\mathsf{T} \widehat{U})(U_0^\mathsf{T} \widehat{U})^{-1}$ when the inverse exists. The server applies the same chart and ellipsoid checks at this corrected point, using its existing summaries.

Direct one-round recentring replaces the Newton center by this exact fit's coordinate $\widehat{v}_N$, retaining the initial-chart covariance $\widehat{\Sigma}_N^{\mathrm{A}}$ and the reporting checks. A dedicated two-round protocol instead collects ambient second moments, broadcasts their leading eigenspace, and collects score second moments in that fitted chart; the first-round matrix supplies the Hessian and the pooled score is zero. The latter needs no initial-sample collection.

For fixed dimension, a positive population eigengap, finite fourth moment and $Q_\star \succ 0$, both yield limiting coverage $1 - \alpha$. The one-round version requires only a consistent independent initial estimate: exact recentring removes the Newton approximation and its stronger initial-rate condition. The two central-centered Wald statistics are asymptotically equivalent, also allowing an observable choice between them (Proposition F.2).

4 Simulations

We compare Newton+A, surrogate+H and centralized fitting on shared worker observations. Nominal coverage is 0.95; every no-report event counts as noncoverage. Point errors condition on return and measure geodesic distance to the target; centralized gaps are $\sqrt{N}$ times distance to the worker-sample fit. Smaller errors and gaps indicate better approximation and accuracy; coverage is assessed against its nominal level. Each design uses 500 replications, giving coverage Monte Carlo standard error at most 0.023 (about 0.010 at nominal coverage). Sections B and C give distributions, reporting counts and matched-budget details.

4.1 Localization and Hessian estimation

Spherical intrinsic means test the general nonlinear construction. At fixed $T = 4000$ and $m = 20$, we compare initial-sample sizes 20 and 400 under isotropic and asymmetric tangent distributions (Section C.1 and Table 4). All methods return all 500 regions per cell. One-round coverage differs

Table 1: Coverage on $N = 16000$ worker observations with 500 replications per setting; $m = 16000/n$ and the eigengap is Δ. Nominal coverage is 0.95 and failures count as noncoverage. The three methods use the same worker draws.

Law	n	n_0	Δ	Central	Newton	Surrogate
Gaussian	500	2000	0.5	0.952	0.924	0.178
Gaussian	2000	4000	0.5	0.936	0.934	0.856
Gaussian	500	1000	0.5	0.956	0.846	0.038
t_8	500	1000	0.5	0.924	0.682	0.008
Gaussian	500	1000	1	0.928	0.908	0.268

from worker-centralized coverage by at most 0.010. Updating the surrogate anchor in a second round reduces its centralized gaps by more than fortyfold, with little change in target root mean squared error (RMSE). Thus a much closer empirical fit need not materially improve statistical accuracy once its approximation error is small.

Table 1 shows the more demanding $\mathrm{Gr}(20, 3)$ problem. At eigengap 0.5, increasing the initial-sample size from 1000 to 2000 with fixed $(n, m) = (500, 32)$ raises Newton coverage from 0.846 to 0.924 and reduces its mean scaled gap from 7.49 to 2.27. At initial-sample size 1000, increasing the eigengap from 0.5 to 1 raises coverage from 0.846 to 0.908. These changes agree with the role of localization and the inverse Hessian in the one-round error.

The same-worker centralized comparison separates baseline Wald calibration from one-round approximation. At initial-sample size 1000 and eigengap 0.5, Gaussian and t_8 observations give Newton coverage 0.846 and 0.682, while central coverage is 0.956 and 0.924. Both laws satisfy the required fourth-moment condition. Tail sensitivity therefore includes both centralized Wald error and additional approximation error. The surrogate reaches 0.856 when $n = 2000$ and $n_0 = 4000$, where both the worker block and initial sample increase. For Gaussian data with $n_0 = 1000$ and eigengap 0.5, its 230 returned regions cover only 19 times, showing substantial miscalibration among reported regions.

A fixed-total-sample comparison separates better localization from increased data use. At $T = 19200$ and $m = 32$, increasing the initial-sample fraction from 5% to 20% raises Newton coverage from 0.804 to 0.924, even as observations per worker decrease from 570 to 480 (Section C.2 and Table 5). Newton RMSE falls from 0.126 to 0.091, while worker-centralized RMSE rises from 0.080 to 0.088 as less data remain for inference. The surrogate still reports only 304 regions with coverage 0.382 at the largest initial sample.

4.2 Exact subspace recovery: rounds versus bytes

At $T = 19200$, $n_0 = 3840$ and $m = 32$, an independent paired comparison tests the exact PCA protocols (Table 2). On Gaussian samples, recentring raises coverage from Newton's 0.914 to 0.938 without another report and recovers the worker-centralized point. Dedicated two-round PCA also matches the fitted-chart covariance, using 0.40 MB versus 2.38 MB for raw worker data. The table separates locally available and remotely collected initial samples.

For t_5, recentring returns 462 regions with coverage 0.906, versus 492 and 0.890 for two rounds; its initial-chart radii are larger. Section A.1 studies an adaptive covariance update. With a local initial sample, recentring saves one upload stage at an extra 0.29 MB; communication favors it when that stage's latency exceeds the extra byte cost. With a remote initial sample, both paths require two uploads and recentring costs 0.90 MB more. Section C.3 gives the complete paths and linear communication-cost model.

5 Dry Bean application

The UCI Dry Bean data contain 13611 grains from seven varieties and 16 morphological measurements, with no missing entries (Dry Bean, 2020; Koklu and Ozkan, 2020). We estimate the pooled covariance subspace of the natural logarithms of the six positive, unstandardized features in Table 3. For independent copies X_1, X_2 of the pooled log-feature vector,

$$Y = (X_1 - X_2)/\sqrt{2}, \qquad \mathbb{E}Y = 0, \qquad \mathrm{Cov}(Y) = \mathrm{Cov}(X_1).$$

Table 2: PCA protocols on $N = 15360$ worker observations, with 500 paired replications per law. Cov. includes no-report events. Uploads counts sequential upload stages to the final output; MB includes broadcasts. Entries a/b denote an initial sample already at the coordinator / collected remotely. Dedicated PCA uses no initial sample.

Law	Method	Regions	Cov.	Uploads	MB
Gaussian	Newton+A	500	0.914	1/2	0.69/1.30
	Recenter, one round	500	0.938	1/2	0.69/1.30
	PCA, two rounds	500	0.924	2	0.40
	Raw central workers	500	0.924	1	2.38
t_5	Newton+A	459	0.842	1/2	0.69/1.30
	Recenter, one round	462	0.906	1/2	0.69/1.30
	PCA, two rounds	492	0.890	2	0.40
	Raw central workers	492	0.890	1	2.38

Disjoint pairs therefore allow second-moment PCA without estimating a common mean. We allocate 500 pairs to the initial sample and 6000 pairs to 20 workers, using 13000 grains and excluding 611 to form equal blocks. The target is a rank-two subspace on $\mathrm{Gr}(6, 2)$; each Newton worker sends 80 scalars. Section D gives preprocessing and the functional Jacobian.

For an orthonormal representative U, the projection leverage $h_a(U) = (UU^{\mathsf{T}})_{aa}$ is the squared length of feature axis a projected onto the subspace, with $\sum_a h_a = 2$. Table 3 gives pointwise delta-method intervals from the common-anchor Newton estimate. Area and both axis lengths have the largest leverages; Solidity has little alignment on this unstandardized log scale.

Table 3: Dry Bean application: diagonal entries of the estimated rank-two projection and pointwise delta-method uncertainty. Features are analyzed on the log scale; paired differences remove the unknown mean.

Feature	Projection leverage	Standard error	95% interval
Area	0.56635	0.00032	[0.56572, 0.56699]
Perimeter	0.18436	0.00068	[0.18303, 0.18569]
MajorAxisLength	0.52758	0.00344	[0.52084, 0.53432]
MinorAxisLength	0.58376	0.00427	[0.57539, 0.59212]
Extent	0.13785	0.00767	[0.12281, 0.15289]
Solidity	0.00010	0.00001	[0.00008, 0.00012]

The centralized worker fit explains 0.988 of the log-feature variance and has second–third eigengap 0.0083. Newton's geodesic distance to it is 0.000154, compared with 0.00142 for projection averaging and 0.00634 for the master surrogate. These distances measure approximation to the worker-sample fit. The intervals describe the pooled subspace under an i.i.d. working model; the dataset supplies no sampling design for design-based uncertainty.

6 Concluding Remarks

A common independent anchor connects one-round manifold estimation to centralized uncertainty by separating localization error from master-Hessian error. The same summaries yield frame-invariant Wald regions; in PCA they also recover the exact empirical subspace. Nonlinear experiments support centralized calibration when approximation is small; the subspace results translate initial-sample and covariance choices into explicit communication costs.

Cross-fitting could reuse the initial sample for inference, so observations used to locate the common chart also contribute to estimating the target within it. This would require a joint expansion accounting for dependence between anchors and worker scores induced by shared observations. A higher-order correction could address the boundary $\sqrt{N}\delta_N^{1+\rho} \asymp 1$, where the localization remainder contributes on the same scale as the centralized sampling fluctuation.

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

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

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

Bhattacharya, R. and Lin, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proceedings of the American Mathematical Society*, **145**, 413–428.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *The Annals of Statistics*, **31**, 1–29.

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

Brunel, V.-E. (2023). Geodesically convex M-estimation in metric spaces. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, 2188–2210.

Dry Bean [Dataset]. (2020). UCI Machine Learning Repository.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, **20**(2), 303–353.

Fan, J., Guo, Y., and Wang, K. (2023). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, **118**, 1000–1010.

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

Hotz, T., Le, H., and Wood, A. T. A. (2024). Central limit theorem for intrinsic Fréchet means in smooth compact Riemannian manifolds. *Probability Theory and Related Fields*, **189**, 1219–1246.

Huang, L.-K. and Pan, S. (2020). 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*, 4465–4474.

Huang, C. and Zhang, A. R. (2026). High-order accurate inference on manifolds. *The Annals of Statistics*, **54**(4), 1712–1739.

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

Koklu, M. and Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, **174**, 105507.

Kurisu, D. and Otsu, T. (2026). Empirical likelihood for manifolds. *Journal of the Royal Statistical Society: Series B*, **88**, 91–119.

Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, **25**(1), 127–154.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, **5**(4), 379–404.

Saparbayeva, B., Zhang, M. M., and Lin, L. (2018). Communication efficient parallel algorithms for optimization on manifolds. In *Advances in Neural Information Processing Systems 31*, 3574–3584.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, 1000–1008.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, **14**, 3321–3363.

The appendices begin with the reporting protocol (Section A), followed by experimental designs, fixed-budget comparisons and application implementation (Sections B to D). Additional theoretical results and model calculations precede the supporting lemmas and proofs (Sections E to I).

A IMPLEMENTATION DETAILS

Inputs: independent initial sample $\mathcal{D}_0$; worker blocks $\mathcal{D}_1, \ldots, \mathcal{D}_m$; a local retraction/frame and chart radius r_0; confidence level $1-\alpha$; estimator mode N or S; covariance mode A or H.

1. Compute a feasible initial estimate p_0 by a procedure satisfying (2.6). If the initial estimate or its declared chart cannot be computed, return a no-report status; otherwise broadcast p_0 and the chart convention.
2. Every worker computes g_j and S_j. In Newton mode or pooled-Hessian mode it also computes H_j. The master retains f_1 and H_1.
3. Aggregate g_N, S_N, Q_N and, when requested, A_N according to (2.7).
4. In mode N, form the candidate $\widetilde{v} = -A_N^{-1} g_N$ only if A_N is positive definite. In mode S, solve (2.10) for its designated local minimizer $\widetilde{v}$. The numerical initialization and acceptance checks are specified below.
5. Use (2.9) in covariance mode A and (2.11) in mode H, provided the required inverse and the resulting covariance are positive definite.
6. Assess point and region reporting separately. A point is reportable only if its required Hessian inversion or local solve succeeds and the point lies in the declared chart. It is Wald-reportable only if, in addition, the covariance is positive definite and the containment check (2.12) passes. On a point failure, return a no-report status; a computable off-chart candidate may be retained only as a diagnostic. On a Wald-only failure, return the point with a localization warning but no full-dimensional confidence region.
7. Otherwise return $\widetilde{\theta} = \mathcal{R}_{p_0}(\widetilde{v})$ and

$$\mathcal{C}_{1-\alpha} = \left\{ \mathcal{R}_{p_0}(v) : v \in B_{r_0}(0),\ N(v-\widetilde{v})^{\mathsf{T}} \widehat{\Sigma}^{-1} (v-\widetilde{v}) \le \chi^2_{d,1-\alpha} \right\}.$$

Failure rule. A no-report event counts as noncoverage. The numerical implementation checks eigenvalues before inversion and records solver failures separately from chart failures. It does not replace a failed inverse by a regularized matrix when reporting the stated Wald region.

Algorithm 1: Common-anchor one-round manifold inference.

Numerical surrogate selection. The exploratory sphere and Grassmann implementations initialize BFGS at $-H_1^{-1} g_N$ when the master Hessian passes the positive-definiteness check, otherwise at $-A_N^{-1} g_N$ if that inverse passes, and otherwise at zero. They penalize trial points outside the declared coordinate ball and accept only finite interior solutions with successful optimizer status. The sphere routine uses radius 1.5, numerical derivatives, relative objective tolerance 10^{-10} and at most 150 iterations; it has no additional gradient-norm acceptance check. The Grassmann routine uses radius 2, an analytic gradient and relative tolerance 10^{-14}, restarting from the last iterate for at most 12 passes of 250 iterations and requiring final gradient norm below 10^{-5}. The fixed-budget sphere comparison instead requires the master inverse for initialization and uses the analytic-gradient and residual checks in Section C.1; a failed master inverse gives no report. Point return and Wald-region return then obey the separate checks in Algorithm 1. These numerical checks do not establish that a returned stationary point is the theoretical local branch: identification with that branch uses the local strong-convexity and localization events in Sections G and H.

Whole-sample convention. On a point failure in Algorithm 1, assign coordinate zero for theoretical statements and return a no-report status. On a Wald-only failure, retain the point and define its confidence set to be empty. Extend $v_\star(p_0)$ by zero when $p_0 \notin V$. These extensions define random

elements on the whole sample space; the point and Wald theorems show that their respective failure probabilities vanish. In uniform-law proofs, p_0° denotes p_0 on $\{p_0 \in V\}$ and $\theta_\star$ otherwise, so chart-indexed quantities remain defined and $p_0^\circ = p_0$ with probability tending to one.

Remark A.1 (Initial-sample cost and the estimand). The worker-only estimator has first-order covariance $\Sigma_\star/N$, whereas an ideal centralized estimator using all $T = N + n_0$ observations has first-order covariance $\Sigma_\star/T$. The relative first-order variance factor is therefore $T/N = 1 + n_0/N$. Discarding the initial sample is asymptotically free only when $n_0/N \to 0$. Recycling initial-sample observations is possible through cross-fitting or a joint expansion, but it changes the dependence structure and is not claimed by the present one-round theorem.

A.1 Optional PCA covariance update

The optional covariance update increases the t_5 return count from 462 to 492 and changes coverage from 0.906 to 0.890. For Gaussian data it adds 220 second rounds and changes coverage from 0.938 to 0.932. An optional adaptive protocol compares the Newton point with the worker-centralized PCA point recovered from its summaries by Proposition F.1, in estimated standard-error units:

$$\eta_N = \left[N(\widetilde{v}_N^{\mathrm{N}} - \widehat{v}_N)^{\mathsf{T}} (\widehat{\Sigma}_N^{\mathrm{A}})^{-1} (\widetilde{v}_N^{\mathrm{N}} - \widehat{v}_N) \right]^{1/2}. \tag{A.1}$$

For a tolerance $\varepsilon > 0$ chosen before analysis, it keeps the initial-chart covariance when $\eta_N \le \varepsilon$ and the ellipsoid centered at $\widehat{v}_N$ passes the chart check. Otherwise it broadcasts the recovered fit and collects fitted-chart score moments in a second round. Missing first-round inverses or graph coordinates also trigger this covariance update. We study $\varepsilon = 0.25$, a quarter-standard-error displacement. The statistic measures the one-round point approximation; the paired numerical comparison assesses what its use as a rule for updating covariance achieves.

B Experimental design and reporting

The primary profiles are Newton with the pooled Hessian (Newton+A) and the corrected master objective with the master Hessian (surrogate+H). All comparisons use the same observations within a replication. Worker-centralized fits exclude initial-sample data; all-sample fits include them. Centralized reporting is independent of whether a distributed method returns. The original simulation family uses analytic derivatives in R 4.4.3. Method settings and reporting thresholds were fixed after a 20-replication preliminary run and before evaluation; the matched designs retain those settings. The accompanying code and saved replicate records specify seeds and numerical tolerances. Point return requires the method-specific Hessian inversion or solver to succeed and the point to lie in its declared chart. Region return additionally requires positive-definite covariance and (2.12). The region radius is 1.5 on the sphere and 2 in the Grassmann graph chart. These reporting checks use fitted quantities.

Unconditional coverage divides the number of containing regions by all 500 replications. Conditional coverage divides it by the number of returned regions. Target RMSE and mean target distance condition on point return. Fixed-budget centralized gaps also condition on point return; the exploratory comparisons in Table 1 average gaps over finite candidates, including ones subsequently rejected. This distinction matters when reportability varies. For an estimated coverage $\widehat{p}$, its Monte Carlo standard error is $\{\widehat{p}(1-\widehat{p})/500\}^{1/2}$, at most 0.023 and about 0.010 at 0.95.

The Grassmann target is the leading rank-three subspace of a 20-dimensional covariance with spectrum

$$(5, 4, 3, 3 - \Delta, \lambda_5, \ldots, \lambda_{20}), \qquad \lambda_5 = \min\{2.2, 2.8 - \Delta\},$$

where $\lambda_5, \ldots, \lambda_{20}$ are equally spaced down to 0.8. The relevant eigengap is exactly Δ. Gaussian observations have this covariance. Standardized t_ν observations are obtained by multiplying the Gaussian vector by $\sqrt{(\nu-2)/W}$, independently with $W \sim \chi^2_\nu$. The t_8 and t_5 comparisons both satisfy the finite-fourth-moment condition. Initial subspaces are leading eigenspaces of their own second moments. Surrogate optimization uses the master-Hessian, pooled-Hessian, then zero initialization sequence described in Section A.

The G11–G12 settings are exploratory; the tail, eigengap and initial-sample comparisons G14–G19 were specified after inspecting exploratory settings and evaluated on new replications with fixed method settings. G12 jointly enlarges n and n_0. The separate numerical supplement retains complete point and region counts. These distinguish failure to supply an interval from miscalibration of intervals actually supplied. For example, G14 Newton has 423 containing regions out of 486 returned, whereas its surrogate has 19 out of 230. The corresponding unconditional coverages include no-report events. The fixed-total-sample and exact-PCA comparisons below use separate paired experiments. In the condition-exceeding t_3 control G17, Newton and the surrogate each have zero coverage; the worker-centralized coverage is 0.124. The oracle-anchor G13 gives surrogate coverage 0.700, or $350/369 = 0.949$ among reported regions, separating reporting failure from conditional calibration.

Table 4: Complete fixed-total-sample comparison. Coverage is unconditional; returned gives point/Wald counts out of 500. RMSE is root mean squared geodesic error to the target; gap is mean $\sqrt{N}$-scaled distance to the worker-centralized fit. Radius is the mean largest coordinate semi-axis of returned regions. RMSE and gap condition on point return.

Method	Coverage	Returned	RMSE	Gap	Radius
Isotropic, $n_0 = 20$					
Newton+A	0.954	500/500	0.0185	0.0798	0.0325
Surrogate+H	0.948	500/500	0.0186	0.231	0.0327
Two-round surrogate	0.960	500/500	0.0182	0.00505	0.0318
Centralized workers	0.958	500/500	0.0182	0	0.0315
Centralized all	0.962	500/500	0.0181	0.069	0.0315
Isotropic, $n_0 = 400$					
Newton+A	0.976	500/500	0.0177	0.000849	0.0332
Surrogate+H	0.972	500/500	0.0177	0.0512	0.0336
Two-round surrogate	0.970	500/500	0.0176	0.00118	0.0336
Centralized workers	0.976	500/500	0.0177	0	0.0332
Centralized all	0.964	500/500	0.0167	0.312	0.0315
Asymmetric, $n_0 = 20$					
Newton+A	0.948	500/500	0.0111	0.00474	0.0263
Surrogate+H	0.944	500/500	0.0112	0.0188	0.0263
Two-round surrogate	0.950	500/500	0.0111	0.00019	0.0263
Centralized workers	0.948	500/500	0.0111	0	0.0263
Centralized all	0.948	500/500	0.0112	0.0405	0.0262
Asymmetric, $n_0 = 400$					
Newton+A	0.930	500/500	0.0124	0.000258	0.0277
Surrogate+H	0.930	500/500	0.0124	0.00437	0.0277
Two-round surrogate	0.934	500/500	0.0124	4.45×10^{-5}	0.0277
Centralized workers	0.932	500/500	0.0124	0	0.0277
Centralized all	0.934	500/500	0.0119	0.181	0.0262

C Fixed-budget comparisons

C.1 Sample and communication budgets on the sphere

At a fixed total sample, enlarging the initial sample improves localization while leaving fewer observations for inference. The worker-centralized fit isolates the communication approximation for each allocation; comparison with the all-sample fit measures the cost of assigning observations to the initial sample.

A separate paired comparison fixes the total sample at $T = 4000$ and the worker count at $m = 20$. Initial-sample sizes $n_0 = 20$ and 400 leave $n = 199$ and 180 observations per worker. Each allocation uses 500 replications under each of two tangent distributions at the north pole: the uniform disk of radius 1.35, and the asymmetric mean-zero mixture used in the spherical designs. In the latter, the first coordinate is shifted by $0.35r$ with probability $2/3$ or by $-0.70r$ otherwise, and independent uniform-disk noise has radius $0.15r$, with $r = 1.35$. The exponential map gives the observations; both distributions have a regular intrinsic-mean target at the north pole. All five methods use the same realized observations within each replication.

The two-round surrogate broadcasts the first surrogate estimate as a new anchor, recomputes the worker scores and the master Hessian there, and solves the corrected master objective again. Its final covariance uses the master Hessian and the pooled score covariance at this second anchor. The first round needs only scores to compute its point; score second moments are sent in the final round. This comparator measures the empirical gain from an anchor updated using worker data. The conditional-independence analysis in the main theorem applies to the original independent anchor. Centralized fits use either the N worker observations or all T observations and evaluate the covariance in their own fitted chart. All regions use the same radius-1.5 containment rule; the target enters only the coverage calculation.

Every method returns a point and region in every replication. At nominal coverage 0.95, the Monte Carlo standard error from 500 replications is 0.010. The extra surrogate round reduces approximation error in all four cells; at $n_0 = 400$, the all-sample centralized fit also recovers the information assigned to localization, reducing its mean region radius by about 5% relative to the worker-only fit.

Initial and centralized means start from the normalized ambient sample mean and iterate logarithm averaging, stopping at gradient norm 10^{-10} within 100 iterations. Surrogate objectives use BFGS quasi-Newton optimization with an analytic pullback gradient, relative objective tolerance 10^{-10}, and at most 150 iterations; a returned solution also has gradient norm at most 10^{-5}. Newton uses one linear solve after aggregation. The one-round methods require one worker pass at the anchor, while the two-round surrogate requires two; the initial fit and master optimization are separate computations.

Communication counts 64-bit scalars, with worker 1 colocated with the coordinator. Collecting the remote initial-sample costs $3n_0$ scalars. An anchor broadcast sends three coordinates to each of the other 19 workers; a deterministic frame requires no further transmission. Newton workers return two score entries and two symmetric 2×2 matrices, whereas surrogate workers omit the Hessian. Thus Newton, one-round surrogate, and two-round surrogate respectively cost $8\{3n_0 + 11(m-1)\}$, $8\{3n_0 + 8(m-1)\}$, and $8\{3n_0 + 13(m-1)\}$ bytes. Centralized worker-only fitting uploads $24(m-1)n$ bytes of raw observations; the all-sample fit adds $24n_0$. These byte counts include anchor broadcasts and initial-sample collection, exclude transport headers, and compare the actual communication required at a common sample budget.

C.2 Initial-sample allocation at a fixed Grassmann sample budget

This comparison examines the same localization–information trade-off for a subspace, where a small eigengap makes initial-estimate error consequential. Shared total samples allow changes across allocations to be compared against an unchanged all-sample fit.

We use the Gaussian $\mathrm{Gr}(20, 3)$ distribution with eigengap 0.5 from Section B, fixing $T = 19200$ and $m = 32$. Initial-sample fractions 5%, 10%, and 20% give $n_0 = 960, 1920, 3840$ and $n = 570, 540, 480$. Each of 500 independent total samples is shared across methods and allocations. Its first n_0 observations supply the initial eigenspace; the rest form equal worker blocks, with worker 1 as master. Initial sample and workers are independent within each allocation.

Newton+A and surrogate+H use analytic Grassmann derivatives and the radius-2 rule. The surrogate starts from its master-Hessian step when available and from zero otherwise. Centralized fits evaluate their covariance at the fitted subspace using worker-only or all-sample data. Reporting uses fitted quantities; the target enters only the coverage calculation.

Communication uses 64-bit scalars and colocates the master with the coordinator. Initial-sample collection costs $20n_0$ scalars; broadcasting a 20×3 anchor costs 60 per remote worker, which reconstructs its complementary frame. With $d = 51$, Newton sends 51 score entries and two symmetric matrices of 1326 distinct entries each; the surrogate omits the Hessian. Centralized methods collect raw data from 31 remote workers, adding the initial sample for the all-sample fit. PCA also permits exact worker-sample point estimation by aggregating 210 ambient second-moment entries per remote worker, costing 52 kB; the centralized columns include raw-data collection for fitted-chart covariance.

As the initial-sample fraction grows, worker-centralized RMSE rises from 0.080 to 0.088, while Newton's falls from 0.126 to 0.091. Surrogate coverage improves but retains point failures and miscalibration among returned regions. Unconditional coverage has Monte Carlo standard error at most 0.023, and 0.010 at nominal 0.95. The all-sample fit is identical across allocations within each replication.

C.3 PCA recentring and fitted-chart covariance

The independent comparison in Table 2 uses $T = 19200$, $n_0 = 3840$, $m = 32$, $D = 20$, $k = 3$ and eigengap 0.5. Gaussian and standardized t_5 laws have the spectrum in Section B; each has 500 independent paired replications. The first n_0 observations form the initial sample, and the remaining $N = 15360$ are partitioned into equal worker blocks. No method uses the population subspace for its reporting decision. Direct recentring uses the recovered exact worker PCA fit, initial-chart covariance

Table 5: Fixed-total Grassmann comparison: $T = 19200$, $m = 32$, Gaussian observations and eigengap 0.5, with 500 paired replications per allocation. Points and Regions are return counts; coverage includes every no-return event, and Cond. is coverage among returned regions. RMSE and the mean $\sqrt{N}$-scaled gap to the worker-centralized fit use returned points. C_N and C_T use worker-only and all-sample centralized fits. Communication cost includes initial-sample collection and broadcasts for the one-round methods and raw-data collection for centralized fitting; 1000 bytes per kB.

n_0	Method	Points	Regions	Coverage	Cond.	RMSE	Gap	kB
960	Newton+A	495	489	0.804	0.822	0.126	7.85	839
960	Surrogate+H	268	234	0.040	0.085	0.159	16.82	510
960	C_N	500	500	0.944	0.944	0.080	0.00	2827
960	C_T	500	500	0.942	0.942	0.078	2.38	2981
1920	Newton+A	500	499	0.892	0.894	0.096	2.81	992
1920	Surrogate+H	340	310	0.162	0.261	0.145	13.63	664
1920	C_N	500	500	0.940	0.940	0.083	0.00	2678
1920	C_T	500	500	0.942	0.942	0.078	3.40	2986
3840	Newton+A	500	500	0.924	0.924	0.091	0.90	1300
3840	Surrogate+H	328	304	0.382	0.628	0.135	10.20	971
3840	C_N	500	500	0.938	0.938	0.088	0.00	2381
3840	C_T	500	500	0.942	0.942	0.078	4.72	2995

and the radius-2 containment check. Dedicated two-round PCA collects ambient second moments and then score moments at their leading subspace, without requiring initial-sample collection. It uses the same N worker observations, so its results isolate covariance and communication choices rather than additional inferential data.

Table 6: Additional point and region summaries for the fixed PCA protocols. RMSE conditions on point return; Radius is the mean largest coordinate semiaxis over returned regions. Each metric uses its stated return denominator and coordinate chart.

Law	Method	Points	Regions	RMSE	Radius
Gaussian	Newton+A	500	500	0.092	0.399
	Recenter, one round	500	500	0.087	0.399
	PCA, two rounds	500	500	0.087	0.373
	Raw central all	500	500	0.078	0.332
t_5	Newton+A	484	459	0.184	0.781
	Recenter, one round	498	462	0.149	0.785
	PCA, two rounds	500	492	0.157	0.678
	Raw central all	500	495	0.138	0.607

Figure 1 shows how the communication path depends on the initial sample's location. If the independent initial sample is already at the coordinator, it broadcasts the anchor and collects worker scores, Hessians and score second moments in one upload stage. The communication cost is $8 \cdot 31(60 + 51 + 2 \cdot 1326)$ bytes (0.69 MB). If all n_0 initial-sample vectors must first be collected remotely, their upload precedes initial fitting and anchor broadcast. This adds a sequential upload stage and $8 \cdot 20n_0$ bytes, giving $8\{20n_0 + 31(60 + 51 + 2 \cdot 1326)\}$ bytes (1.30 MB). The initial sample remains an information cost in either deployment; the same N worker observations determine the recovered point.

Dedicated two-round PCA first uploads 210 distinct ambient second-moment entries per remote worker, then broadcasts the 60-entry fitted anchor and uploads 1326 score second-moment entries. Its two upload stages and one broadcast cost $8 \cdot 31(210 + 60 + 1326)$ bytes (0.40 MB). At the fitted subspace the pooled score is zero and the Hessian follows from the ambient matrix. Raw worker collection uses one upload stage and $8 \cdot 20 \cdot 31 \cdot 480$ bytes (2.38 MB); collecting the initial sample concurrently gives 3.00 MB for the all-sample fit. All counts use 64-bit scalars, include required

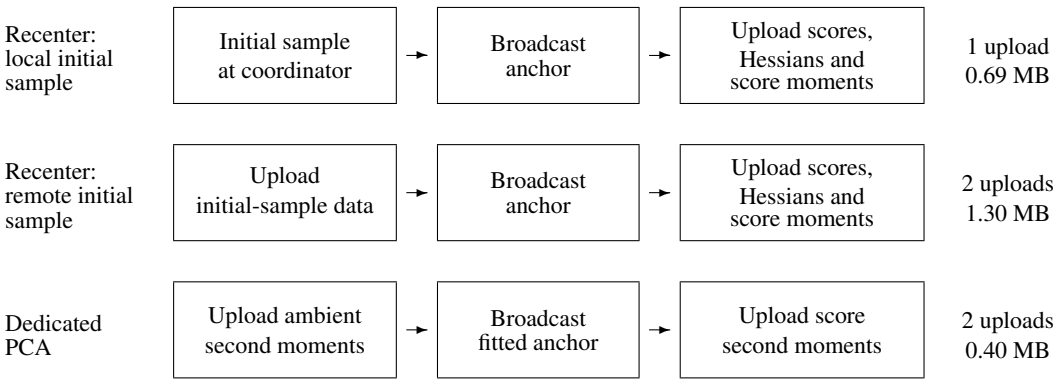


Figure 1: Communication dependencies for the PCA comparison in Table 2. Uploads go to the coordinator; broadcasts go to the workers. In the bottom two paths, the first upload supplies the data used to fit the anchor before broadcasting it. Counts give sequential upload stages; byte totals include broadcasts. Recentring retains the initial-chart covariance; dedicated PCA estimates covariance in the fitted chart.

broadcasts, and exclude transport headers. A single ambient second-moment upload recovers the exact PCA point; covariance inference uses the additional score moments.

For a communication-only comparison, charge equal fixed latency per sequential upload stage and a linear cost per transmitted byte. Both PCA paths have one broadcast. With a ready local initial sample, recentring saves one upload stage but sends 0.29 MB more; with remote initial-sample collection, it has the same two upload stages and sends 0.90 MB more. These differences use the unrounded byte counts. Local fitting time is separate from this communication-cost model.

For the optional covariance update in Section A.1, an additional worker-reporting round adds $8 \cdot 31(60 + 1326)$ bytes. Its mean communication cost is 1.45 MB for Gaussian data and 1.63 MB for t_5, exceeding the fixed two-round protocol because it already collected the initial sample and first-round Newton summaries. The t_5 covariance update adds 30 region returns but gains 26 containing regions and loses 34, changing coverage by -0.016 (paired Monte Carlo standard error 0.015). For Gaussian data it loses three containing regions and gains none, changing coverage by -0.006 (standard error 0.0035). Updating covariance thus trades coverage against reportability in these paired comparisons. Proposition F.2 establishes asymptotic coverage for the covariance choice.

D Application implementation

The six positive measurements are Area, Perimeter, MajorAxisLength, MinorAxisLength, Extent and Solidity. We take natural logarithms and retain the resulting feature scales; standardization would define a different, correlation-based target. A fixed permutation assigns 1000 grains to 500 initial-sample pairs and 12000 grains to 6000 worker pairs, leaving 611 unused. Each retained grain appears once. Under the pooled independent-sampling model, disjoint pairs supply independent mean-zero differences with the original log-feature covariance. The intervals are interpreted under this working model.

On $\mathrm{Gr}(6, 2)$, the chart dimension is eight. A worker reports eight score entries and two symmetric matrices with 36 distinct entries each. The initial estimate is the leading rank-two subspace of its 500 pair differences. The functional Jacobian for each projection leverage is computed at the Newton fit by centered finite differences with step 10^{-6}. With estimated standard error $\{\widehat{J}_N \widehat{\Sigma}_N^{\mathrm{A}} \widehat{J}_N^{\mathrm{T}}/N\}^{1/2}$, each interval in Table 3 uses the corresponding scalar normal approximation. The resulting intervals are pointwise. Projection averaging computes the leading two eigenvectors of the mean worker projection matrix; tangent averaging averages local-subspace coordinates in the common initial chart. All approximation comparisons exclude initial-sample pairs from the central worker fit. A 500-replication pairs bootstrap with the initial estimate held fixed gives projection-leverage standard errors 0.95–1.06 times the delta-method estimates.

E Further Methodological and Theoretical Details

E.1 Identification and initial-estimate construction

Proposition E.1. *Under Assumptions 2.1 to 2.3, the neighborhoods can be reduced, without changing notation, so that for some $\lambda > 0$,*

$$A_p(v) \succeq 4\lambda \mathrm{I}_d, \qquad p \in V, v \in B_{r_0}(0). \tag{E.1}$$

For each $p \in V$, $v_\star(p)$ is then the unique minimizer of F_p on $B_{r_0}(0)$. Moreover,

$$\|v_\star(p)\| \le c_R^{-1} d(p, \theta_\star). \tag{E.2}$$

Positive definiteness of the target Hessian extends to a common neighborhood of anchors and chart coordinates. This gives every worker the same locally identified target and supplies the Hessian lower bound used to control the Newton and surrogate steps. The proof is in Section G.1.

Proposition E.2. *Under Assumptions 2.1 to 2.3, fix any deterministic $p_{\mathrm{ref}} \in V$ for which $v_\star(p_{\mathrm{ref}})$ is an interior point of the local ball. Let*

$$f_0(v) = \frac{1}{n_0} \sum_{i=1}^{n_0} \phi_{p_{\mathrm{ref}}, Z_{0,i}}(v),$$

and let $\widehat{v}_0$ be its local minimizer in the admissible ball. Then

$$\|\widehat{v}_0 - v_\star(p_{\mathrm{ref}})\| = O_p(n_0^{-1/2}), \qquad p_0 = \mathcal{R}_{p_{\mathrm{ref}}}(\widehat{v}_0) \quad \Longrightarrow \quad d(p_0, \theta_\star) = O_p(n_0^{-1/2}).$$

The proposition verifies the initial-estimate rate when a fixed reference chart in the target's regular basin is available: ordinary local M-estimation in that chart supplies the anchor (van der Vaart, 1998, Chapter 5). Its proof is in Section G.4. The general initial-estimate assumption also covers model-specific constructions. For principal subspaces, the Grassmann experiments obtain the anchor directly from the leading eigenspace of the initial-sample second-moment matrix. Once the anchor is computed, workers need only the broadcast chart to form their summaries.

Remark E.3. The score second moment is required by the multivariate central limit theorem and by consistency of the score covariance. A Hessian second moment gives the $n^{-1/2}$ master-Hessian fluctuation. The Hölder remainder itself only uses an empirical average of L, so $\mathbb{E}L < \infty$ is enough. Requiring every envelope to have the same high-order moment would be convenient but unnecessarily strong.

Remark E.4. For a specified model, geometry and tail assumptions verify the derivative envelopes, while a positive-definite target Hessian establishes local identification. Section 2.5 carries out this verification for intrinsic means and principal subspaces. The reporting checks in Section 2.4 assess whether the fitted point and covariance can be used within the chosen chart. These observable checks complement the model conditions.

E.2 Smooth functions of the manifold parameter

Corollary E.5. *Let $h : \mathcal{M} \to \mathbb{R}^q$ be continuously differentiable near $\theta_\star$, and define*

$$J_\star = D(h \circ \mathcal{R}_{\theta_\star})(0), \qquad \widehat{J}_N = D(h \circ \mathcal{R}_{p_0})(\widetilde{v}_N^{\bullet}).$$

Under the rate condition for $\bullet$ in Theorem 3.3,

$$\sqrt{N}\{h(\widetilde{\theta}_N^{\bullet}) - h(\theta_\star)\} \Longrightarrow N(0, J_\star \Sigma_\star J_\star^{\mathsf{T}}), \tag{E.3}$$

and

$$\widehat{J}_N \widehat{\Sigma}_N \widehat{J}_N^{\mathsf{T}} \xrightarrow{p} J_\star \Sigma_\star J_\star^{\mathsf{T}}. \tag{E.4}$$

The Jacobian can be evaluated at the reported estimate by continuity. Section I.2 supplies the delta-method proof.

E.3 CENTRALIZED LOCAL EXPANSION

We first separate the statistical error of the centralized fit from the error of approximating it at the initial anchor. This distinction determines where initial-estimate accuracy enters the distributed procedure.

Theorem E.6 (Centralized worker-sample M-estimator)**.** *Under Assumptions 2.1 to 2.4, the following statements hold.*

(a) With probability tending to one, f_N is strongly convex on the localized chart, and $\widehat{v}_N$ in (2.8) *is its unique interior minimizer.*

(b) The centralized estimator is root-N around the target coordinate:

$$\|\widehat{v}_N - v_\star(p_0)\| = O_p(N^{-1/2}). \tag{E.5}$$

It has the asymptotic linear representation

$$\sqrt{N}\{\widehat{v}_N - v_\star(p_0)\} = -A_\star^{-1}\frac{1}{\sqrt{N}}\sum_{j=1}^{m}\sum_{i=1}^{n} s_{\theta_\star}(0, Z_{j,i}) + o_p(1). \tag{E.6}$$

(c) The same estimator admits the anchor-centered one-step expansion

$$\widehat{v}_N = -A_N^{-1} g_N + R_N^{\mathrm{C}}, \qquad \left\|R_N^{\mathrm{C}}\right\| = O_p(\delta_N^{1+\rho}). \tag{E.7}$$

Consequently,

$$\sqrt{N}\{\widehat{v}_N - v_\star(p_0)\} \Longrightarrow N(0, \Sigma_\star). \tag{E.8}$$

The root-N statement in (E.5) holds with any consistent initial estimate. The anchor changes the chart, while the centralized fit remains root-N accurate for the same geometric target. Initial-estimate accuracy enters the single Newton correction through its Taylor remainder. For Lipschitz Hessians and a root-n initial estimate, this remainder is $O_p(n^{-1})$.

The linear representation (E.6) is a local chart version of standard regular M-estimation; see van der Vaart (1998, Chapter 5) and the manifold mean literature cited above. The new role it plays here is to provide the centralized benchmark against which the communication approximation is measured.

E.4 SHARPNESS OF THE ANCHOR REMAINDER

Proposition E.7. *Fix $\rho \in (0, 1]$ and $c > 0$. On $\mathbb{R}$, consider*

$$F(x) = \frac{x^2}{2} + \frac{c}{2+\rho}\,|x|^{2+\rho}, \qquad \theta_\star = 0.$$

For a positive anchor $p \downarrow 0$, one exact population Newton step is

$$p - \frac{F'(p)}{F''(p)} = \frac{c\rho\, p^{1+\rho}}{1 + c(1+\rho)p^\rho} \sim c\rho p^{1+\rho}. \tag{E.9}$$

Thus the generic $a_{n_0}^{1+\rho}$ anchor term in (3.1) *cannot be replaced by a smaller order under only ρ-Hölder Hessian regularity.*

E.5 FRAME INVARIANCE

Proposition E.8. *Let $E_p' = E_p U_p$, where $U_p \in \mathbb{R}^{d\times d}$ is orthogonal, and write the same retraction in the new coordinates as $\mathcal{R}_p'(v') = \mathcal{R}_p(U_p v')$. Then, for each realized anchor p_0,*

$$g_N' = U_{p_0}^{\mathsf{T}} g_N, \qquad A_N' = U_{p_0}^{\mathsf{T}} A_N U_{p_0}, \qquad H_1' = U_{p_0}^{\mathsf{T}} H_1 U_{p_0},$$
$$Q_N' = U_{p_0}^{\mathsf{T}} Q_N U_{p_0}, \qquad \widetilde{v}_N^{\bullet\prime} = U_{p_0}^{\mathsf{T}} \widetilde{v}_N^{\bullet}, \qquad \widehat{\Sigma}_N' = U_{p_0}^{\mathsf{T}} \widehat{\Sigma}_N U_{p_0},$$

whenever the displayed inverses exist and the surrogate uses the corresponding unique local minimizer. The Newton and surrogate report status, manifold estimates, Wald quadratic form, and mapped confidence set are therefore unchanged. Without uniqueness, the two coordinate systems still map the sets of surrogate minimizers bijectively, but an arbitrary tie-breaking rule need not be invariant.

Coordinates are needed for communication, but the inferential object is not tied to a particular orthonormal basis. A change to a genuinely different retraction need not give the same finite-sample point estimate. Because all smooth retractions have the same first derivative at the origin, regular choices are first-order equivalent; the exact statement above is intentionally limited to frame changes.

E.6 Formal reductions and comparison with established procedures

Proposition E.9. *The common-anchor construction has the following special cases.*

(i) If $\mathcal{M} = \mathbb{R}^d$ and $\mathcal{R}_p(v) = p + v$, the surrogate objective (2.10) *agrees, up to an additive constant, with the low-dimensional communication-efficient surrogate likelihood objective of Jordan et al. (2019); the Newton estimator is the corresponding one-step estimator.*

(ii) If $m = 1$, then $g_N = g_1$, $\widetilde{f}_N = f_1$, and $\widetilde{v}_N^{\mathrm{S}} = \widehat{v}_N$ exactly.

(iii) If every worker chart loss is quadratic in v, then $H_p(v, z)$ does not depend on v, $R_N^{\mathrm{C}} = 0$, and

$$\widetilde{v}_N^{\mathrm{N}} = \widehat{v}_N$$

exactly whenever the minimizer lies in the chart. No initial-rate or machine-growth condition is then needed for Newton–centralized equivalence.

(iv) Under quadratic losses, the surrogate is also exact whenever $H_1 = A_N$. More generally,

$$\widetilde{v}_N^{\mathrm{S}} - \widehat{v}_N = (A_N^{-1} - H_1^{-1}) g_N$$

without a Taylor remainder.

The reductions distinguish nonlinear localization error from variation in worker Hessians. Quadratic losses eliminate the former, and a common Hessian eliminates the latter. They also connect the common-chart construction directly to Euclidean surrogate likelihood and explain the exact agreement expected in the quadratic simulation.

F Verification for the Two Main Manifold Examples

F.1 Intrinsic Fréchet means

Let $K \subset \mathcal{M}$ be a compact set containing the support of Z, and suppose there is a neighborhood V of $\theta_\star$ such that no point of K lies in the cut locus of any point in the chart images used over V. Then

$$(\theta, z) \longmapsto \frac{1}{2} d^2(\theta, z)$$

is smooth on the relevant compact subset of $\mathcal{M} \times \mathcal{M}$. All derivatives through order three are bounded there. Consequently, for any smooth local retraction, the score and Hessian envelopes in Assumption 2.2 can be chosen deterministic, and $\rho = \tau_s = \tau_H = 1$ is valid. If the population Fréchet Hessian at $\theta_\star$ is positive definite, Assumption 2.3 holds.

A sufficiently small strongly convex normal ball containing K supplies a transparent sufficient condition for uniqueness and local convexity; see Afsari (2011). This compact separation condition is stronger than what is needed for a centralized intrinsic-mean central limit theorem. In particular, Hotz et al. (2024) allow certain distributions whose support meets the cut locus. The stronger condition here is tied to communicating ordinary sample Hessians at a shared anchor. If those Hessians are not well defined or have nonintegrable singularities, the common-anchor Wald procedure requires modification rather than a formal appeal to the weaker centralized theorem.

F.2 Grassmann principal subspaces

Let $U_\star \in \mathbb{R}^{D \times k}$ contain the leading eigenvectors of $\Sigma_X = \mathbb{E}(XX^{\mathsf{T}})$. Work in an eigenbasis and write

$$U_\star = \begin{pmatrix} \mathrm{I}_k \\ 0 \end{pmatrix}, \qquad \Sigma_X = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix},$$

where $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\Lambda_2 = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_D)$. A standard graph chart for a nearby subspace is

$$U(K) = \begin{pmatrix} \mathrm{I}_k \\ K \end{pmatrix} (\mathrm{I}_k + K^\mathsf{T} K)^{-1/2}, \qquad K \in \mathbb{R}^{(D-k)\times k}.$$

For the population version of (2.13), a second-order expansion at $K = 0$ gives

$$\begin{aligned} -\frac{1}{2}\,\mathrm{tr}\{U(K)^\mathsf{T} \Sigma_X U(K)\} &= -\frac{1}{2}\,\mathrm{tr}(\Lambda_1) + \frac{1}{2}\,\mathrm{tr}(K^\mathsf{T} K \Lambda_1) - \frac{1}{2}\,\mathrm{tr}(K^\mathsf{T} \Lambda_2 K) + O(\|K\|^3) \\ &= -\frac{1}{2}\,\mathrm{tr}(\Lambda_1) + \frac{1}{2}\sum_{a=1}^{k}\sum_{b=1}^{D-k} (\lambda_a - \lambda_{k+b}) K_{ba}^2 + O(\|K\|^3). \end{aligned}$$

This proves (2.14). If $\lambda_k > \lambda_{k+1}$, every coefficient is positive and

$$\lambda_{\min}(A_\star) = \lambda_k - \lambda_{k+1}.$$

Write $X = (X_\parallel^\mathsf{T}, X_\perp^\mathsf{T})^\mathsf{T}$ in the same eigenbasis. Differentiating the sample loss at $K = 0$ yields the tangent score

$$s_\star(X) = -\,\mathrm{vec}(X_\perp X_\parallel^\mathsf{T}),$$

up to the fixed vectorization convention. Hence

$$Q_\star = \mathrm{Cov}\{\mathrm{vec}(X_\perp X_\parallel^\mathsf{T})\}.$$

On a fixed local chart, every first, second, and third derivative of the projection $U(K)U(K)^\mathsf{T}$ is bounded by a deterministic chart constant. The sample loss is quadratic in X, so

$$B(X) + M(X) + L(X) + C_s(X) + C_H(X) \le C\,\|X\|^2$$

for a deterministic C. Therefore $\mathbb{E}\,\|X\|^4 < \infty$ verifies all second-moment envelopes in (2.5); $\mathbb{E}L < \infty$ follows already from the second moment. This establishes the stated specialization for fixed D and k.

F.3 Exact PCA recovery and inference

The quadratic ambient loss permits an observable check of the Newton approximation. Let U_0 represent the initial subspace and let V_0 be an orthonormal complementary matrix. Set $q = D - k$ and use column-wise vectorization of the $q \times k$ graph coordinate. In the basis $[U_0, V_0]$, partition the empirical worker second moment as

$$\widehat{\Sigma}_X = \begin{pmatrix} \widehat{\Sigma}_{yy} & \widehat{\Sigma}_{yz} \\ \widehat{\Sigma}_{zy} & \widehat{\Sigma}_{zz} \end{pmatrix}.$$

Here this matrix is expressed in the initial basis; it is not the tangent covariance. Write $G_N = \mathrm{mat}_{q\times k}(g_N)$, and let $(A_N)_{ab}$ be the $q \times q$ block of A_N for graph columns a, b.

Proposition F.1. *Define a $k \times k$ matrix B and a $q \times q$ matrix C by*

$$B_{ab} = q^{-1}\,\mathrm{tr}\{(A_N)_{ab}\}, \qquad C = k^{-1}\sum_{a=1}^{k}\{B_{aa}\mathrm{I}_q - (A_N)_{aa}\}.$$

Then

$$\widehat{\Sigma}_X^\circ := \begin{pmatrix} B & -G_N^\mathsf{T} \\ -G_N & C \end{pmatrix} = \widehat{\Sigma}_X - \frac{\mathrm{tr}(\widehat{\Sigma}_{zz})}{q}\mathrm{I}_D.$$

Consequently its leading eigenspace, mapped by $[U_0, V_0]$, equals the worker-centralized PCA subspace whenever the empirical eigengap is positive.

The reconstruction is proved in Section I.5.

The same reconstruction supplies the Hessian at the recovered fit. Rotate $\widehat{\Sigma}_X^\circ$ into an orthonormal basis $[\widehat{U}_N, \widehat{V}_N]$ of that fit and its complement and apply the same block-Hessian formula. The omitted scalar identity cancels. The pooled score is zero at an exact empirical PCA fit, so workers need only return their uncentered score second moments in the extra round. Their pooled mean equals the empirical score covariance there. Numerical implementations check the final Hessian and covariance before returning a region.

Proposition F.2. *Suppose D, k are fixed, the population eigengap is positive, $\mathbb{E}\|X\|^4 < \infty$, and $Q_\star \succ 0$. Use either covariance protocol in Section 3.3, or select between their central-centered regions by an observable rule, with the stated reporting checks. For the initial-chart protocol and adaptive selection, let the independent initial estimate be consistent. The returned point equals the worker-centralized PCA point whenever it is reportable, and the complete returned confidence set satisfies*

$$\mathbb{P}\{\theta_\star \in \mathcal{C}_{1-\alpha}\} \longrightarrow 1-\alpha.$$

No root-N equivalence condition on the original Newton step is required.

Section I.5 proves the coverage and reporting assertions.

G Supporting Local Results

This appendix proves the geometric reduction, the deterministic Hölder localization inequality, the conditional uniform laws used with a random anchor, and the initial-estimate verification result.

G.1 Proof of the local-identification proposition

Proof. Let $\lambda_\star = \lambda_{\min}(A_\star) > 0$. We first show that the population chart Hessian is continuous at $(\theta_\star, 0)$ with a quantitative bound. By (2.3) and (2.4),

$$\begin{aligned}\|A_p(v) - A_\star\|_{\mathrm{op}} &\le \|A_p(v) - A_p(0)\|_{\mathrm{op}} + \|A_p(0) - A_{\theta_\star}(0)\|_{\mathrm{op}} \\ &\le (\mathbb{E}L)\,\|v\|^\rho + (\mathbb{E}C_H) d(p,\theta_\star)^{\tau_H}. \end{aligned} \tag{G.1}$$

The expectations are finite by (2.5). Choose a smaller neighborhood V and a smaller radius r_0 so that the right-hand side of (G.1) is at most $\lambda_\star/2$ for every $p \in V$ and $v \in B_{r_0}(0)$. Weyl's eigenvalue inequality then gives

$$\lambda_{\min}\{A_p(v)\} \ge \lambda_\star/2.$$

Taking $\lambda = \lambda_\star/8$ yields (E.1).

We next verify stationarity in every nearby chart. The equality $b_{\theta_\star}(0) = 0$ says that the differential of F at $\theta_\star$ is the zero covector in one nonsingular chart. Vanishing of a differential is coordinate invariant. Because $\mathcal{R}_p\{v_\star(p)\} = \theta_\star$ and $D\mathcal{R}_p\{v_\star(p)\}$ is nonsingular after the neighborhoods are reduced if necessary,

$$b_p\{v_\star(p)\} = D\mathcal{R}_p\{v_\star(p)\}^{\mathsf{T}} dF_{\theta_\star} = 0.$$

By Assumption 2.1, $v_\star(p)$ lies in the interior of the ball. The bound (E.1) implies that F_p is strongly convex on that ball. Hence its stationary point $v_\star(p)$ is its unique minimizer there.

Finally, the lower bi-Lipschitz inequality gives

$$c_R\,\|v_\star(p)\| \le d\{\mathcal{R}_p(v_\star(p)), \mathcal{R}_p(0)\} = d(\theta_\star, p),$$

which proves (E.2). □

G.2 A deterministic Hölder localization lemma

Lemma G.1. *Let f be twice continuously differentiable on an open neighborhood of $\overline{B}_r(0) \subset \mathbb{R}^d$, and put*

$$g = \nabla f(0), \qquad H_0 = \nabla^2 f(0).$$

Suppose that, for constants $\lambda > 0$, $L_f \ge 0$, and $\rho \in (0,1]$,

$$\lambda_{\min}(H_0) \ge \frac{3\lambda}{2}, \qquad \left\|\nabla^2 f(u) - \nabla^2 f(v)\right\|_{\mathrm{op}} \le L_f\,\|u-v\|^\rho. \tag{G.2}$$

Define

$$r_f = \begin{cases} \min\{r, (\lambda/(2L_f))^{1/\rho}\}, & L_f > 0, \\ r, & L_f = 0. \end{cases}$$

If $2\|g\|/\lambda \le r_f$, then f has a unique stationary minimizer u_f in $B_{r_f}(0)$ and

$$\|u_f\| \le \lambda^{-1}\|g\|, \tag{G.3}$$

$$\left\|u_f + H_0^{-1} g\right\| \le \frac{L_f}{(1+\rho)\lambda^{2+\rho}} \|g\|^{1+\rho}. \tag{G.4}$$

When $L_f = 0$, the remainder in (G.4) *is zero.*

Proof. We divide the argument into four steps.

Step 1: Hessian lower bound on the localization ball. For $u \in B_{r_f}(0)$,

$$\left\|\nabla^2 f(u) - H_0\right\|_{\mathrm{op}} \le L_f \|u\|^\rho \le L_f r_f^\rho \le \lambda/2.$$

Weyl's inequality and (G.2) imply

$$\nabla^2 f(u) \succeq \lambda \mathrm{I}_d, \qquad u \in B_{r_f}(0). \tag{G.5}$$

Thus f is λ-strongly convex on this ball.

Step 2: existence of an interior minimizer. Set $R = 2\|g\|/\lambda$. By assumption, $R \le r_f$. For every u with $\|u\| = R$, the fundamental theorem of calculus and (G.5) give

$$\begin{aligned}\langle \nabla f(u), u\rangle &= \langle g, u\rangle + \int_0^1 u^\mathsf{T} \nabla^2 f(tu) u \, dt \\ &\ge -\|g\| R + \lambda R^2 = 2\|g\|^2/\lambda.\end{aligned}$$

If $g \neq 0$, this quantity is strictly positive. Hence a minimizer of the continuous function f over the compact ball $\overline{B}_R(0)$ cannot lie on its boundary: moving a small distance inward decreases f. It is therefore an interior stationary point. If $g = 0$, the origin is already stationary. Strong convexity makes the stationary point unique in $B_{r_f}(0)$.

Step 3: norm of the minimizer. The first-order equation and another application of the fundamental theorem of calculus yield

$$0 = g + \left\{\int_0^1 \nabla^2 f(tu_f)\, dt\right\} u_f.$$

The matrix in braces has smallest eigenvalue at least λ by (G.5). Consequently,

$$\|g\| \ge \lambda \|u_f\|,$$

which is (G.3).

Step 4: one-step remainder. Write

$$0 = g + H_0 u_f + r_f^\circ, \qquad r_f^\circ = \int_0^1 \{\nabla^2 f(tu_f) - H_0\} u_f \, dt.$$

The Hölder condition gives

$$\begin{aligned}\left\|r_f^\circ\right\| &\le \int_0^1 L_f (t\|u_f\|)^\rho \|u_f\| \, dt \\ &= \frac{L_f}{1+\rho} \|u_f\|^{1+\rho}.\end{aligned}$$

Since $\left\|H_0^{-1}\right\|_{\mathrm{op}} \le \lambda^{-1}$ and (G.3) holds,

$$\begin{aligned}\left\|u_f + H_0^{-1} g\right\| &= \left\|H_0^{-1} r_f^\circ\right\| \\ &\le \frac{L_f}{(1+\rho)\lambda} \left(\frac{\|g\|}{\lambda}\right)^{1+\rho},\end{aligned}$$

which is (G.4). □

G.3 CONDITIONAL UNIFORM LAWS AT A RANDOM ANCHOR

Lemma G.2. *Let p_k be a random element of V, independent of i.i.d. observations $Z_1, \dots, Z_k$. Under Assumption 2.2,*

$$\sup_{v \in B_{r_0}(0)} \left\| \frac{1}{k} \sum_{i=1}^{k} H_{p_k}(v, Z_i) - A_{p_k}(v) \right\|_{\mathrm{op}} = o_p(1). \tag{G.6}$$

The result is uniform over the distribution of the independent anchor p_k.

Proof. Because d is fixed, it is enough to control the finitely many matrix entries. Fix a mesh size $\eta > 0$. Let $\{v_1, \dots, v_J\}$ be a finite η-net of the compact ball $\overline{B}_{r_0}(0)$.

Condition on p_k. At each net point v_j, every matrix entry of $H_{p_k}(v_j, Z_i)$ has variance bounded by $\mathbb{E}M^2$, uniformly in p_k and j. Chebyshev's inequality followed by a union bound over the finite set of entries and net points gives

$$\max_{1 \le j \le J} \left\| \frac{1}{k} \sum_{i=1}^{k} H_{p_k}(v_j, Z_i) - A_{p_k}(v_j) \right\|_{\mathrm{op}} = o_p(1) \tag{G.7}$$

for each fixed η.

For arbitrary v, choose v_j with $\|v - v_j\| \le \eta$. By (2.3),

$$\begin{aligned} &\left\| \left[\frac{1}{k} \sum_{i=1}^{k} H_{p_k}(v, Z_i) - A_{p_k}(v) \right] - \left[\frac{1}{k} \sum_{i=1}^{k} H_{p_k}(v_j, Z_i) - A_{p_k}(v_j) \right] \right\|_{\mathrm{op}} \\ &\qquad \le \left\{ \frac{1}{k} \sum_{i=1}^{k} L(Z_i) + \mathbb{E}L \right\} \eta^{\rho}. \end{aligned}$$

The expression in braces is $O_p(1)$ by the weak law of large numbers. First let $k \to \infty$ in (G.7), and then let $\eta \downarrow 0$. This proves (G.6). All probability bounds used only the deterministic envelope moments, so they are uniform in the independent anchor. □

Lemma G.3. *Under Assumptions 2.1 to 2.4, let $v_0^\star = v_\star(p_0)$. Then*

$$\|v_0^\star\| = O_p(a_{n_0}), \tag{G.8}$$

$$\|b_{p_0}(0)\| = O_p(a_{n_0}), \tag{G.9}$$

$$g_N - b_{p_0}(0) = O_p(N^{-1/2}), \tag{G.10}$$

$$A_N - A_{p_0}(0) = O_p(N^{-1/2}), \tag{G.11}$$

$$H_1 - A_{p_0}(0) = O_p(n^{-1/2}), \tag{G.12}$$

where vector and matrix O_p statements use the corresponding Euclidean or operator norm. In particular,

$$\|g_N\| = O_p(\delta_N). \tag{G.13}$$

Moreover,

$$\sup_{v \in B_{r_0}(0)} \left\| \nabla^2 f_N(v) - A_{p_0}(v) \right\|_{\mathrm{op}} = o_p(1), \tag{G.14}$$

$$\sup_{v \in B_{r_0}(0)} \left\| \nabla^2 f_1(v) - A_{p_0}(v) \right\|_{\mathrm{op}} = o_p(1), \tag{G.15}$$

$$\frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{n} L(Z_{j,i}) = O_p(1), \qquad \frac{1}{n} \sum_{i=1}^{n} L(Z_{1,i}) = O_p(1). \tag{G.16}$$

Proof. For chart-indexed calculations, first replace p_0 by the V-valued p_0° from the whole-sample convention. Because $\mathbb{P}(p_0^\circ = p_0) \to 1$, this replacement does not change any asserted O_p or o_p order.

We write p_0 below for notational economy. The first assertion follows from (E.2) and (2.6). Since $b_{p_0}(v_0^\star) = 0$, the fundamental theorem of calculus gives

$$b_{p_0}(0) = -\left\{\int_0^1 A_{p_0}(tv_0^\star)\,dt\right\} v_0^\star.$$

The Hessian envelope implies $\sup_{p,v} \|A_p(v)\|_{\mathrm{op}} \le \mathbb{E}M$, so (G.9) follows from (G.8).

Condition on the initial estimate. The N worker observations remain i.i.d. and

$$\mathbb{E}\{s_{p_0}(0, Z) \mid p_0\} = b_{p_0}(0), \qquad \mathbb{E}\{H_{p_0}(0, Z) \mid p_0\} = A_{p_0}(0).$$

The score and Hessian envelopes and their second moments give, entrywise,

$$\mathbb{E}\{\|g_N - b_{p_0}(0)\|^2 \mid p_0\} \le C/N, \qquad \mathbb{E}\{\|A_N - A_{p_0}(0)\|_{\mathrm{op}}^2 \mid p_0\} \le C/N$$

for a deterministic C. Conditional Chebyshev inequalities prove (G.10) and (G.11); the same argument with one worker proves (G.12). Combining (G.9) and (G.10) proves (G.13).

Apply Lemma G.2 with $k = N$ and $p_k = p_0$ to obtain (G.14), and with $k = n$ to the master sample to obtain (G.15). The two statements in (G.16) follow from the weak law of large numbers and $\mathbb{E}L < \infty$. □

Lemma G.4. *Under the assumptions of Lemma G.3, define*

$$\zeta_{p_0}(Z) = s_{p_0}\{v_\star(p_0), Z\}, \qquad \zeta_\star(Z) = s_{\theta_\star}(0, Z).$$

Then, conditionally on p_0,

$$\mathbb{E}\{\zeta_{p_0}(Z) \mid p_0\} = 0, \qquad \frac{1}{N}\sum_{j,i} \zeta_{p_0}(Z_{j,i}) = O_p(N^{-1/2}). \tag{G.17}$$

Furthermore,

$$\mathbb{E}\{\|\zeta_{p_0}(Z) - \zeta_\star(Z)\|^2 \mid p_0\} = o_p(1), \tag{G.18}$$

and

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{m}\sum_{i=1}^{n}\{\zeta_{p_0}(Z_{j,i}) - \zeta_\star(Z_{j,i})\} = o_p(1). \tag{G.19}$$

Proof. The first conditional expectation is zero because $v_\star(p_0)$ represents the stationary point $\theta_\star$ in the p_0 chart. By the fundamental theorem of calculus and (2.1)–(2.2),

$$\|\zeta_{p_0}(Z)\| \le B(Z) + M(Z)\,\|v_\star(p_0)\| \le B(Z) + r_0 M(Z).$$

The right-hand side has a finite second moment. Conditional Chebyshev therefore gives the second assertion in (G.17).

For the transfer bound, decompose

$$\begin{aligned}\|\zeta_{p_0}(Z) - \zeta_\star(Z)\| &\le \|s_{p_0}(v_0^\star, Z) - s_{p_0}(0, Z)\| + \|s_{p_0}(0, Z) - s_{\theta_\star}(0, Z)\| \\ &\le M(Z)\,\|v_0^\star\| + C_s(Z) d(p_0, \theta_\star)^{\tau_s}.\end{aligned}$$

After squaring and taking the conditional expectation, (G.18) follows from (G.8), (2.6), and the second moments of M and C_s.

Both ζ_{p_0} and $\zeta_\star$ have mean zero. Conditional on p_0, the second moment of the normalized sum in (G.19) equals the conditional second moment in (G.18), up to a fixed dimension constant. Conditional Chebyshev and then dominated probability convergence prove (G.19). □

G.4 Proof of the initial-rate proposition

Proof. Write $v_{\mathrm{ref}}^\star = v_\star(p_{\mathrm{ref}})$ and define the shifted initial-sample objective

$$h_0(u) = f_0(v_{\mathrm{ref}}^\star + u)$$

on a fixed ball around $u = 0$ contained in the admissible chart. Its gradient at the origin is

$$\nabla h_0(0) = \frac{1}{n_0} \sum_{i=1}^{n_0} s_{p_{\mathrm{ref}}}(v^\star_{\mathrm{ref}}, Z_{0,i}).$$

The summands have mean zero and a finite second moment by the argument in Lemma G.4. Hence

$$\|\nabla h_0(0)\| = O_p(n_0^{-1/2}). \tag{G.20}$$

By Proposition E.1, the population Hessian in this neighborhood is bounded below by $4\lambda \mathrm{I}_d$. By Lemma G.2, the initial-sample empirical Hessian converges uniformly to the population Hessian, so with probability tending to one it is bounded below by $3\lambda \mathrm{I}_d$ throughout the shifted ball. Its empirical Hölder constant is at most $n_0^{-1} \sum_i L(Z_{0,i}) = O_p(1)$.

Apply Lemma G.1 to h_0. The localization condition holds with probability tending to one because of (G.20). The resulting unique stationary minimizer $\widehat{u}_0$ satisfies

$$\|\widehat{u}_0\| = O_p(n_0^{-1/2}).$$

Since $\widehat{v}_0 = v^\star_{\mathrm{ref}} + \widehat{u}_0$,

$$\|\widehat{v}_0 - v^\star_{\mathrm{ref}}\| = O_p(n_0^{-1/2}).$$

Finally, the upper bi-Lipschitz inequality gives

$$\begin{aligned} d(p_0, \theta_\star) &= d\{\mathcal{R}_{p_{\mathrm{ref}}}(\widehat{v}_0), \mathcal{R}_{p_{\mathrm{ref}}}(v^\star_{\mathrm{ref}})\} \\ &\le C_R \, \|\widehat{v}_0 - v^\star_{\mathrm{ref}}\| = O_p(n_0^{-1/2}). \end{aligned}$$

□

H Proofs of the Centralized and Distributed Results

H.1 Proof of the centralized theorem

Proof. We prove the three assertions in a sequence that separates localization from the root-N expansion.

Step 1: high-probability Hessian lower bound. By Proposition E.1, $A_{p_0}(v) \succeq 4\lambda \mathrm{I}_d$ for $p_0 \in V$ and $v \in B_{r_0}(0)$. Assumption 2.4 implies $\mathbb{P}(p_0 \in V) \to 1$. By (G.14),

$$\sup_{v \in B_{r_0}(0)} \left\| \nabla^2 f_N(v) - A_{p_0}(v) \right\|_{\mathrm{op}} = o_p(1).$$

It follows that, with probability tending to one,

$$\nabla^2 f_N(v) \succeq 3\lambda \mathrm{I}_d, \qquad v \in B_{r_0}(0). \tag{H.1}$$

Thus f_N is strongly convex on the whole local ball.

Step 2: localization at the chart origin. By (G.13), $\|g_N\| = O_p(\delta_N) = o_p(1)$. The empirical Hölder constant of $\nabla^2 f_N$ is

$$\frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{n} L(Z_{j,i}) = O_p(1)$$

by (G.16). On an event whose probability tends to one, the hypotheses of Lemma G.1 hold for f_N on a fixed subball, with deterministic Hessian lower bound. The lemma gives an interior stationary point, and (H.1) makes it the unique minimizer on the full local ball. This proves part (a). It also gives

$$\|\widehat{v}_N\| = O_p(\|g_N\|) = O_p(\delta_N) \tag{H.2}$$

and, by (G.4),

$$\widehat{v}_N = -A_N^{-1} g_N + O_p(\|g_N\|^{1+\rho}) = -A_N^{-1} g_N + O_p(\delta_N^{1+\rho}).$$

This is part (c).

Step 3: root-N distance from the target coordinate. Put $v_0^\star = v_\star(p_0)$ and

$$G_N^\star = \nabla f_N(v_0^\star) = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n}\zeta_{p_0}(Z_{j,i}).$$

By (G.17), $G_N^\star = O_p(N^{-1/2})$. The first-order condition at the empirical minimizer is

$$\begin{aligned} 0 &= G_N^\star + \left\{\int_0^1 \nabla^2 f_N\big(v_0^\star + t(\widehat{v}_N - v_0^\star)\big)\,dt\right\}(\widehat{v}_N - v_0^\star) \\ &= G_N^\star + \bar{A}_N(\widehat{v}_N - v_0^\star). \end{aligned} \tag{H.3}$$

Here $\bar{A}_N$ denotes the integral matrix. The segment in (H.3) lies in the local ball with probability tending to one: $v_0^\star = o_p(1)$ by (G.8) and $\widehat{v}_N = o_p(1)$ by (H.2). On the Hessian event (H.1), $\lambda_{\min}(\bar{A}_N) \geq 3\lambda$. Hence

$$\|\widehat{v}_N - v_0^\star\| \leq (3\lambda)^{-1}\,\|G_N^\star\| = O_p(N^{-1/2}),$$

which proves (E.5).

Step 4: convergence of the averaged Hessian. Decompose

$$\begin{aligned} \big\|\bar{A}_N - A_\star\big\|_{\mathrm{op}} \leq &\sup_{v\in B_{r_0}(0)} \big\|\nabla^2 f_N(v) - A_{p_0}(v)\big\|_{\mathrm{op}} \\ &+ \sup_{0\leq t\leq 1} \big\|A_{p_0}\big(v_0^\star + t(\widehat{v}_N - v_0^\star)\big) - A_\star\big\|_{\mathrm{op}}. \end{aligned}$$

The first term is $o_p(1)$ by (G.14). For the second, use (G.1), $p_0 \to_p \theta_\star$, $v_0^\star = o_p(1)$, and (E.5). It is also $o_p(1)$. Therefore

$$\bar{A}_N \xrightarrow{p} A_\star, \qquad \bar{A}_N^{-1} \xrightarrow{p} A_\star^{-1}. \tag{H.4}$$

Step 5: fixed-chart asymptotic linearity. Multiplying (H.3) by $\sqrt{N}$ gives

$$\sqrt{N}(\widehat{v}_N - v_0^\star) = -\bar{A}_N^{-1}\frac{1}{\sqrt{N}}\sum_{j=1}^{m}\sum_{i=1}^{n}\zeta_{p_0}(Z_{j,i}).$$

By (H.4) and (G.19),

$$\sqrt{N}(\widehat{v}_N - v_0^\star) = -A_\star^{-1}\frac{1}{\sqrt{N}}\sum_{j=1}^{m}\sum_{i=1}^{n}\zeta_\star(Z_{j,i}) + o_p(1),$$

which is (E.6).

Finally, the summands $\zeta_\star(Z)$ are i.i.d., centered, and have covariance $Q_\star$ and a finite second moment. The multivariate central limit theorem and Slutsky's theorem yield (E.8). □

H.2 Proof of the one-round approximation theorem

Proof. Step 1: aggregated Newton estimator. By (G.11), (G.1), and $p_0 \to_p \theta_\star$,

$$A_N \xrightarrow{p} A_\star.$$

Thus A_N is invertible with probability tending to one. The anchor expansion (E.7) gives directly

$$\widetilde{v}_N^{\mathrm{N}} - \widehat{v}_N = -A_N^{-1}g_N - \widehat{v}_N = -R_N^{\mathrm{C}},$$

so the coordinate bound in (3.1) follows. On the event that both points lie in the chart, the upper bi-Lipschitz inequality gives

$$d(\widetilde{\theta}_N^{\mathrm{N}}, \widehat{\theta}_N) \leq C_R\,\big\|\widetilde{v}_N^{\mathrm{N}} - \widehat{v}_N\big\|,$$

which proves the manifold bound.

Step 2: existence and expansion of the surrogate. The surrogate gradient and Hessian at the origin are

$$\nabla\widetilde{f}_N(0) = g_1 - (g_1 - g_N) = g_N, \qquad \nabla^2\widetilde{f}_N(0) = H_1.$$

The linear correction in (2.10) does not alter higher derivatives, so the empirical Hölder constant is

$$\frac{1}{n}\sum_{i=1}^{n} L(Z_{1,i}) = O_p(1).$$

By (G.15) and (E.1), the master objective, and hence the surrogate, is strongly convex on the local ball with probability tending to one. Also $H_1 \to_p A_\star$ and $\|g_N\| = O_p(\delta_N)$. Applying Lemma G.1 to $\widetilde{f}_N$ yields a unique local minimizer and

$$\widetilde{v}_N^{\mathrm{S}} = -H_1^{-1} g_N + O_p(\delta_N^{1+\rho}),$$

which is (3.2).

Step 3: surrogate–centralized discrepancy. Subtract (E.7) from (3.2):

$$\widetilde{v}_N^{\mathrm{S}} - \widehat{v}_N = (A_N^{-1} - H_1^{-1}) g_N + O_p(\delta_N^{1+\rho}). \tag{H.5}$$

Because A_N and H_1 converge to the same positive-definite matrix, their inverse norms are bounded in probability. The inverse identity gives

$$A_N^{-1} - H_1^{-1} = A_N^{-1}(H_1 - A_N)H_1^{-1}.$$

By (G.11) and (G.12),

$$\begin{aligned}\|H_1 - A_N\|_{\mathrm{op}} &\le \|H_1 - A_{p_0}(0)\|_{\mathrm{op}} + \|A_N - A_{p_0}(0)\|_{\mathrm{op}} \\ &= O_p(n^{-1/2}) + O_p(N^{-1/2}) = O_p(n^{-1/2}).\end{aligned}$$

Together with (G.13), this yields

$$\big\|(A_N^{-1} - H_1^{-1}) g_N\big\| = O_p(n^{-1/2}\delta_N).$$

Substitution into (H.5) proves (3.3). The manifold-distance result again follows from the upper bi-Lipschitz inequality. Finally, subtracting the exact Newton definition from (3.2) proves (3.4). □

H.3 Proof of the root-N equivalence corollary

Proof. Multiply (3.1) and (3.3) by $\sqrt{N}$. The general statements follow immediately from (3.5) and (3.6).

For a root-n_0 initial estimate, substituting $\delta_N = O(n_0^{-1/2} + N^{-1/2})$ gives

$$\begin{aligned}\sqrt{N}\,\delta_N^{1+\rho} &= O\{(N/n_0^{1+\rho})^{1/2} + N^{-\rho/2}\}, \\ \sqrt{N}\,n^{-1/2}\delta_N &= O\{(m/n_0)^{1/2} + n^{-1/2}\}.\end{aligned}$$

Thus $N = o(n_0^{1+\rho})$ suffices for the Newton condition, and adding $m = o(n_0)$ suffices for the surrogate condition. If instead $a_{n_0} = O(N^{-1/2})$, the two expressions are respectively $O(N^{-\rho/2})$ and $O(n^{-1/2})$.

Now suppose $n_0 = n$ and $a_{n_0} = O(n^{-1/2})$. Since $N = mn \ge n$,

$$\delta_N = a_{n_0} + N^{-1/2} = O(n^{-1/2}).$$

Therefore

$$\sqrt{N}\,\delta_N^{1+\rho} = O\{(mn)^{1/2} n^{-(1+\rho)/2}\} = O\{m^{1/2} n^{-\rho/2}\},$$

which converges to zero under $m = o(n^\rho)$. The additional surrogate term satisfies

$$\sqrt{N}\,n^{-1/2}\delta_N = O\{(m/n)^{1/2}\}.$$

Because $0 < \rho \le 1$, $m = o(n^\rho)$ implies $m = o(n)$, so the additional term also vanishes. □

H.4 Proof of the sharpness proposition

Proof. For $p > 0$,

$$F'(p) = p + cp^{1+\rho}, \qquad F''(p) = 1 + c(1+\rho)p^{\rho}.$$

Consequently,

$$\begin{aligned} p - \frac{F'(p)}{F''(p)} &= \frac{p\{1 + c(1+\rho)p^{\rho}\} - p - cp^{1+\rho}}{1 + c(1+\rho)p^{\rho}} \\ &= \frac{c\rho p^{1+\rho}}{1 + c(1+\rho)p^{\rho}}. \end{aligned}$$

The denominator converges to one, proving (E.9). The second derivative is ρ-Hölder near zero, so this example belongs to the stated regularity class. □

I Proofs for Inference, Invariance, and Reductions

I.1 Proof of the inference theorem

Proof. Part (a): asymptotic linearity. For either estimator,

$$\sqrt{N}\{\widetilde{v}_N^{\bullet} - v_\star(p_0)\} = \sqrt{N}\{\widehat{v}_N - v_\star(p_0)\} + \sqrt{N}(\widetilde{v}_N^{\bullet} - \widehat{v}_N).$$

The first term has representation (E.6). By Theorem 3.1, the norm of the second term is $O_p(\sqrt{N} r_N^{\bullet}) = o_p(1)$. This proves (3.7); the multivariate central limit theorem gives (3.8).

Part (b), aggregate and master Hessians. By (G.11),

$$A_N - A_\star = \{A_N - A_{p_0}(0)\} + \{A_{p_0}(0) - A_\star\}.$$

The first term is $O_p(N^{-1/2})$. The second is bounded by

$$(\mathbb{E}C_H) d(p_0, \theta_\star)^{\tau_H} = o_p(1)$$

using (2.4) and consistency of the initial estimate. Thus $A_N \to_p A_\star$. The same decomposition, with (G.12), gives $H_1 \to_p A_\star$.

Part (b), score covariance. Write $s_p(Z) = s_p(0, Z)$ and

$$\Gamma(p) = \mathbb{E}\{s_p(Z) s_p(Z)^{\mathsf{T}}\}.$$

We first show $S_N - \Gamma(p_0) = o_p(1)$. Condition on p_0. Each matrix entry is an average of i.i.d. variables dominated in absolute value by $B^2(Z)$. To avoid imposing a fourth score moment, fix $K > 0$ and split each entry into the event $\{B^2 \le K\}$ and its complement. The bounded part has conditional variance at most K^2/N, uniformly in p_0, and hence converges by conditional Chebyshev. The expectation and empirical average of the tail are uniformly small as $K \to \infty$ because $\mathbb{E}B^2 < \infty$; Markov's inequality controls the empirical tail. Taking $N \to \infty$ and then $K \to \infty$ yields

$$S_N - \Gamma(p_0) = o_p(1).$$

Next, by the identity $aa^{\mathsf{T}} - bb^{\mathsf{T}} = (a-b)a^{\mathsf{T}} + b(a-b)^{\mathsf{T}}$,

$$\begin{aligned} \|\Gamma(p_0) - Q_\star\|_{\mathrm{op}} &\le \mathbb{E}\left[\|s_{p_0}(Z) - s_{\theta_\star}(Z)\| \left\{\|s_{p_0}(Z)\| + \|s_{\theta_\star}(Z)\|\right\} \mid p_0\right] \\ &\le 2\{\mathbb{E}C_s^2\}^{1/2}\{\mathbb{E}B^2\}^{1/2} d(p_0, \theta_\star)^{\tau_s} = o_p(1). \end{aligned}$$

Here $Q_\star = \Gamma(\theta_\star)$ because the target score is centered. Therefore $S_N \to_p Q_\star$. Also $g_N = o_p(1)$ by (G.13), so

$$Q_N = S_N - g_N g_N^{\mathsf{T}} \xrightarrow{p} Q_\star.$$

This proves (3.9). Since $A_\star$ is nonsingular, the continuous mapping theorem gives (3.10).

Part (c): Wald limit and mapped coverage. Under $Q_\star \succ 0$, $\Sigma_\star$ is positive definite. Combine (3.8) with either covariance convergence in (3.10). Slutsky's theorem yields (3.11).

The algorithmic no-report event has probability tending to zero. Indeed, the required Hessian matrix is positive definite with probability tending to one by (3.9); $\widetilde{v}_N^{\bullet} = o_p(1)$ by (3.8) and consistency of the initial estimate; and the largest Euclidean radius of the untruncated Wald ellipsoid is $O_p(N^{-1/2})$ because the extreme eigenvalues of $\widehat{\Sigma}_N$ are bounded away from zero and infinity with probability tending to one. Their sum is $o_p(1)$, so (2.12) holds with probability tending to one. Thus the implemented sufficient check has the same asymptotic reporting probability as full ellipsoid containment.

With probability tending to one, $p_0 \in V$ and $v_\star(p_0) \in B_{r_0/4}(0)$. On the intersection with the reporting event,

$$\theta_\star \in \mathcal{C}_{1-\alpha}^{\bullet}$$

if and only if the Wald quadratic form at $v_\star(p_0)$ does not exceed $\chi^2_{d,1-\alpha}$. The limit in (3.11) therefore gives (3.12). □

I.2 Proof of the smooth-functional corollary

Proof. Let $G_p(v) = h\{\mathcal{R}_p(v)\}$. By (3.8),

$$\|\widetilde{v}_N^{\bullet} - v_\star(p_0)\| = O_p(N^{-1/2}).$$

A mean-value expansion around the correct random target coordinate gives

$$G_{p_0}(\widetilde{v}_N^{\bullet}) - G_{p_0}\{v_\star(p_0)\} = DG_{p_0}\{v_\star(p_0)\}(\widetilde{v}_N^{\bullet} - v_\star(p_0)) + o_p(N^{-1/2}). \tag{I.1}$$

The remainder is uniform on a sufficiently small compact neighborhood because $(p, v) \mapsto DG_p(v)$ is continuous. Since $p_0 \to_p \theta_\star$ and $v_\star(p_0) \to_p 0$,

$$DG_{p_0}\{v_\star(p_0)\} \xrightarrow{p} DG_{\theta_\star}(0) = J_\star.$$

Also $G_{p_0}\{v_\star(p_0)\} = h(\theta_\star)$. Combining (I.1) with (3.8) proves (E.3).

The estimator itself satisfies $\widetilde{v}_N^{\bullet} \to_p 0$, so continuity gives

$$\widehat{J}_N = DG_{p_0}(\widetilde{v}_N^{\bullet}) \xrightarrow{p} J_\star.$$

Together with (3.10), this proves (E.4). □

I.3 Proof of the frame-invariance proposition

Proof. Fix the realized anchor and abbreviate $U = U_{p_0}$. Since $\mathcal{R}'_{p_0}(v') = \mathcal{R}_{p_0}(Uv')$, the chain rule gives

$$s'_{p_0}(v', z) = U^{\mathsf{T}} s_{p_0}(Uv', z), \qquad H'_{p_0}(v', z) = U^{\mathsf{T}} H_{p_0}(Uv', z)U.$$

At $v' = 0$, aggregation proves the displayed transformations for g_N, A_N, H_1, and Q_N.

For the Newton estimator,

$$\widetilde{v}_N^{\mathrm{N}\prime} = -(U^{\mathsf{T}} A_N U)^{-1} U^{\mathsf{T}} g_N = -U^{\mathsf{T}} A_N^{-1} g_N = U^{\mathsf{T}} \widetilde{v}_N^{\mathrm{N}}.$$

For the surrogate, write $v = Uv'$. Its transformed objective is

$$\begin{aligned} f'_1(v') - \langle g'_1 - g'_N, v' \rangle &= f_1(Uv') - \left\langle U^{\mathsf{T}}(g_1 - g_N), v' \right\rangle \\ &= f_1(v) - \langle g_1 - g_N, v \rangle . \end{aligned}$$

Hence minimizers correspond by $v' = U^{\mathsf{T}} v$. The covariance transformation follows by direct substitution into (2.9) or (2.11).

Finally,

$$(v' - \widetilde{v}')^{\mathsf{T}} (\widehat{\Sigma}')^{-1} (v' - \widetilde{v}') = (v - \widetilde{v})^{\mathsf{T}} \widehat{\Sigma}^{-1} (v - \widetilde{v}),$$

and $\mathcal{R}'_{p_0}(v') = \mathcal{R}_{p_0}(v)$. Orthogonal changes preserve the coordinate norm and covariance eigenvalues, so they also preserve (2.12). Thus the reporting decision, mapped estimate, and confidence set are identical. □

I.4 Proof of the reduction proposition

Proof. In Euclidean space, $\phi_{p,z}(v) = \ell(p+v; z)$ and the correction in (2.10) is exactly the difference between the master and aggregate gradients at the common initial estimator p. This is the low-dimensional communication-efficient surrogate likelihood objective after the change of variable $\theta = p + v$, up to an additive constant.

If $m = 1$, then $g_N = g_1$, so the linear correction vanishes and $\widetilde{f}_N = f_1 = f_N$. Uniqueness of the local minimizer gives $\widetilde{v}_N^{\mathrm{S}} = \widehat{v}_N$.

If every chart loss is quadratic, $\nabla^2 f_N(v) = A_N$ for all v. The first-order equation is exactly $g_N + A_N v = 0$, so $\widehat{v}_N = -A_N^{-1} g_N = \widetilde{v}_N^{\mathrm{N}}$. The surrogate first-order equation is $g_N + H_1 v = 0$, giving $\widetilde{v}_N^{\mathrm{S}} = -H_1^{-1} g_N$. The remaining claims follow by subtraction. □

I.5 Proofs for exact PCA recovery and covariance choice

Exact recovery.

Proof. The first two graph derivatives of the empirical loss give $G_N = -\widehat{\Sigma}_{zy}$ and $(A_N)_{ab} = (\widehat{\Sigma}_{yy})_{ab}\mathrm{I}_q - \mathbf{1}\{a = b\}\widehat{\Sigma}_{zz}$. Taking block traces yields $B = \widehat{\Sigma}_{yy} - c\mathrm{I}_k$ with $c = \mathrm{tr}(\widehat{\Sigma}_{zz})/q$. Substitution in the definition of C yields $C = \widehat{\Sigma}_{zz} - c\mathrm{I}_q$. Adding a scalar identity leaves all eigenspaces unchanged. □

Covariance and coverage.

Proof. The population eigengap and empirical second-moment consistency imply consistency of the leading worker eigenspace and a positive empirical eigengap with probability tending to one. The graph loss has the integrable fourth-moment envelopes established above. Thus the centralized expansion in Theorem E.6 holds for any consistent independent initial estimate. Also, Theorem 3.3(b) gives $\widehat{\Sigma}_N^{\mathrm{A}} \to_p \Sigma_\star$ in a smooth local frame, irrespective of the Newton approximation rate.

For the fixed fitted-chart protocol, the centralized expansion may be established in a deterministic reference chart near the target; no initial sample is used by that protocol. The following comparison handles the initial-chart and adaptive choices together.

Let $\widehat{\Sigma}_N^{\mathrm{C}}$ denote the covariance computed at the recovered fit in its own graph chart. In smoothly aligned frames, the difference between score outer products at this fitted anchor and at the target is bounded by a fixed constant times $\|X\|^4\, d(\widehat{\theta}_N, \theta_\star)$ on a fixed compact neighborhood. Averaging this bound gives $o_p(1)$ by consistency and the fourth-moment law of large numbers, even though the fitted anchor depends on the workers. The analogous Hessian bound uses $\|X\|^2$. The fixed-target empirical moments converge to $Q_\star$ and $A_\star$, and the pooled score at the fitted anchor is zero. Hence $\widehat{\Sigma}_N^{\mathrm{C}}$ is consistent, and both covariance estimates have eigenvalues bounded away from zero and infinity with probability tending to one.

For the comparison of coordinates, use smoothly aligned local frames and define

$$\psi_N(u) = \mathcal{R}_{p_0}^{-1}\{\mathcal{R}_{\widehat{\theta}_N}(u)\}, \qquad T_N = D\psi_N(0), \qquad u_\star = \mathcal{R}_{\widehat{\theta}_N}^{-1}(\theta_\star).$$

These are proof coordinates; implementation can use any orthonormal representatives by frame invariance. Since $\psi_N(0) = \widehat{v}_N$ and $u_\star = O_p(N^{-1/2})$, smoothness gives

$$v_\star(p_0) - \widehat{v}_N = T_N u_\star + O_p(N^{-1}), \qquad \widehat{\Sigma}_N^{\mathrm{A}} - T_N \widehat{\Sigma}_N^{\mathrm{C}} T_N^{\mathsf{T}} = o_p(1).$$

Here $T_N \widehat{\Sigma}_N^{\mathrm{C}} T_N^{\mathsf{T}}$ expresses the fitted-chart covariance in initial-chart coordinates. The second relation follows from covariance consistency and $T_N \to_p \mathrm{I}_d$ in the aligned frames. The two central-centered Wald statistics consequently differ by $o_p(1)$, and each converges to χ_d^2. Its continuous distribution implies that the two coverage events differ with probability $o(1)$ at the common critical value. Selecting either region by (A.1), or by any other observable rule, therefore changes its coverage by $o(1)$.

Finally, the initial and central points converge to the target, their covariance estimates remain bounded, and both ellipsoid radii shrink as $N^{-1/2}$. The required graph inverses, positivity and containment checks succeed with probability tending to one. Empty regions caused by final reporting failure contribute $o(1)$ to noncoverage. Removing these events proves the assertion for the complete procedure, without conditioning on the selected round. □