TRANSPORT-AWARE INFERENCE ON VECTOR BUNDLES: WHEN IS PARALLEL TRANSPORT STATISTICALLY BENEFICIAL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Many learning problems attach a vector to every point of a curved space: a velocity at each point of a sphere, a feature at each vertex of a mesh. The vectors at each location live in their own vector space, the *fibre*, and together the fibres form a *vector bundle*. Vectors in different fibres cannot be averaged directly; one must first decide how to carry a vector from one location to another. Geometric machine learning usually uses *parallel transport*, on the grounds that it is the geometrically canonical choice. We ask when it is also the statistically better one. We show that the error of replacing transport by a fixed reference rule depends on the rule; two standard rules on the sphere have biases of opposite sign, so this error is not an intrinsic curvature effect. Curvature enters invariantly through holonomy, the rotation of a vector carried around a closed loop. An exact mean-squared-error identity shows that at low curvature a biased reference that contracts noise beats transport, and that for varying fields the answer depends on curvature relative to the field's own variation. To make the comparison operational for incomplete data we develop Bundle-GEM, a generalized EM procedure whose parameters live in moving fibres. In controlled experiments the predicted crossover matches the exact theory, a plug-in estimate of the identity picks the better estimator in 94% of cases, full Bundle-GEM on data with up to 80% missing fibre values comes within 0.006 of the true-parameter error where fixed references and global frames fail, and a surface without a global frame shows the predicted breakdown of coordinate baselines. On real global wind fields (NCEP/NCAR reanalysis), a gauge-equivariant network beats all non-transport baselines on every test field, and the plug-in identity picks the better local estimator at 98% of locations.

1 INTRODUCTION

The usual answer in geometric machine learning to "how should I combine features attached to different points of a curved space?" is parallel transport, typically defended as the geometrically correct choice. We ask instead when inserting transport into an estimator reduces its risk, and find that being geometrically canonical does not make it statistically optimal. The advantage of transport is a bias reduction that must outweigh a variance cost, and the curvature, the departure of the latent field from the radially parallel model, the aggregation radius, the sample size and the accuracy of the transport paths determine both terms. We make each of these dependencies explicit and test them.

Formally, vector spaces $\mathcal{E}_b$ attached to the points b of a base manifold $\mathcal{M}$ form a *vector bundle* $\pi : \mathcal{E} \to \mathcal{M}$, and a data point $z = (b, v)$ pairs a base location with a fibre value $v \in \mathcal{E}_b$. To combine a feature at b with one at b_0, it has to be carried to $\mathcal{E}_{b_0}$. Parallel transport $P_{b_0 \leftarrow b}$ does this along a path using a connection. A coordinate implementation instead fixes a reference identification $\tau_{b_0 \leftarrow b} : \mathcal{E}_b \to \mathcal{E}_{b_0}$, a smooth family of linear maps with $\tau_{b_0 \leftarrow b_0} = \mathrm{Id}$ that need not be isometric. EM itself is coordinate-free; the question is whether the sufficient statistics a model uses agree with the model connection. The relevant open-path error is $\delta_{\mathrm{ref}}(b, v) := \tau_{b_0 \leftarrow b} v - P^{\nabla}_{b_0 \leftarrow b} v$ (Figure 1), which is zero when τ and ∇ encode the same identification.

Two discrepancies need to be kept apart. Comparing two paths around a loop gives the holonomy, a gauge-covariant quantity whose leading term is F^{∇} integrated over the enclosed oriented area

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.

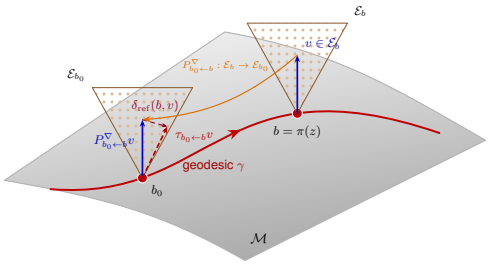

Figure 1: The two discrepancies. A vector $v \in \mathcal{E}_b$ is transported to $\mathcal{E}_{b_0}$ by $P^{\nabla}_{b_0 \leftarrow b}$; the dashed vector is $\tau_{b_0 \leftarrow b} v$ from a reference identification, and $\delta_{\mathrm{ref}}(b, v)$ is their difference, which depends on the reference. The open triangle is not a holonomy loop; curvature is measured separately by closed-loop holonomy (Thm. 5).

(Ambrose & Singer, 1953; Kobayashi & Nomizu, 1963); its conjugacy class is gauge invariant. Comparing transport with a single reference identification gives a reference-dependent error instead. Mixing the two suggests, wrongly, that curvature contributes a second-order term $F^{\nabla}_{jk} x^j x^k$ along a radial path (it vanishes by antisymmetry), and that the resulting single-path bias is a canonical curvature effect. The precise statement is that the leading single-path bias depends on the reference and is not a gauge-invariant curvature quantity; curvature enters invariantly through holonomy, and it influences the reference bias only through the identification that is chosen. Appendix B discusses a further example, image-subspace bundles (Basri & Jacobs, 2003; Hamm & Lee, 2008).

Motivation. Before we can combine vectors attached to different points of a curved space, we must decide how to carry one to the other: by parallel transport, or by a fixed reference such as a coordinate frame or ambient projection. We face this decision in message passing on spherical and mesh networks (Cohen & Welling, 2016; Cohen et al., 2019; de Haan et al., 2021; Fuchs et al., 2020; Satorras et al., 2021), in Riemannian field models (Mathieu & Nickel, 2020; De Bortoli et al., 2022), and when we smooth or impute wind and diffusion-tensor fields (Pennec et al., 2006; Arsigny et al., 2006). Each option has a price. Transport asks for a connection, paths and extra computation, and on meshes it inherits path error; fixed references are cheaper but biased, and on the sphere every global frame breaks down somewhere. Transport is usually chosen for its geometric correctness, not its risk, so practitioners have had no principled way to decide for a given dataset. We show that the data themselves can settle it: on real winds, transport was the better rule at only 67% of locations, yet a plug-in estimate of our exact risk identity picked the better rule at 98%.

Contributions. **(1) When transport reduces risk.** We separate the reference-dependent single-path error (Prop. 7), which has opposite signs for two standard references on TS^2 (Prop. 6), from the gauge-invariant loop holonomy (Thm. 5), and give an exact MSE identity that keeps the transformed-noise term (Thm. 4). Because ambient projection contracts noise, there is a threshold $\kappa^\star$ below which the biased reference has lower risk (Cor. 1). **(2) Diagnostics** (Section 5): an optimal aggregation radius (Thm. 6), the transport-relevance ratio Ξ (Cor. 2), curvature as an obstruction to parallel fields (Prop. 8), path and connection error bounds, and the non-existence of global reference frames on nontrivial bundles (Prop. 1). **(3) Bundle-GEM** (Section 3), the likelihood machinery that makes the comparison operational for incomplete data. Its ascent and consistency arguments are classical (Thms. 1–3); what is new is that the sufficient statistics must be transported into common fibres, that the parameter space is itself a bundle, so anchors and fibre parameters must move together along a transported retraction (Lem. 1), and that the base normalizer depends on the anchor on a bounded region. **(4) Experiments** (Section 6), with uncertainty estimates: the crossover, full Bundle-GEM on incomplete data, validation of the diagnostics, a surface without a global frame, the torus mechanism, and real winds with a gauge-equivariant network.

Organization. Every result is stated in the main text with its key formula. Formal statements and assumptions are in Appendices K, L and H (Table 11), proofs in Appendices N and O, protocols in Appendix T, full experimental tables in Appendices F, Q, R and V, and related work in Appendix J.

2 SETUP AND BUNDLE-VALUED LIKELIHOOD

Geometry. Let $\pi : \mathcal{E} \to \mathcal{M}$ be a smooth rank-r vector bundle over a d-dimensional Riemannian base $(\mathcal{M}, g_{\mathcal{M}})$, with fibre metric $g^{\mathcal{E}}$ and a metric-compatible connection ∇. The connection defines parallel transport $P_{b\leftarrow b'} : \mathcal{E}_{b'} \to \mathcal{E}_b$ along minimizing geodesics and a curvature form $F^{\nabla} \in \Omega^2(\mathcal{M}; \operatorname{End}\mathcal{E})$ (Kobayashi & Nomizu, 1963). General statements below are made for F^{∇}. The scalar κ used from Section 4 onward is the Gaussian curvature of the constant-curvature surfaces in our worked examples, where $F^{\nabla}_{jk} = \kappa\,\epsilon_{jk} J$ with J the oriented quarter-turn; only fully worked constants such as $\kappa^\star$ depend on constant curvature. We write $\operatorname{Exp}_b, \operatorname{Log}_b, d_{\mathcal{M}}$ for the base exponential, logarithm and distance, and equip $\mathcal{E}$ with $d\mu_{\mathcal{E}}(b, v) = d\mathrm{vol}_{\mathcal{M}}(b)\, d\nu_b(v)$. A topologically trivial bundle can still carry a non-flat connection. The effects below are driven by the connection; topology only decides whether a global reference identification exists (Proposition 1). EM iterates $Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z|Y,\theta^{(t)}}[\log p_\theta(Z, Y)]$ and ascends because $D_{\mathrm{KL}} \geq 0$ (Dempster et al., 1977; Wu, 1983). This argument is coordinate-free; geometry only determines whether the E- and M-step quantities can be computed intrinsically on $\mathcal{E}$.

A latent bundle variable $Z_i = (b_i, v_i) \sim f_\theta$ is observed through a Markov kernel $\mathcal{K}_\psi(Y \mid z)$, which gives $L(\theta, \psi) = \prod_i p_{\theta,\psi}(Y_i)$ with $p_{\theta,\psi}(Y) = \int_{\mathcal{E}} \mathcal{K}_\psi(Y \mid z) f_\theta(z)\, d\mu_{\mathcal{E}}(z)$ (Definition 1, Appendix G). We use two kernels: normalized coarsening with factor $\mathbf{1}\{z \in Y\}$ (Matheron, 1975; Molchanov, 2005), and a smooth kernel whose fibre residual is measured after transport, so that the kernel itself depends on ∇. Our running examples are TS^2 with the Levi-Civita connection (topologically nontrivial) and the tangent bundle of a torus with a discrete Levi-Civita connection (trivial, with curvature of both signs). A secondary SPD-valued field on a flat grid has manifold-valued fibres and therefore falls outside the vector-bundle theorems (Appendix G).

A transported Gaussian mixture. As a concrete model, consider the K-component transported Gaussian mixture

$$f_\theta(b, v) = \sum_{k=1}^{K} \pi_k\, c_k(m_k, \beta_k)^{-1} \exp\left[-\frac{d_{\mathcal{M}}(b, m_k)^2}{2\beta_k^2}\right] \varphi(P_{m_k\leftarrow b} v; \mu_k, \Sigma_k)\,, \tag{1}$$

where $m_k \in \mathcal{M}$ is an anchor, $\beta_k > 0$ a base bandwidth, $\mu_k \in \mathcal{E}_{m_k}$, Σ_k is positive-definite and self-adjoint on $\mathcal{E}_{m_k}$, and φ is the Gaussian density with respect to the Lebesgue measure of the fibre metric. We work on a fixed, geodesically convex observation region, so transport paths are unique and differentiation involves no moving boundary. The normalizer c_k does not depend on the anchor on a full homogeneous base, but in general it does on a fixed region; dropping its derivative breaks Theorem 1, and Appendix I.3 gives exact and approximate ways to compute it. We constrain $\pi_k \geq \pi_{\min} > 0$ and $\lambda_{\min} I \preceq \Sigma_k \preceq \lambda_{\max} I$. With latent component C_i and posterior q_{ik}, the E-step gives the transported sufficient statistics $N_k = \sum_i \int q_{ik}$, $s_k = \sum_i \int q_{ik}\, w$ and $T_k = \sum_i \int q_{ik}\, ww^\top$ with $w = P_{m_k\leftarrow b} v$, evaluated at the current anchor, and for fixed m_k the unconstrained M-step is closed form: $\mu_k = s_k/N_k$, $\Sigma_k = T_k/N_k - \mu_k\mu_k^\top$, $\pi_k = N_k/n$, while (m_k, β_k) are updated by retraction ascent on the same Q (Appendix M). The M-step is therefore computed directly from the transported moments. The robust inlier/outlier model of the sphere experiment uses the same latent-label construction with a separately specified outlier law, and active constraints use the constrained updates of Appendix M.

3 THE BUNDLE-GEM ALGORITHM AND ITS GUARANTEES

Given iterates $(\theta^{(t)}, \psi^{(t)})$, the posterior over latent bundle points and the minorizer are

$$q_i^{(t)}(z) = \frac{\mathcal{K}_{\psi^{(t)}}(Y_i \mid z)\, f_{\theta^{(t)}}(z)}{\int_{\mathcal{E}} \mathcal{K}_{\psi^{(t)}}(Y_i \mid u)\, f_{\theta^{(t)}}(u)\, d\mu_{\mathcal{E}}(u)}, \quad Q = \sum_{i=1}^{n} \int_{\mathcal{E}} q_i^{(t)}(z)\, \log\{f_\theta(z)\mathcal{K}_\psi(Y_i \mid z)\}\, d\mu_{\mathcal{E}}(z). \tag{2}$$

The posterior is coordinate-free; geometry enters because its sufficient statistics lie in different fibres. For component k of equation 1, Bundle-GEM evaluates equation 20 at the current anchor $m_k^{(t)}$. The E-step also yields a base Fréchet mean $\bar{b}_k$ and a transported fibre mean $\bar{v}_k$ (Appendix M), which form the base M-step when the base density is decoupled. With a positive-definite preconditioner G_ϑ and a first-order retraction Retr (Absil et al., 2008; Boumal, 2023), the generalized M-step is $\vartheta^{(t+1)} = \mathrm{Retr}_{\vartheta^{(t)}}(\eta_t G^{-1}_{\vartheta^{(t)}} \operatorname{grad}_\vartheta Q(\vartheta \mid \vartheta^{(t)})|_{\vartheta=\vartheta^{(t)}})$.

The parameter space is a bundle. In equation 1 the fibre parameters $\mu_k \in \mathcal{E}_{m_k}$ and $\Sigma_k \in \mathrm{Sym}^{++}(\mathcal{E}_{m_k})$ are not free Euclidean variables. The parameter space $\Theta = \{(\pi_k, m_k, \beta_k, \mu_k, \Sigma_k)\}_{k \le K}$ is therefore the total space of a bundle over $\mathcal{M}^K$ rather than a product, and when an anchor moves the fibre parameters have to move with it. We use the transported retraction $m_k^{(t+1)} = \mathrm{Exp}_{m_k^{(t)}}(\eta_t \xi_k)$, $\mu_k^{(t+1)} = P_k(\mu_k^{(t)} + \eta_t \zeta_k)$ and $\Sigma_k^{(t+1)} = P_k \widetilde{\Sigma}_k P_k^{-1}$, with $P_k := P_{m_k^{(t+1)} \leftarrow m_k^{(t)}}$ and $\widetilde{\Sigma}_k$ the in-fibre update (eq. equation 12, Appendix K). Lemma 1 shows that this update keeps every parameter in its fibre and is a first-order retraction on Θ. It does not preserve Q when the anchor moves, so ascent comes from a line search on the full Q.

Bundle-EM and Bundle-GEM. We write *Bundle-EM* for an exact M-step ($\theta^{(t+1)} = \arg\max_\theta Q$) and *Bundle-GEM* for the retraction or natural-gradient implementation above. The guarantees below only require the sufficient-ascent condition

$$Q(\vartheta_{t+1} \mid \vartheta_t) - Q(\vartheta_t \mid \vartheta_t) \ge c \left\| G^{-1}_{\vartheta^{(t)}} \operatorname{grad}_\vartheta Q \right\|^2, \qquad c > 0, \tag{3}$$

with $\vartheta_t = (\theta^{(t)}, \psi^{(t)})$, which an Armijo line search ensures under Assumption 1. This is the standard GEM requirement (Wu, 1983; McLachlan & Krishnan, 2008) stated for retractions. It holds for any positive-definite G_ϑ and any first-order retraction, so these choices affect conditioning, not correctness (Appendix I.2). Closed-form block updates followed by Armijo steps form a valid generalized ECM cycle provided the whole cycle satisfies equation 3. Algorithm 1 (Appendix C) states the loop; one iteration costs $O(nKPr^2)$ more than the corresponding product-space iteration (Appendix P), and the deterministic results below assume an exact E-step.

Guarantees. Under Assumption 1 (Appendix D: posterior base support in a geodesic ball small enough for a unique Fréchet mean, a compact likelihood superlevel set, Lipschitz gradients, and Armijo steps with a bounded preconditioner), Bundle-GEM inherits the classical EM guarantees in intrinsic form: monotone ascent of ℓ_{obs}, stationary limit points for interior iterates, identifiability of the fixed-anchor mixture up to label permutations, consistency of the constrained global maximizer with ψ fixed, and a local linear rate $\rho(\mathcal{I}_{\mathrm{mis}}\mathcal{I}_{\mathrm{com}}^{-1})$, the spectral radius of the missing-information fraction, with no separate curvature term. Identifiability with moving anchors, and of the observed model, remains an assumption, because anchors, transport and fibre law all act on the same observable. These are Theorems 1–3 and Propositions 4–5 (Appendix K; proofs in Appendix N). The bundle-specific parts are the transported sufficient statistics, the transported retraction and the anchor-dependent normalizer; the inequalities themselves are classical.

4 Reference-dependent aggregation error and invariant holonomy

This section derives the single-path error of a reference identification, shows that it depends on the reference, identifies the loop holonomy as the gauge-invariant curvature effect, and gives an exact MSE identity with a regime in which the reference estimator wins. Let ∇^0 be the flat reference connection induced by $\tau_{b_0 \leftarrow b}$, set $a = \nabla - \nabla^0$, and assume first-order alignment $a(b_0) = 0$; without it the relative map has a first-order term. For $b = \mathrm{Exp}_{b_0}(x)$ the single-path discrepancy is

$$\Phi(x) \;=\; \tau_{b_0 \leftarrow b}\, P^{\nabla}_{b \leftarrow b_0} \;=\; I + A_2(x) + O(\|x\|^3), \qquad A_2(x) = -\tfrac{1}{2}\partial_{(k}a_{j)}(b_0)x^j x^k. \tag{4}$$

The coefficient is the symmetric part of ∂a and depends on ∇^0, so it is not a gauge-invariant curvature contraction; the antisymmetric curvature term cancels identically, $F^{\nabla}_{jk}x^j x^k \equiv 0$. Curvature appears only when two paths enclose an area. Lemma 2 (Appendix L) states equation 4 for a general smooth family τ, including references that do not come from a flat connection, and Table 6 (Appendix E) lists the identifications we consider.

The invariant object is the loop holonomy. By Theorem 5 (Appendix N), for a small geodesic loop at b_0 bounding oriented area A^{jk}, $\mathrm{Hol}(\partial S) = \mathrm{Id} - \frac{1}{2}F^{\nabla}_{jk}(b_0)A^{jk} + O(|A|^{3/2})$, which on TS^2 is a rotation by the spherical excess $\kappa|A|$. Under a gauge change the holonomy is conjugated, so its trace, spectrum and rotation angle are gauge invariant. Our loop experiment recovers rotation $= \kappa \times$area to three decimals.

Proposition 7 (Appendix L) turns equation 4 into a statement about estimators. Take n observations $y_i = P_{b_i \leftarrow b_0} v^\star + \varepsilon_i$ at $b_i = \mathrm{Exp}_{b_0}(x_i)$, with $\mathbb{E}x_i = 0$, $\mathrm{Cov}x_i = \Sigma_b$, $\|x_i\| \le \rho$ and noise covariance $\sigma_v^2 I$. The transport-aware average $\hat{v}_{\mathrm{bundle}} = \frac{1}{n}\sum_i P_{b_0 \leftarrow b_i} y_i$ is unbiased, while the reference-based average $\hat{v}_{\mathrm{prod}} = \frac{1}{n}\sum_i \tau_{b_0 \leftarrow b_i} y_i$ has leading bias $\mathrm{Bias}_{\mathrm{ref}} = -\frac{1}{2}\partial_{(k} a_{j)}(b_0)\Sigma_b^{jk} v^\star + O(\rho^3)$, which for the ambient reference on a constant-curvature surface equals $-\frac{1}{2}\kappa\Sigma_b v^\star$.

Scope of the unbiasedness claim. It holds when the latent field is a parallel-transported copy of $v^\star$ from b_0, as assumed throughout this section, and not for a general smooth section. We call it *model-unbiasedness*. Section 5 gives the bias $\mathrm{Bias}_{\mathrm{mod}}$ of transport for a general field V and shows that curvature rules out a nonzero globally parallel field, although the radial extension exists locally.

Two references with opposite biases. Proposition 6 (Appendix L) shows that this bias depends on the reference. On TS^2 with ∇ Levi-Civita, ambient projection τ^{amb} gives $\Phi_{\mathrm{amb}}(x) = I - \frac{1}{2}\kappa(x \otimes x) + O(\|x\|^4)$, hence $\mathrm{Bias}_{\mathrm{ref}}^{\mathrm{amb}} = -\frac{1}{2}\kappa\Sigma_b v^\star$; the normal-coordinate frame τ^{nc}, which keeps coefficients in the frame $d(\mathrm{Exp}_{b_0})_x$, gives $\Phi_{\mathrm{nc}}(x) = I + \frac{\kappa}{6}(\|x\|^2 I - x \otimes x) + O(\|x\|^4)$, hence $\mathrm{Bias}_{\mathrm{ref}}^{\mathrm{nc}} = +\frac{\kappa}{6}(\mathrm{tr}\,\Sigma_b I - \Sigma_b)v^\star$. For isotropic $\Sigma_b = \sigma_b^2 I_2$ these are $-\frac{1}{2}\kappa\sigma_b^2 v^\star$ and $+\frac{1}{6}\kappa\sigma_b^2 v^\star$: opposite in sign, and three times smaller in magnitude for τ^{nc}. The reference is therefore part of the estimator and should be reported, and switching to τ^{nc} is a cheap alternative to inserting transport (Appendix E).

Theorem 4 (Appendix L) gives the exact identity

$$\mathrm{MSE}(\hat{v}_{\mathrm{prod}}) - \mathrm{MSE}(\hat{v}_{\mathrm{bundle}}) = \Delta_{\mathrm{noise}} + \tfrac{1}{n}\,\mathrm{tr}\,\mathrm{Cov}\big(\Phi(x)v^\star\big) + \|\mathrm{Bias}_{\mathrm{ref}}\|^2, \quad \Delta_{\mathrm{noise}} := \tfrac{\sigma_v^2}{n}\,\mathrm{tr}\{\mathbb{E}[\tau\tau^*] - I\}, \tag{5}$$

whose three terms are transformed noise, base-sampling variation, and squared bias. If τ is fibrewise isometric, $\Delta_{\mathrm{noise}} = 0$ and the gap is nonnegative. For a non-isometric reference such as ambient projection the noise term must be kept, and the total can be negative. The product estimator is biased, so the unbiased Cramér–Rao bound does not apply to it; the MSE gap is the appropriate comparison (Appendix N.9).

A projection reference shrinks noise. Ambient projection is a contraction, with singular values $(\cos(\|x\|/R), 1)$ in a basis adapted to x. Hence $\tau^{\mathrm{amb}}(\tau^{\mathrm{amb}})^* \preceq I$ and $\Delta_{\mathrm{noise}} = -\sigma_v^2\kappa\,\mathrm{tr}\,\Sigma_b/n + O(\rho^4)$ (eq. equation 30, Appendix N). The reference estimator is thus a shrinkage estimator that trades bias for variance, and together with Propositions 6 and 7 this gives a threshold.

Corollary 1 (Leading-order curvature crossover)**.** *Under the model of Proposition 7 on TS^2 with the ambient reference, isotropic base dispersion $\Sigma_b = \sigma_b^2 I_d$, $d = 2$, and isotropic fibre noise, to leading order*

$$\mathrm{MSE}(\hat{v}_{\mathrm{prod}}) - \mathrm{MSE}(\hat{v}_{\mathrm{bundle}}) = \underbrace{\tfrac{1}{4}\kappa^2\sigma_b^4\|v^\star\|^2}_{\textit{squared bias}} - \underbrace{\tfrac{d\,\sigma_v^2\,\kappa\,\sigma_b^2}{n}}_{\textit{noise shrinkage}} + R_{\kappa,\rho,n}. \tag{6}$$

The leading approximation therefore favors the transport-aware estimator for $\kappa > \kappa^\star$, where

$$\kappa^\star \approx \frac{4d\,\sigma_v^2}{n\,\sigma_b^2\,\|v^\star\|^2}, \tag{7}$$

and the reference-based estimator for $0 < \kappa < \kappa^\star$. For a rotationally symmetric base law the gap is exactly $\|v^\star\|^2\big(\frac{\beta}{2n} + \frac{n-1}{4n}\alpha^2\big) - \frac{\sigma_v^2}{n}(2\alpha - \beta)$ with $\alpha = \mathbb{E}[1 - \cos(\|x\|/R)]$ and $\beta = \mathbb{E}[(1 - \cos(\|x\|/R))^2]$ (eq. equation 31).

Table 7 (Appendix F) compares equation 6 and the exact gap with simulation over two decades of curvature. The corollary turns the advice "use transport because it is correct" into an inequality that can be checked, and it identifies the regime of low curvature, small n and narrow dispersion in which transport does not pay for itself. It relies on the model of Proposition 7, in which the latent field is parallel along radial geodesics. Section 5 shows that dropping this assumption changes the answer.

5 When does transport matter? Misspecification, discretization, and topology

Section 4 assumed that the latent field is a transported copy of a single value. On a curved base this model makes the transport-aware estimator exactly unbiased, but it is special, since curvature prevents a nonzero field from being parallel in every direction. Below we drop the assumption, show that curvature rules out globally parallel fields, bound the effect of discrete paths and misspecified connections, and show that a global coordinate baseline may not exist. Table 1 summarizes the resulting regimes; proofs are in Appendix O.

A bias–variance tradeoff in the aggregation radius. Let V be a C^3 section over $B_\rho(b_0)$ with $y_i = V(b_i) + \varepsilon_i$ at $b_i = \mathrm{Exp}_{b_0}(x_i)$, $\mathbb{E}[x_i] = 0$, $\mathrm{Cov}(x_i) = \Sigma_b$, $\|x_i\| \le \rho$, $\mathrm{Cov}(\varepsilon_i) = \sigma_v^2 I_r$, and target $v^\star = V(b_0)$. Let $L_j = \sup_{B_\rho} \|\nabla^j V\|$ for $j = 1, 3$, and let $L_2 = \sup_{b,\|u\|=1} \|\nabla^2_{u,u} V(b)\|$ bound the symmetric Hessian. The main tool is the covariant Taylor expansion $P_{b_0 \leftarrow b} V(b) = V(b_0) + (\nabla_x V)(b_0) + \frac{1}{2}(\nabla^2_{x,x} V)(b_0) + O(L_3\|x\|^3)$ (Lemma 3); transport is what turns covariant derivatives into ordinary ones.

By Theorem 6 (Appendix O), $\mathrm{Bias}_{\mathrm{mod}} = \frac{1}{2}\Sigma_b^{jk}(\nabla^2_{jk} V)(b_0) + O(L_3\rho^3)$, so $\|\mathrm{Bias}_{\mathrm{mod}}\| \le \frac{1}{2} L_2 \operatorname{tr}\Sigma_b + O(L_3\rho^3)$, against a noise variance $r\sigma_v^2/n$ plus base-sampling variance. With $n = n_0\omega_d\rho^d$ points in the ball (ω_d the unit-ball volume) and $\operatorname{tr}\Sigma_b = c_d\rho^2$, the leading bound is minimized at $\rho^\star = (d\, r\sigma_v^2/(L_2^2 c_d^2 n_0 \omega_d))^{1/(d+4)}$, which gives the classical d-dimensional local-averaging rate (Tsybakov, 2009). To second order the reference estimator has bias $\mathrm{Bias}_{\mathrm{mod}} + \mathrm{Bias}_{\mathrm{ref}}$, and its noise and base-sampling variance change as well. The aggregation radius therefore has an optimum, and widening the window beyond $\rho^\star$ hurts. Since the two biases are of the same order in ρ, a dimensionless ratio decides between them.

Corollary 2 (Appendix O) makes this precise. Define the *transport-relevance ratio*

$$\Xi \;:=\; |\kappa|\,\|v^\star\| \,/\, L_2. \tag{8}$$

On TS^2 with isotropic sampling, transport has the smaller leading squared bias whenever $\Xi > 4$, whatever the alignment of the Hessian. When $\Xi \lesssim 1$ and $\mathrm{Bias}_{\mathrm{mod}}$ is sufficiently aligned with $\kappa v^\star$, the two errors partly cancel and the reference estimator has the smaller bias. The ratio alone does not determine the MSE, since the bias alignment and the noise also enter. Ξ can be estimated from angle defects or loop holonomy for the curvature, a covariant quadratic fit for $\|\operatorname{Sym}\nabla^2 V\|$, and a noise-corrected $\|v^\star\|$ (Appendix I.1). In Section 6, Ξ describes the bias regimes correctly, but ranking estimators at small n requires a plug-in estimate of the full identity.

Curvature, parallel fields, and anchor dependence. The full covariant Hessian need not be small when $|\kappa|$ is large, even though its symmetric part can vanish, because a covariantly constant section is a parallel section, and curvature obstructs parallel sections. Proposition 8 (Appendix O) has four parts. (i) A nonvanishing parallel section forces $F^\nabla v^\star \equiv 0$, so none exists on a curved patch of S^2 or of the torus. (ii) The radially parallel extension V^{b_0} of $v^\star$ exists locally, with $\|\nabla^2 V^{b_0}\| = O(|\kappa|\|v^\star\|)$ and matching lower bounds when $F^\nabla v^\star \neq 0$; this bound on the full Hessian does not force the symmetric contraction, and hence the model bias, to be nonzero. (iii) Extensions anchored at b_0 and b_1 differ by the holonomy of the triangle $b_0 b_1 b$. (iv) $\Sigma_b^{jk}(\nabla^2_{jk} V^{b_0})(b_0) = 0$, so the unbiasedness in Proposition 7 holds when the estimation anchor equals the anchor of the radial extension; another anchor can introduce bias, possibly with cancellations that depend on the sampling law. Curvature thus creates reference bias and rules out globally parallel fields, but it does not rule out the local radial model or bound the transport advantage uniformly. In our TS^2 and torus controls the estimation anchor coincides with the anchor of the field by design, and part (iii) shows that moving the anchor costs exactly a holonomy term.

Discretized paths and misspecified connections. Appendix H quantifies two differences between the theory and an implementation. Transport along a discrete path instead of the minimizing geodesic incurs the holonomy of the loop the two paths enclose. If the path stays within $O(h)$ of the geodesic, this gives an $O(|\kappa| h\rho\|v^\star\|)$ bias bound that grows with the window and does not average out (Prop. 2). Graph shortest paths need not satisfy the tracking assumption. On icospheres the enclosed rotation scales like ρ^2 and does not shrink with h, and averaging over two paths lowers

Table 1: Regimes; h is the mesh resolution and ρ the aggregation radius. The conditions are local diagnostics; ranking estimators also requires the noise, alignment and path assumptions.

Regime	Condition	Dominant term	Recommendation
Noise dominated	$0 < \kappa < \kappa^\star$	$\Delta_{\text{noise}} < 0$	leading approximation favors the reference (Cor. 1)
Misspecification dominated	$\Xi \lesssim 1$	Bias_{mod}	shrink ρ towards $\rho^\star$; compare alignment and variance
Discretization dominated	$h \gtrsim \rho$	holonomy of the path defect	fix the paths first (Prop. 2)
Frame-singular	$\mathcal{E}$ nontrivial (e.g. $e(\mathcal{E}) \neq 0$)	no global τ	reference baseline undefined (Prop. 1)
Transport dominated	$\kappa > \kappa^\star, \Xi \gg 1, h \ll \rho$	$\|\text{Bias}_{\text{ref}}\|^2$	**compare full risk**

it by about a third (Appendix I.4). A misspecified connection $\nabla + \delta$ with $\|\delta\|_{C^1} \leq \eta \leq 1$ costs $O(\eta\rho^2)$ under centred sampling but $O(\eta\rho)$ when $\mathbb{E}[x] \neq 0$, which is the situation at boundaries and in masked regions (Prop. 3).

Topology: when the coordinate baseline does not exist. Our two curved testbeds differ topologically, and this affects how their results should be read.

Proposition 1 (Existence of a global reference identification)**.** *A family $\{\tau_{b' \leftarrow b}\}$ of linear isomorphisms that is globally defined, continuous, and a cocycle ($\tau_{b'' \leftarrow b'} \tau_{b' \leftarrow b} = \tau_{b'' \leftarrow b}$, $\tau_{b \leftarrow b} = I$) exists on $\mathcal{E} \to \mathcal{M}$ if and only if $\mathcal{E}$ is trivial. Consequently:*

(i) The tangent bundle $T\mathbb{T}^2$ of the torus is trivial (∂_u, ∂_v form a global frame), so the coordinate-frame baseline is defined everywhere. This is the best case for a pipeline built on a trivialization.

(ii) Since $\chi(S^2) \neq 0$, TS^2 admits no global frame and no such τ exists. A global coordinate baseline must then be singular, patched, or replaced by a non-cocycle family such as ambient projection, which is smooth but loses rank at geodesic distance $\pi R/2$ from its anchor. Local frame singularities need a separate analysis.

The $O(\rho^2)$ law of Section 4 is therefore local, and on a nontrivial bundle (for example one with nonvanishing Euler class) it has no uniform global version.

What Proposition 1 does not rule out. It concerns a single globally consistent trivialization, which a pipeline assumes when it expresses every fibre value in one reference frame. Gauge-equivariant networks route messages through local gauges and avoid the obstruction, and patched or anchor-local references handle the multi-chart case (Appendix I.5). The torus is the easy case topologically; Section 6 tests a surface with no global frame.

6 Experiments

The sphere studies isolate the reference bias and the crossover (Section 6), exercise the full Bundle-GEM loop on incomplete data (Section 6) and test whether the diagnostics predict which estimator is better (Section 6). A genus-zero surface tests the topological obstruction, and the torus studies test variable curvature and a gauge-equivariant encoder (Section 6). A secondary study with SPD fibres on a flat base is in Appendix S. Code is at `https://anonymous.4open.science/r/transport-aware-bundles-73F3/`; all intervals are two standard errors over independent seeds or paired trials.

Reference bias and the crossover on TS^2. We place a unit vector $v^\star$ at b_0 on a sphere of radius R and observe noisy transported copies $y_i = P_{b_i \leftarrow b_0} v^\star + \varepsilon_i$ at base points drawn uniformly from the tangent disk of radius 0.6, so Σ_b does not change with κ ($n = 40$, $\sigma_v = 0.15$, 10^5 paired trials per curvature). Transporting to b_0 before averaging is unbiased, while averaging ambient coordinates and projecting has a bias that grows linearly in κ (slope 0.044, $R^2 = 0.9999$, against the leading-order 0.045 of Prop. 7); at $\kappa = 1.56$ its MSE is 5.2 times larger (Fig. 2 and Table 8, Appendix F). The measured MSE gap agrees with the exact formula equation 31 within two standard errors at every curvature (Table 7). It changes sign between $\kappa = 0.02$ and 0.05, as predicted ($\kappa^\star = 0.048$ exactly, 0.050 from equation 7). At $\kappa = 0.02$ the biased reference estimator has the lower MSE, by nine

Table 2: Full Bundle-GEM on TS^2 with 50% of the fibre values missing and 30% of the rest observed through one component ($\sigma = 0.1$): RMSE of the imputed missing fibres, mean over 20 seeds. The best method in each column is in bold (ties at the displayed precision are all marked); the oracle row is a reference and is not ranked. Two standard errors of the paired differences to Bundle-GEM are at most 0.015, except for ambient EM at $\kappa = 2.25$; all settings are in Appendix V.1.

	data straddling a pole			data on the equator		
Curvature κ	0.25	1.00	2.25	0.25	1.00	2.25
Bundle-GEM (transport)	**0.306**	**0.316**	**0.357**	**0.322**	**0.320**	**0.324**
EM, ambient reference	0.307	0.326	1.97	**0.322**	0.330	2.14
EM, global coordinate frame	0.572	0.566	0.543	**0.322**	0.322	0.337
One-shot transport (no EM)	0.323	0.337	0.383	0.339	0.348	0.360
True parameters (oracle)	0.304	0.315	0.353	0.318	0.319	0.321

standard errors. Matching the model connection to the flat connection that induces τ removes the gap, and transport around geodesic triangles rotates vectors by $\kappa\times$area to three decimals (Table 9, Appendix F).

Full Bundle-GEM on incomplete data. We fit the two-component transported mixture equation 1 with moving anchors to $n = 400$ points on TS^2 whose fibre values are fully observed, observed through one random linear component, or missing (20–80%), with Gaussian noise. The E-step is exact, while the M-step combines closed-form updates of (π, μ, Σ) with Armijo steps of the transported retraction for the anchors. We compare with EM under an ambient projection or global spherical-coordinate frame, and with one-shot transport imputation, over 20 settings of curvature, placement, missingness and noise with 20 seeds each (Appendix V.1). The observed log-likelihood never decreased in any of the $1{,}200$ runs. Table 2 shows imputation error at 50% missingness. Bundle-GEM is within 0.004 of the true-parameter error here and within 0.006 across the full grid. Ambient-reference EM matches it at low curvature, is worse at $\kappa = 1$, and breaks down at $\kappa = 2.25$, where data reach $\pi R/2$ and projection loses rank. The coordinate frame is close to Bundle-GEM on the equator but fails whenever data straddle a pole, at every curvature, as Proposition 1 predicts. One-shot transport, which skips the likelihood, is worse than Bundle-GEM in all 20 settings, by 0.014–0.072.

Predicting the better estimator. We construct fields on the sphere with prescribed curvature, symmetric Hessian and alignment (105 settings with Ξ from 0.25 to 16; Appendix V.2) and measure both estimators with 5×10^4 paired trials each. Corollary 2 holds in all 30 settings with $\Xi > 4$, while in all 15 settings with $\Xi \leq 1$ and aligned Hessian the reference estimator has smaller bias, as predicted. At $n = 40$, however, the sampling-variance term of equation 5 is as large as these effects, so Ξ alone is a poor decision rule, and a threshold at 4 picks the better estimator in 49% of resolved cases. A plug-in estimate of all terms in the exact identity, fitted on an independent calibration sample of the same field, picks the better estimator in 94% of resolved cases, with mean excess MSE 0.1%, against 18% for always transporting and 52% for always using the reference.

A surface without a global frame. On an icosphere with 2562 vertices, we reconstruct 50% masked values of smooth random tangent fields from observed neighbours (20 fields; Appendix V.3). A single global frame fails near its singular poles (RMSE 0.34 against 0.05 for transport at radius 0.25). A two-chart atlas, which avoids its singularities, removes the failure and stays within 0.004 of transport. The normal-coordinate reference is slightly but consistently better than transport (-0.0012 ± 0.0001).

Tangent fields on a torus. The torus tangent bundle is trivial, so a global coordinate frame exists (Proposition 1); its discrete Levi-Civita connection has curvature of both signs (Appendix Q). On data from the radial model, transport is exact and ambient error grows with $|K|$ and ρ (Table 13, Appendix Q). In the imputation task of Table 14, one-shot transport did worst. A mechanism test (Appendix V.4) shows why. The field is generated by transport along a spanning tree, whose own

Table 3: Masked reconstruction of real 300 hPa winds on the four held-out months: RMSE relative to the field RMS, mean over 3 training seeds. Two standard errors over test fields are at most 0.009 (random orientations, 64 fields) and 0.023 (natural orientation, 4 fields); the gauge network is better than every other model on every test field.

Network	random orientation	near frame poles	natural orientation
Gauge-equivariant (Levi-Civita transport)	**0.152**	**0.150**	**0.144**
Standard, 3-D ambient components	0.172	0.171	0.164
Standard, east/north components	0.189	0.366	0.160
Gauge network without transport	0.237	0.214	0.200

paths recover it to the noise floor, whereas every local rule (one-shot transport, the coordinate frame and ambient projection) errs where tree and local paths enclose holonomy, with correlation 0.95 between one-shot error and holonomy. On a smooth field, one-shot transport is within 0.003 of the best rule, so the poor result reflects the test field.

Real wind fields and a gauge-equivariant network. We use NCEP/NCAR reanalysis (Kalnay et al., 1996) long-term monthly mean winds at 300 hPa, a global tangent field on a 642-vertex icosphere (Appendix V.5). Networks are trained on 8 months (randomly rotated copies) and tested on 4 held-out months in their natural and 16 random orientations, with 50% of vertices masked. All use four message-passing layers over six neighbours, see the mask and positions, and are trained for 3000 Adam steps with 3 seeds; they differ only in how neighbouring fibres are compared (Table 3). The gauge-equivariant network transports messages with the Levi-Civita connection (equivariance error 2×10^{-7}) and is best on every test field. Removing only the transport phases raises error by 38–56%, isolating the connection's contribution. A latitude–longitude network is competitive in the natural orientation, where flow is mostly zonal, but its error near the frame poles more than doubles after rotation. Diagnostics give $\widehat{\Xi} < 1$ at all 642 locations (median 0.08), placing real winds in the field-variation-dominated regime; transport averaging nevertheless beats ambient averaging at only 67% of locations. The plug-in identity estimated on training months predicts the better estimator on held-out months at 97.7% of locations, whereas a threshold on $\widehat{\Xi}$ does not.

7 Related work

EM and coarsened data (Dempster et al., 1977; Wu, 1983; McLachlan & Krishnan, 2008; Heitjan & Rubin, 1991) supply our ascent arguments; intrinsic statistics (Karcher, 1977; Pennec et al., 2006; Smith, 2005) works on a single nonlinear space rather than fibres. Closest in statistical theory is Das & Snášel (2026), who give concentration and minimax bounds for the transported mean, with a floor set by ambiguity between two transport rules. We study instead the error from replacing transport by a non-transport reference, which depends on the reference and, since the reference need not be an isometry, can beat the unbiased transported estimator. Sheaf and connection learning, Lie-group uncertainty, Fréchet regression, gauge-equivariant networks, geometric Gaussian processes, and Riemannian mixture optimization (Hansen & Gebhart, 2020; Bodnar et al., 2022; Barfoot & Furgale, 2014; Petersen & Müller, 2019; de Haan et al., 2021; Park et al., 2023; Hutchinson et al., 2021; Hosseini & Sra, 2015) are compared in Appendix J and Table 10; we give criteria for when their transport step helps.

8 Conclusion

Our analysis shows that the benefit of parallel transport depends on the relative sizes of the reference bias it removes and the variance reduction it forgoes. This trade-off is controlled by curvature, field variation, sample size, and transport-path accuracy. The exact MSE identity provides a practical way to estimate it, and in our experiments it selected the lower-error estimator in 94% of the synthetic cases and at 98% of the real-wind locations. For geometric deep learning, this turns transport from a default design choice into one that can be checked on data. Our formulas hold when neighbourhoods are small relative to curvature, as in message passing.

REPRODUCIBILITY STATEMENT

Assumptions and proofs are in Appendix N, and experimental protocols in Appendices T, V.1–V.4, V.5, S and R. The anonymized repository at `https://anonymous.4open.science/r/transport-aware-bundles-73F3/` contains the code, random seeds and scripts for every table and figure.

ETHICS STATEMENT

This is a methods paper on geometric statistical inference, and we see no direct ethical concerns. Applications to medical imaging should follow standard data-governance practice.

AI USE STATEMENT

We used an AI assistant for language editing, literature search, and checking theorem statements and assumptions. The assistant also contributed to the experimental design, synthetic data, and wrote and ran code for some experiments; all code is in the released repository. The authors verified the code and results and take full responsibility for the content of the paper.

REFERENCES

P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007. doi: 10.1007/s10208-005-0179-9.

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

Bijan Afsari. Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011. doi: 10.1090/S0002-9939-2010-10541-5.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Warren Ambrose and Isadore M. Singer. A theorem on holonomy. *Transactions of the American Mathematical Society*, 75(3):428–443, 1953. doi: 10.1090/S0002-9947-1953-0063739-1.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006. doi: 10.1002/mrm.20965.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017. doi: 10.1214/16-AOS1435.

Timothy D. Barfoot and Paul T. Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Transactions on Robotics*, 30(3):679–693, 2014. doi: 10.1109/TRO.2014.2298059.

Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. doi: 10.1109/TPAMI.2003.1177153.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003. doi: 10.1214/aos/1046294456.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18527–18541, 2022. doi: 10.52202/068431-1346. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/75c45fca2aa416ada062b26cc4fb7641-Abstract-Conference.html.

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

Guillaume Bourmaud, Rémi Mégret, Marc Arnaudon, and Audrey Giremus. Continuous-discrete extended Kalman filter on matrix Lie groups using concentrated Gaussian distributions. *Journal of Mathematical Imaging and Vision*, 51(1):209–228, 2015. doi: 10.1007/s10851-014-0517-0.

Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6e69ebbfad976d4637bb4b39de261bf7-Abstract.html.

Rudrasis Chakraborty, Jose Bouza, Jonathan Manton, and Baba C. Vemuri. ManifoldNet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. doi: 10.1109/TPAMI.2020.3003846.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19d.html.

Keenan Crane, Mathieu Desbrun, and Peter Schröder. Trivial connections on discrete surfaces. *Computer Graphics Forum*, 29(5):1525–1533, 2010. doi: 10.1111/j.1467-8659.2010.01761.x.

Swagatam Das and Václav Snášel. Sharp concentration bounds for bundle-valued statistics on manifolds. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2026. doi: 10.48550/arXiv.2607.10592. URL https://arxiv.org/abs/2607.10592.

Andrew Dawson. Windspharm: A high-level library for global wind field computations using spherical harmonics. *Journal of Open Research Software*, 4(1):e31, 2016. doi: 10.5334/jors.129.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2406–2422, 2022. doi: 10.52202/068431-0175. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/105112d52254f86d5854f3da734a52b4-Abstract-Conference.html.

Pim de Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh CNNs: Anisotropic convolutions on geometric graphs. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Jnspzp-oIZE.

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

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6): 3161–3182, 2020. doi: 10.1214/19-AOS1924.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

Benjamin Eltzner. Geometrical smeariness—a new phenomenon of Fréchet means. *Bernoulli*, 28 (1):239–254, 2022. doi: 10.3150/21-BEJ1340.

P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8): 995–1005, 2004. doi: 10.1109/TMI.2004.831793.

Gersende Fort and Eric Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259, 2003. doi: 10.1214/aos/1059655912.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 376–383. ACM, 2008. doi: 10.1145/1390156.1390204.

Jakob Hansen and Thomas Gebhart. Sheaf neural networks. In *NeurIPS Workshop on Topological Data Analysis and Beyond*, 2020. doi: 10.48550/arXiv.2012.06333. URL https://arxiv.org/abs/2012.06333.

Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19 (4):2244–2253, 1991. doi: 10.1214/aos/1176348396.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for Gaussian mixtures. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/dbe272bab69f8e13f14b405e038deb64-Abstract.html.

Michael Hutchinson, Alexander Terenin, Viacheslav Borovitskiy, So Takao, Yee Whye Teh, and Marc Deisenroth. Vector-valued Gaussian processes on Riemannian manifolds via gauge-independent projected kernels. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17160–17169, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8e7991af8afa942dc572950e01177da5-Abstract.html.

Sungkyu Jung, Brian Rooks, David Groisser, and Armin Schwartzman. Averaging symmetric positive-definite matrices on the space of eigen-decompositions. *Bernoulli*, 31(2):1552–1578, 2025. doi: 10.3150/24-BEJ1783.

Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, Yuejian Zhu, Muthuvel Chelliah, Wesley Ebisuzaki, Wayne Higgins, John Janowiak, Kingtse C. Mo, Chester Ropelewski, Julian Wang, Ants Leetmaa, Richard Reynolds, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996. doi: 10.1175/1520-0477(1996)077⟨0437:TNYRP⟩2.0.CO;2.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977. doi: 10.1002/cpa.3160300502.

Felix Knöppel, Keenan Crane, Ulrich Pinkall, and Peter Schröder. Globally optimal direction fields. *ACM Transactions on Graphics*, 32(4):59:1–59:10, 2013. doi: 10.1145/2461912.2462005.

Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry, Volume I*. Wiley Interscience, 1963.

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 44(2):226–233, 1982. doi: 10.1111/j.2517-6161.1982.tb01203.x.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Georges Matheron. *Random Sets and Integral Geometry*. John Wiley and Sons, New York, 1975.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2503–2515, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1aa3d9c6ce672447e1e5d0f1b5207e85-Abstract.html.

Emile Mathieu, Vincent Dutordoir, Michael Hutchinson, Valentin De Bortoli, Yee Whye Teh, and Richard Turner. Geometric neural diffusion processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53475–53507, 2023. doi: 10.52202/075280-2327. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Abstract-Conference.html.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley, 2nd edition, 2008.

Ilya Molchanov. *Theory of Random Sets*. Springer, London, 2005. doi: 10.1007/1-84628-150-4.

Josep M. Oller and José M. Corcuera. Intrinsic analysis of statistical estimation. *The Annals of Statistics*, 23(5):1562–1581, 1995.

Jung Yeon Park, Lawson L. S. Wong, and Robin Walters. Modeling dynamics over meshes with gauge equivariant nonlinear message passing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15277–15302, 2023. doi: 10.52202/075280-0672. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/317470b3fde29f3bb8d6dee563afffc4-Abstract-Conference.html.

Xavier Pennec. Curvature effects on the empirical mean in Riemannian and affine manifolds: a non-asymptotic high concentration expansion in the small-sample regime. *arXiv preprint arXiv:1906.07418*, 2019. doi: 10.48550/arXiv.1906.07418. URL https://arxiv.org/abs/1906.07418.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. doi: 10.1007/s11263-005-3222-z.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019. doi: 10.1214/17-AOS1624.

Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H. Manton. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017. doi: 10.1109/TIT.2017.2653803.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Steven T. Smith. Covariance, subspace, and intrinsic Cramér–Rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963. doi: 10.1214/aoms/1177703862.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: Image processing in Python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.

Joachim Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart, 1998.

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

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Ho Yun and Byeong U. Park. Exponential concentration for geometric-median-of-means in non-positive curvature spaces. *Bernoulli*, 29(4):2927–2960, 2023. doi: 10.3150/22-BEJ1569.

Appendix

Organization. Notation is collected in App. A. *Setting and algorithm:* image-subspace motivation (App. B), algorithm and cost (App. C, P), standing assumptions (App. D), observation kernels (App. G), and E-step statistics (App. M). *Theory:* guarantees (App. K), single-path expansion and MSE identity (App. L), discretized paths and misspecified connections (App. H), and proofs (App. N, O). *Practice and positioning:* reporting guidance (App. E), diagnostics and design choices (App. I), and related work with a positioning table (App. J). *Experiments:* TS^2 tables (App. F), torus mesh (App. Q), gauge-equivariant pipeline (App. R), SPD fields (App. S), protocols and code (App. T), full Bundle-GEM on incomplete data (App. V.1), validation of the diagnostics (App. V.2), a surface without a global frame (App. V.3), the torus mechanism test (App. V.4), real wind fields (App. V.5), and additional analyses (App. U).

A Notation and symbol disambiguation

Table 4 collects the principal symbols; L_2 bounds the symmetric covariant Hessian.

Table 4: Principal notation. A target–source subscript is used consistently for transport.

Symbol	Meaning	Space / qualification
$\mathcal{M}, \mathcal{E}, \pi$	base manifold, total space, bundle projection	$\pi : \mathcal{E} \to \mathcal{M}$
$b, v, z = (b, v)$	base point, fibre value, bundle-valued latent state	$v \in \mathcal{E}_b$
∇, ∇^0, a	model connection, reference connection, difference	$a = \nabla - \nabla^0$
$P_{b' \leftarrow b}$	parallel transport from b to b'	$\mathcal{E}_b \to \mathcal{E}_{b'}$
$\tau_{b' \leftarrow b}$	specified smooth product/reference map	linear; need not be isometric
$\delta_{\mathrm{ref}}(b, v)$	$\tau_{b_0 \leftarrow b} v - P^{\nabla}_{b_0 \leftarrow b} v$	open-path; reference dependent
$F^{\nabla}, A^{jk}, \mathrm{Hol}$	curvature, oriented area, loop holonomy	operator gauge covariant; conjugacy data invariant
θ, ψ, ϑ	latent-model, observation, joint parameters	$\vartheta = (\theta, \psi)$
q_{ik}, N_k, s_k, T_k	responsibilities and transported sufficient statistics	mixture component k
Σ_k, Σ_b	fibre covariance, base-displacement covariance	distinct objects
$\Phi(x), \mathrm{Bias}_{\mathrm{ref}}$	relative single-path map and its aggregation bias	reference dependent
ρ, r, d, K	support radius, fibre rank, base dimension, components	scalars

B Image-subspace models and the gauge-theory connection

Modelled structure. Low intrinsic dimension does not supply a vector bundle. We assume a smooth base $\mathcal{M}$ and a family of rank-r feature spaces with smooth local trivializations and linear transition maps. The feature attached to b lies in $\mathcal{E}_b$; a deterministic feature field is a section $V : \mathcal{M} \to \mathcal{E}$ satisfying $\pi \circ V = \mathrm{Id}$. A conditional distribution on $\mathcal{E}_b$ describes uncertainty about that feature. Tangent directions due to pose or illumination, local appearance spaces, and oriented derivative channels represent different choices of fibre; the model must specify which it uses. Approximate illumination subspaces motivate local linear descriptions under restricted imaging assumptions (Basri & Jacobs, 2003); they do not establish a global manifold model for unrestricted natural images.

An explicit image-subspace bundle. Represent a vectorized image or patch in $\mathbb{R}^D$. Let b be a rank-r local appearance subspace, estimated, for example, from a controlled image family. Assume the selected subspaces form a smooth submanifold $\mathcal{M} \subseteq \mathrm{Gr}(r, D)$. Then

$$\mathcal{E} = \{(b, v) : b \in \mathcal{M},\ v \in b\}, \qquad \pi(b, v) = b, \qquad \mathcal{E}_b = b$$

is the tautological bundle restricted to $\mathcal{M}$. This makes the base a manifold of subspaces, and the fibre the corresponding appearance space; subspace-based recognition provides relevant precedent (Hamm & Lee, 2008; Edelman et al., 1998). For an orthonormal basis $U_b \in \mathbb{R}^{D \times r}$, write $v = U_b a$. Replacing U_b by $U_b g_b$, $g_b \in O(r)$, changes a to $g_b^{-1} a$ and leaves v unchanged. The ambient orthogonal-projection connection $\nabla V = \Pi_b \, dV$ is one possible modeling choice; the paper does not claim that it is uniquely suitable for an image-learning task. The quantities involved are summarized in Table 5, which separates what an empirical implementation must report.

A learned map into one fixed $\mathbb{R}^r$ already defines a trivial feature bundle; identity comparison is available, and introducing another connection needs a task-specific justification. A spectral gap

Table 5: Distinct dimensions and assumptions in the image-subspace example. The image-subspace bundle is a motivating construction; the experiments of this paper use tangent bundles.

Quantity	Meaning
D	Ambient pixel/patch dimension, including channels if used.
$d = \dim \mathcal{M}$	Dimension of the modeled subspace family; at most $r(D-r)$, and not necessarily the dimension of a manifold of individual images.
$r = \dim \mathcal{E}_b$	Constant rank of each local appearance space; independent of d.
$d + r$	Dimension of the total space of the resulting vector bundle.

Algorithm 1 Bundle Generalized Expectation–Maximization (Bundle-GEM)

Require: Observations $\{Y_i\}_{i=1}^n$, initial $(\theta^{(0)}, \psi^{(0)})$, tolerance $\epsilon > 0$.

1: **repeat**
2: **for** each observation i and component k **do**
3: Compute $q_{ik}^{(t)}(z) \propto \mathcal{K}_{\psi^{(t)}}(Y_i \mid z)\pi_k^{(t)} f_k(z; \theta_k^{(t)})$.
4: **end for**
5: Compute N_k, s_k, T_k by transporting fibre statistics to $\mathcal{E}_{m_k^{(t)}}$.
6: Update (π_k, μ_k, Σ_k) using feasible block maximizers (Appendix M); update (m_k, β_k, ψ) by an Armijo transported retraction step equation 12 that increases the full Q; reevaluate transported moments at each trial anchor.
7: **until** relative likelihood increment and the feasible-gradient norm are both below tolerance
8: **return** $(\theta^\star, \psi^\star)$.

can stabilize a local PCA rank, but a finite collection of estimated subspaces does not itself prove smoothness. Changing rank, occlusion boundaries, and mixtures of classes can require variable-rank, stratified, or piecewise models outside the smooth constant-rank analysis.

Gauge transformations preserve a fixed estimator. For local orthonormal frames with gauge matrices g_b, the coordinate matrices transform as

$$v_b' = g_b^{-1} v_b, \qquad P'_{b_0 \leftarrow b} = g_{b_0}^{-1} P_{b_0 \leftarrow b} g_b, \qquad \tau'_{b_0 \leftarrow b} = g_{b_0}^{-1} \tau_{b_0 \leftarrow b} g_b.$$

Thus $\Phi' = g_{b_0}^{-1} \Phi g_{b_0}$ and $\hat{v}' - v^{\star\prime} = g_{b_0}^{-1}(\hat{v} - v^\star)$, so the metric loss and statistical risk are unchanged. For general invertible frames, the metric matrix must transform as well. Changing τ as a geometric map, while keeping the model and target fixed, instead changes the estimator. Its bias is therefore reference dependent, although the bias vector is gauge covariant and its metric norm is gauge invariant under consistent re-expression.

A local frame is a gauge choice, a representation of the frame group determines an associated vector bundle, and a connection specifies comparison of neighbouring fibres. Holonomy transforms by conjugation around a based loop; trace, spectrum, and metric-invariant conjugacy data are representation independent. This is the mathematically relevant physics analogy. It supplies neither a physical law for image features nor a proof that a chosen connection is optimal or identifiable. Learning the connection requires additional observation assumptions and validation.

C Algorithm and computational cost

Computational considerations. The base Fréchet mean is obtained by a few Riemannian gradient steps, each an $\mathrm{Exp}/\mathrm{Log}$ pair, and the fibre statistics cost one transport per posterior sample. With P particles per observation, the additional dense-transport cost is $O(nKPr^2)$, excluding path construction. Posterior evaluation and sampling are shared with the corresponding product model. Posterior integrals may use the same Monte-Carlo or Laplace scheme as ordinary EM; the geometry enters only through $\mathrm{Exp}, \mathrm{Log}, P_{\leftarrow}$, which are available for every bundle we treat. The deterministic theorems below assume an exact E-step. A finite-particle approximation instead requires increasing sample sizes or stochastic-approximation conditions from MCEM/SAEM theory (Fort & Moulines,

2003); it does not inherit deterministic ascent automatically. A per-operation breakdown (posterior evaluation, Karcher iterations, transport, fibre aggregation, Monte-Carlo E-step, and M-step) is given in Appendix P; the general posterior case is dominated by the number of latent particles and inner Karcher iterations, not by the single extra transport.

D STANDING REGULARITY ASSUMPTIONS

Assumption 1 (Geometric and algorithmic regularity)**.** (a) $(\mathcal{M}, g_{\mathcal{M}})$ is complete and the base marginal of each posterior $q_i^{(t)}$ is supported (up to null sets) within a common geodesic ball of radius $\rho < \frac{1}{2}\min\{\mathrm{inj}(\mathcal{M}), \pi/\sqrt{\Delta}\}$, where $\Delta \geq 0$ upper-bounds sectional curvature (Karcher–Kendall condition), guaranteeing a unique base Fréchet mean (Afsari, 2011). (b) Parallel transport uses the selected minimizing geodesics. (c) The observed likelihood superlevel set containing the iterates is compact; Q and ℓ_{obs} are C^1 with Lipschitz Riemannian gradients there. (d) The preconditioner satisfies $0 < mI \preceq G_\vartheta \preceq MI$, and the smooth retraction with Armijo backtracking yields equation 3 and a step bounded by $C\|\operatorname{grad}\ell_{\mathrm{obs}}\|$. (e) The smooth stationarity theorem concerns interior iterates in a compact subset of the open parameter manifold. If a weight or spectral constraint is active, feasible updates still give ascent, but the appropriate limit condition is constrained stationarity; the unconstrained proof below does not establish that boundary case. The ball condition concerns base geometry, not bounded Gaussian fibres.

E SCOPE OF THE SINGLE-PATH EXPANSION, AND REPORTING GUIDANCE

Scope of equation 4. For a fixed anchor, use the smooth relative map $\Phi(x) = \tau_{b_0 \leftarrow b} P_{b \leftarrow b_0}$ and the invariant first-order condition $D\Phi(0) = 0$. A local frame defines a cocycle on its chart, hence a flat connection there; an orthonormal frame makes that connection metric compatible. Even a non-cocycle pairwise family can locally induce a flat reference by fixing an invertible anchor map and defining all other maps through that anchor. This induced cocycle need not reproduce the original pairwise family.

Table 6: Fibre identifications. Cocycle means composition for endpoint-selected pairwise maps, not composition along concatenated paths.

Identification	Isometric?	Endpoint cocycle?	Scope
Selected-geodesic ∇ transport	yes	generally no if curved	normal neighbourhood
Flat metric connection, trivial holonomy	yes	yes	local; global if holonomy trivial
Local frame identification	only for orthonormal frames	yes on chart	chart dependent
Ambient projection	generally no	generally no	global linear map; may lose rank

Remark 1 (Consequences for reporting). The reference and the path rule are part of the estimator and should be reported. On TS^2 with isotropic sampling, τ^{nc} has one third of τ^{amb}'s leading bias magnitude, with opposite sign; its leading squared bias is one ninth as large. The MSE advantage is smaller, because the normal-coordinate reference expands transverse noise while ambient projection contracts radial noise; the exact risk identity compares them.

F CONTROLLED TS^2 STUDY: FIGURE AND FULL TABLES

Setup and conventions. Tables 7 and 8 and Figure 2(a,b) come from a single sweep over ten curvatures with the protocol of Appendix T; both estimators are evaluated on the same simulated data, so the reported standard errors are those of paired differences. The exact gap equation 31 lies within two standard errors of the measured gap at every curvature, and the leading approximation equation 6 is within about 10% of it wherever the gap is resolved. Because the data are generated by transporting a common vector, exact recovery by the transport-aware estimator checks the implementation and the constants; it does not independently validate the data model.

Robust run with outliers. Figure 2(c) uses $n = 120$ observations at $\kappa = 1$ with $\sigma_v = 0.2$, of which 25% are replaced by outliers drawn uniformly from the disk of radius 1 in the tangent plane at their base point. Each method fits a two-component model (Gaussian inliers, uniform

Table 7: Curvature crossover on TS^2 with the ambient reference ($n = 40$, $\sigma_v = 0.15$, base points uniform on the tangent disk of radius 0.6, $\|v^\star\| = 1$, 10^5 paired trials per curvature). Gaps are $\mathrm{MSE}(\hat{v}_{\mathrm{prod}}) - \mathrm{MSE}(\hat{v}_{\mathrm{bundle}})$ in units of 10^{-6}, with two standard errors of the paired difference. "Exact" is equation 31 for this base law and "leading" is equation 6. The exact crossover is $\kappa^\star = 0.048$.

Curvature κ	0.02	0.05	0.10	0.21	0.50	1.00	1.56
MSE Product-EM	0.001126	0.001128	0.001137	0.001200	0.001589	0.003058	0.005815
MSE Bundle-GEM	0.001127	0.001128	0.001126	0.001128	0.001121	0.001131	0.001122
gap, measured (10^{-6})	-1.26 ± 0.28	$+0.50 \pm 0.69$	$+11.0 \pm 1.4$	$+71.9 \pm 2.9$	$+467 \pm 6.8$	$+1927 \pm 14$	$+4693 \pm 22$
gap, exact (10^{-6})	-1.18	$+0.22$	$+11.0$	$+71.3$	$+468$	$+1933$	$+4681$
gap, leading (10^{-6})	-1.21	0.00	$+10.1$	$+68.0$	$+456$	$+1924$	$+4770$

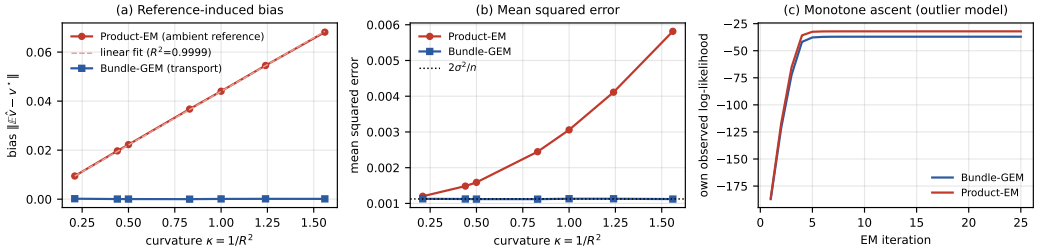

Figure 2: Controlled TS^2 study. (a) The bias of Product-EM grows linearly in $\kappa = 1/R^2$ ($R^2 = 0.9999$), while Bundle-GEM has no measurable bias. (b) The MSE of Bundle-GEM stays at the noise floor $2\sigma_v^2/n$; the MSE of Product-EM is dominated by its bias. (c) Robust EM with 25% outliers: both loops increase their own observed likelihood monotonically; the two likelihoods belong to different models and are not directly comparable.

outliers on the same disk) by EM with closed-form updates of v, σ and the inlier weight. Both loops increase their own observed likelihood at every iteration. The Product-EM loop ends at the larger value because projection shrinks residuals, which is why likelihoods of the two models cannot be compared directly. The estimates can: over 200 seeds the mean error $\|\hat{v} - v^\star\|$ is 0.032 for Bundle-GEM and 0.064 for Product-EM, and Bundle-GEM is closer to $v^\star$ in 95% of the seeds.

G Observation kernels and instantiations

Definition 1 (Bundle-valued observation model)**.** A latent bundle variable $Z_i = (b_i, v_i) \sim f_\theta$ has density f_θ w.r.t. $\mu_{\mathcal{E}}$. For the fixed normal region U, restrict this measure to $\pi^{-1}(U)$ (equivalently set $f_\theta = 0$ outside); positivity assumptions refer to this common support. Let $(\mathcal{Y}, \lambda)$ be the observation space with its dominating measure. An observation Y_i is generated by a Markov density $\mathcal{K}_\psi(Y \mid z)$ satisfying $\int_{\mathcal{Y}} \mathcal{K}_\psi(Y \mid z) d\lambda(Y) = 1$, giving observed likelihood

$$L(\theta, \psi) = \prod_{i=1}^{n} p_{\theta,\psi}(Y_i), \qquad p_{\theta,\psi}(Y) = \int_{\mathcal{E}} \mathcal{K}_\psi(Y \mid z)\, f_\theta(z)\, d\mu_{\mathcal{E}}(z). \tag{9}$$

The formulation covers deterministic set-censoring and soft probabilistic kernels alike:

$$\mathcal{K}(Y \mid z) = h(Y)\mathbf{1}\{z \in Y\} \qquad \text{(normalized coarsening design)}, \tag{10}$$

$$\mathcal{K}(Y \mid b, v) \propto \exp\Big(- \tfrac{d_{\mathcal{M}}(b, b_Y)^2}{2\sigma_b^2} - \tfrac{\|P_{b_Y \leftarrow b} v - v_Y\|_g^2 \mathcal{E}}{2\sigma_v^2} \Big). \tag{11}$$

In equation 11, the proportionality constant is the normalizer with respect to $d\lambda(Y)$; any z- or ψ-dependent normalizer is retained in the likelihood and in Q. For the first, require $\int h(Y)\mathbf{1}\{z \in Y\} d\lambda(Y) = 1$ for every z; a deterministic partition has $h = 1$ under counting measure. Under ignorable coarsening, a known parameter-independent $h(Y)$ can be omitted from likelihood ratios,

Table 8: Controlled TS^2 study across curvatures, from the same sweep as Table 7 (10^5 paired trials per curvature; noise floor $2\sigma_v^2/n = 0.001125$). The error of Product-EM is dominated by its bias, which grows linearly in κ (Proposition 7); Bundle-GEM stays at the floor.

Curvature $\kappa = 1/R^2$	0.21	0.44	0.83	1.00	1.24	1.56
Bundle-GEM bias	0.0002	0.0001	0.0000	0.0001	0.0001	0.0001
Product-EM bias	0.0094	0.0197	0.0368	0.0440	0.0545	0.0681
Bundle-GEM MSE	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
Product-EM MSE	0.0012	0.0015	0.0024	0.0031	0.0041	0.0058

Table 9: Mechanism checks on TS^2. Left: the Product-vs-Bundle gap vanishes when the model connection equals the flat reference connection inducing τ. Right: loop holonomy equals $\kappa\times$area across triangle sizes.

Flat-connection control	gap	Loop area A	0.013	0.064	0.281
curved connection	0.0060	holonomy (rad)	0.013	0.064	0.281
flat connection	0.0000	κA (theory)	0.013	0.064	0.281

leaving the set-probability likelihood (Matheron, 1975; Molchanov, 2005; Heitjan & Rubin, 1991). Arbitrary overlapping sets do not make the bare indicator a normalized Markov kernel. The second kernel is bundle-valued because the fibre residual is measured only after transport $P_{b_Y \leftarrow b}$, so the kernel depends on ∇. The basic EM ascent identity applies to both kernels. The differentiable convergence and consistency results below apply to the smooth positive model equation 11; hard set-censoring is covered only when differentiation and domination hold for its fixed support; an indicator in the latent variable does not itself imply nonsmoothness in the parameters.

Two instantiations. Two cases recur below. *(i) Tangent bundle TS^n.* Take $\mathcal{M} = S^n$ with the round metric, $\mathcal{E} = TS^n$, and ∇ the Levi-Civita connection. Parallel transport of $w \in T_pS^n$ to T_qS^n along the minimizing geodesic is $P_{q\leftarrow p}w = w + (w\cdot\hat{u})\big((\cos\theta - 1)\hat{u} - \sin\theta\, p/R\big)$ with $\hat{u} = \mathrm{Log}_p(q)/d_{\mathcal{M}}(p,q)$, $\theta = d_{\mathcal{M}}(p,q)/R$, and the curvature is the sectional curvature $\kappa = 1/R^2$; this is the controlled setting of §6. TS^2 is topologically nontrivial: by the hairy-ball theorem it admits no global frame. This is the hypothesis of Proposition 1 and it distinguishes TS^2 from the parallelizable torus of §6. *(ii) SPD-valued fibre field.* Take fibres $\mathcal{E}_b = \mathrm{Sym}_r^{++}$ with the affine-invariant metric $d(A,B) = \|\log(A^{-1/2}BA^{-1/2})\|_F$ (Pennec et al., 2006; Arsigny et al., 2006; Said et al., 2017). The SPD cone is a Riemannian manifold, not a vector space; consequently, this experiment is an empirical manifold-fibre extension rather than an instance of the vector-bundle theorems. Its base grid is flat, so it tests intrinsic fibre aggregation only. We retain it as a secondary application and report its complete protocol in Appendix T.

H Discretized transport paths and misspecified connections

H.1 Path dependence: a holonomy error budget for discrete transport

Assumption 1(b) requires minimizing-geodesic transport; implementations rarely comply, and on a mesh one follows a Dijkstra path. Theorem 5 prices the difference, turning the invariant law from a verification target into an error budget.

Proposition 2 (Discretized-path transport bias; proof in Appendix O)**.** *Let $\gamma, \tilde{\gamma}$ share endpoints and bound a small homotopy surface S inside a normal neighborhood of diameter $O(\rho)$. For its oriented area bivector A, $P^{\tilde{\gamma}}_{b_0\leftarrow b} = HP^{\gamma}_{b_0\leftarrow b}$, where H is the loop holonomy based at b_0. Thus, for the radial model,*

$$\|\,\mathrm{Bias}_{\mathrm{path}}\,\| \le C\|F^{\nabla}\|\,\mathbb{E}\,\mathrm{area}(S)\,\|v^\star\| + O\big(\mathbb{E}[\rho\,\mathrm{area}(S) + \mathrm{area}(S)^2]\|v^\star\|\big)\,.$$

If the path tracks a geodesic of length $O(\rho)$ within an $O(h)$ ribbon, then $\mathrm{area}(S) = O(h\rho)$. *The resulting bound is $O(\|F^{\nabla}\| h\rho \|v^{\star}\|)$. This is an upper bound; cancellation can make the actual mean bias zero.*

The tracking hypothesis is additional to a mesh edge-length bound. A Dijkstra path on a restricted set of edge directions need not satisfy it. Consequently, an increasing imputation error with window size does not by itself identify discretization bias, and $h \gtrsim \rho$ compares upper-bound scales rather than proving which error dominates.

H.2 Misspecified connections

Proposition 3 (Connection misspecification; proof in Appendix O)**.** *For fixed connections $\widehat{\nabla} = \nabla + \delta$ in a common local gauge, assume $\|\delta\|_{C^1} \le \eta \le 1$. Under the radial signal model and conditionally zero-mean noise, the connection-induced mean error is $O(\eta\rho^2\|v^{\star}\|)$ for centered sampling and $O(\eta\rho\|v^{\star}\|)$ in general. A C^0 bound alone gives only the latter order.*

When a connection is learned from the same noisy observations, its dependence on the noise also requires analysis; these fixed-connection bounds do not automatically cover that case.

I Practical procedures for the diagnostics and design choices

This appendix answers five implementation questions that the theory raises, but the main text does not have room to settle: how to estimate Ξ from data, how to choose G_ϑ and the retraction, how to compute c_k off a homogeneous base, how to mitigate discretization bias, and how to choose references on a bundle with no global frame. Each procedure is stated so that it can be implemented and tested, and the corresponding measurements are reported in Appendices V.2, V.5 and in this appendix.

I.1 Estimating the transport-relevance ratio Ξ from data

The ratio $\Xi = |\kappa|\|v^{\star}\|/L_2$ uses a bound on the symmetric covariant Hessian. It is a leading-bias diagnostic for the spherical ambient-reference model. A practical estimator should report the following quantities separately.

Curvature and reference. On an appropriate refined surface mesh, angle defects divided by dual area can estimate Gaussian curvature; pointwise consistency needs mesh-quality assumptions. For a general connection, small-loop holonomy estimates the curvature operator only with a specified oriented area and small-loop limit. A single rotation angle cannot identify a general rank-r matrix-valued curvature two-form. For ambient references on the torus, estimate the shape operator as well; Gaussian curvature alone does not determine the reference bias.

Field derivatives. Fit transported observations in a common fibre against $(1, x_i, \frac{1}{2}x_i x_i^{\top})$. This estimates the symmetric covariant Hessian at the anchor, subject to a full-rank design, shrinking bandwidth, sufficient local samples, and controlled noise. It does not recover the antisymmetric curvature term. A median of fitted norms is a robust summary, not an upper bound on the supremum L_2 required by Corollary 2.

Signal and uncertainty. Under the radial model, $\|\hat{v}_{\mathrm{bundle}}\|^2 - r\hat{\sigma}_v^2/n$ is an unbiased signal-energy correction when the noise variance is known. Clipping at zero introduces bias. For general fields, base-sampling variance and smoothing bias also enter; subtracting only the noise term is insufficient.

Decision and validation. Fit diagnostics on training/validation fields and compare held-out losses for the competing estimators. Report bias alignment, effective sample size, transformed noise, and discretization error alongside $\widehat{\Xi}$. The condition $\Xi > 4$ concerns leading squared bias under isotropic spherical sampling; $\Xi < 1$ does not imply that the reference wins. Use independent field bootstrap or appropriate spatial blocks, not an independent-vertex bootstrap on correlated fields. Appendix V.2 validates this procedure on 105 controlled fields and Appendix V.5 on real winds; in both, $\widehat{\Xi}$ identifies the bias regime, and the plug-in estimate of the full identity chooses the better estimator (94% and 98%).

I.2 Sensitivity to the metric G_ϑ and to retraction details

Our guarantees do not depend on G: Theorems 1–2 need only the sufficient-ascent condition equation 3, which holds for any positive-definite G_ϑ obeying Assumption 1(d) with an Armijo line search. So G_ϑ cannot break ascent; it only changes conditioning, the trajectory and, in a nonconvex problem, which stationary point is reached. The same is true of the retraction: any first-order retraction on the parameter bundle Θ preserves the guarantees, and Exp is not required. The transport in equation 12 supplies an intrinsic horizontal/vertical construction. A chart-based retraction is also valid when its derivative is the identity on the represented tangent vector and the full objective is differentiated consistently; changing an anchor is not a gauge change.

Guidance, in increasing cost: identity-plus-Armijo (what our experiments use) is simplest and is usually adequate when the anchors are well separated; an observed-information preconditioner needs positive-definite damping away from a local maximum; the Riemannian/expected Fisher metric (natural gradient) provides reparametrization-invariance at one linear solve per step; for large r or K, diagonal or Kronecker-factored approximations (Martens & Grosse, 2015) may reduce cost; their conditioning benefit requires measurement. This mirrors the situation for Riemannian GMM optimization in the flat-data setting, where reformulating the problem on a matrix manifold gives substantial speed-ups over Euclidean EM (Hosseini & Sra, 2015), and Riemannian trust-region solvers are available as a second-order option (Absil et al., 2007); we expect the same to hold here, with the caveat that our parameter space is a bundle rather than a product, so a trust-region variant must respect equation 12.

Ablation. On the full Bundle-GEM task of Appendix V.1 (data straddling a pole, $\kappa = 1$, 50% missing; 10 data sets, 3 initializations each) we compared the default anchor step (direction β_k^2/N_k times the gradient, exponential-map retraction) with the identity preconditioner and with the first-order retraction $m \mapsto R(m+t)/\|m+t\|$. All three reach the same final log-likelihood on every data set (-213.34 on average), after the same number of iterations (98–99), and every initialization converged to the same stationary point. The identity preconditioner needs more backtracking and doubles the time per iteration (0.033 against 0.016 s); the chart retraction is as fast as the exponential map. The choice of G_ϑ and of the retraction therefore affects cost, not the result, in this setting.

I.3 Computing the normalizer c_k off a homogeneous base

The integration domain is part of the model. On the fixed region U used in equation 1,

$$c_k(m,\beta) = \int_U \exp[-d_{\mathcal{M}}(b,m)^2/(2\beta^2)]\, d\mathrm{vol}_{\mathcal{M}}(b).$$

Homogeneity removes anchor dependence on the full homogeneous manifold, but not generally on a fixed truncated region. Homogeneity does not imply a closed form. On a smooth normal region,

$$\mathrm{grad}_m \log c_k(m,\beta) = \beta^{-2}\mathbb{E}_{b\sim\nu_{m,\beta}}[\mathrm{Log}_m(b)],$$

where $\nu_{m,\beta}$ is the normalized base kernel on U. The fibre likelihood also depends on the anchor through transport. Both contributions belong in the full anchor derivative.

Exact evaluation supports the original ascent theorem. *Shared quadrature* for values and derivatives, or a differentiable precomputed approximation, can provide a consistent approximate objective. Ascent of that approximate objective does not imply ascent of the exact likelihood; retain a known normalization for the discretized model or bound integration error and accept only steps whose exact-Q improvement is certified. Arbitrarily interpolating a normalizer can even produce an unnormalized continuous density. A *frozen normalizer* likewise provides no exact ascent guarantee. Report the approximation, error tolerance, and objective monitored.

I.4 Mitigating discretization bias: symmetrization and loop cancellation

Proposition 2 prices a non-minimizing path by the holonomy of the loop it encloses with the true geodesic, giving an $O(\|F^\nabla\|h\rho\|v^\star\|)$ upper bound under the tracking assumption. The same identity suggests three mitigations, in increasing cost.

(i) Multipath symmetrization. Transport along M distinct paths and average the results in the reference fibre. If the enclosed-area bivectors $A_{(m)}$ have $\sum_m A_{(m)} \approx 0$, which holds when the paths

straddle the geodesic instead of all deviating to one side, then by Theorem 5 the leading $-\frac{1}{2}F^{\nabla}_{jk}A^{jk}$ terms cancel to first order and the residual is $O(|A|^{3/2})$. This is the cheapest useful fix and requires no new geometry. *(ii) Loop cancellation as a diagnostic.* If a reference path is available, transport around the closed loop and measure its holonomy. This measures the transport discrepancy; it does not recover an arbitrary surface area or curvature field from one loop. This converts Proposition 2 from a bound into a per-edge correction, since the measured holonomy can be inverted out of the transported value. *(iii) Geodesic correction.* Replace Dijkstra predecessor paths by straightest-geodesic tracing (unfolding across triangle fans); this attacks the $O(h)$ term at its source and should dominate the other two when affordable.

Measurement. On the unit sphere, transport across a mesh edge by the minimal rotation of normals is exact transport along the great-circle edge, so the rotation between transport along a Dijkstra path and along the minimizing geodesic equals the signed area the two paths enclose. On icospheres with spacing $h = 0.151, 0.076, 0.038$ and radius $\rho = 0.4$, the mean rotation is 0.0058, 0.0054 and 0.0051 rad, and at $\rho = 0.2$ it is about 0.0014 on the two finer meshes: it scales like $0.035\rho^2$ and does not shrink with h. Shortest paths on a graph with a few edge directions do not track the geodesic within $O(h)$, so the $O(h\rho)$ rate of Proposition 2 does not apply to them, as the tracking hypothesis warns; the enclosed-area bound itself holds. Averaging the vectors transported along two different shortest paths (unperturbed and randomly perturbed edge weights) lowers the rotation by 33–42%, as the symmetrization argument predicts.

I.5 References on bundles with no global frame: patching and adaptive choices

Proposition 1 rules out a globally-defined continuous cocycle family $\{\tau\}$ on a nontrivial bundle such as TS^2; it does not rule out useful local references, and the practical question is how to patch them without either a singularity or a large bias.

Atlas with partition of unity. Cover $\mathcal{M}$ with charts $\{U_\alpha\}$ each admitting a frame, and aggregate within charts, combining across overlaps with a smooth partition of unity $\{\varphi_\alpha\}$. Because the transition maps on overlaps are exactly the bundle's cocycle, the combination is well defined; the cost is that the effective reference on an overlap is a φ-weighted blend, whose single-path bias is the corresponding weighted comparison after all outputs are expressed in the same target fibre. No derivative-of-partition term appears solely from this pointwise blending; such terms arise when differentiating the blended field. Smooth, adequately sampled overlaps help numerical stability; they do not, by themselves, guarantee low risk.

Anchor-local references. A cheaper alternative avoids an atlas entirely: use a reference defined only relative to the current aggregation anchor, such as the normal-coordinate frame τ^{nc} at b_0, and never attempt global consistency. This is well defined on every geodesic ball inside the injectivity radius, needs no patching, and by Proposition 6 has a leading isotropic bias magnitude three times smaller than ambient projection on TS^2, with different transformed noise and computational requirements. Its limitation is that values referred to different anchors are not directly comparable, which is acceptable when the downstream quantity is itself anchor-local (as in our imputation task) and unacceptable when a single global field is required.

Adaptive choice. Combining this with Appendix I.1: estimate $\widehat{\Xi}$ per region and use the diagnostics to shortlist estimators and choose their settings by held-out risk. Since Corollary 1 makes the crossover explicit in n, σ_v, σ_b, the same rule can be made sample-size aware, switching to the shrinking reference in sparsely observed regions where $\kappa < \kappa^\star$. Appendix V.5 applies such a rule vertex by vertex on real winds: the plug-in identity estimated on training months chooses between transport and the ambient reference correctly at 97.7% of the locations on held-out months.

J Extended related work

Positioning table. Table 10 compares Bundle-GEM with the works closest to it, by what each assumes about the base, the fibre, transport and the objective.

Table 10: Positioning by base manifold, fibre, transport, and objective. Entries describe the typical formulation of the cited work; individual methods may combine several paradigms.

Work	Base $\mathcal{M}$	Fibre	Transported?	Incomplete-data likelihood	Gauge equiv.	Main guarantee
Standard EM (Dempster et al., 1977)	any	model dependent	n/a	✓	coordinate independent	likelihood ascent
Manifold statistics (Karcher, 1977; Pennec et al., 2006)	any Riemannian	— (base itself)	n/a	not usually	no	estimator asymptotics
Das & Snášel (2026)	any Riemannian	vector bundle	✓	no	intrinsic statistics	concentration + minimax floor
Sheaf/connection learning (Hansen & Gebhart, 2020; Bodnar et al., 2022)	graph	learned stalks	restriction maps	not usually	model dependent	diffusion/classification objective
Fréchet regression (Petersen & Müller, 2019)	Euclidean covariates	metric-space response	n/a	not usually	no	local-averaging rates
Gauge-equivariant networks (Cohen et al., 2019; de Haan et al., 2021; Park et al., 2023)	mesh/manifold	associated bundle	✓	not usually	✓	equivariant prediction
Vector-valued geometric GPs (Hutchinson et al., 2021)	Riemannian	vector bundle	✓ (kernel)	variational/marginal	kernel dependent	probabilistic field prior
Riemannian flows/diffusions (Mathieu & Nickel, 2020; De Bortoli et al., 2022)	Riemannian	— (base itself)	n/a	no	no	density on manifold
Bundle-GEM (ours)	any Riemannian	vector bundle	✓	✓	optional encoder	ascent + scoped MSE law

Relation to gauge-equivariant and mesh CNNs. Gauge-equivariant CNNs (Cohen et al., 2019) and gauge-equivariant mesh CNNs (de Haan et al., 2021; Park et al., 2023) already perform the transport step we analyze: features live in local gauges and are carried along edges by the connection before being mixed. We do not propose a competing architecture. What we add to that line are criteria for when the transport step is worth its cost: the curvature threshold $\kappa^\star$ (Cor. 1), the transport-relevance ratio Ξ (Cor. 2), and the discretization budget of Proposition 2. The last of these is of direct architectural relevance: it predicts an $O(|\kappa|h\rho)$ bias that grows with the aggregation window and does not average away with more data, which is a possible mechanism for triangulation sensitivity under the path assumptions, and it identifies multipath symmetrization and geodesic-corrected paths (Appendix I.4) as remedies. In the same spirit, our holonomy analysis complements discrete-connection and sheaf constructions on surfaces and graphs (Crane et al., 2010; Knöppel et al., 2013; Hansen & Gebhart, 2020; Bodnar et al., 2022) by making the bias attributable to loop enclosures and non-minimizing paths explicit rather than implicit in the discretization.

Relation to vector-valued geometric GPs, and a possible bridge. Gauge-independent vector-valued GPs (Hutchinson et al., 2021) and geometric diffusion processes (Mathieu et al., 2023) share our insistence on treating fibre values gauge-independently, but target Bayesian field priors rather than likelihood-based inference from coarsened observations, and do not price the finite-sample MSE trade-off that Theorem 4 makes exact. The two are complementary, and the combination looks fruitful in two directions. First, the observed likelihood ℓ_{obs} of §3 is a valid likelihood term for bundle-valued observations and could sit inside a Bayesian model whose prior is such a GP, giving a likelihood-based treatment of missingness that a pure field prior does not supply. Second, and more interestingly for the theory here, a GP prior with a controlled smoothness (Matérn-type) directly bounds $\|\nabla^2 V\|$ in expectation and therefore regularizes Ξ: choosing the prior's smoothness parameter is, through equation 8, affecting one ingredient of the bias comparison; it does not by itself determine the risk ranking.

Relation to concentration and minimax results for bundle-valued means. Recent work establishes non-asymptotic concentration for transported empirical means and holonomy-driven error floors (Das & Snášel, 2026). That literature and ours are complementary: it characterizes how fast a transported mean concentrates, whereas we study the split between the reference-dependent single-path bias (Prop. 7, gauge covariant for a fixed reference, and of opposite sign for two standard references by Prop. 6) and the gauge-invariant loop term (Thm. 5), together with the observation that a non-isometric reference contracts noise and can therefore beat an unbiased transported estimator outright below $\kappa^\star$, a shrinkage effect that a concentration bound on the transported estimator alone does not reveal. The two crossovers also answer different questions. Theirs, which equates the stochastic term with the holonomy floor, asks when more data stops helping, whereas $\kappa^\star$ (Cor. 1) asks when transport stops beating a fixed reference. Theirs is intrinsic to transport and compares two isometric transport rules, so no variance-reducing competitor can arise; ours compares transport against an alternative identification, and only the latter admits one.

Relation to Riemannian optimization for mixture models. Riemannian reformulations of GMM fitting (Hosseini & Sra, 2015), together with Riemannian first- and second-order solvers (Absil et al., 2007; 2008; Boumal, 2023), demonstrate that geometric optimization can substantially outperform Euclidean EM, but for Euclidean data, where the geometry enters on the parameter side (positive-definite covariances) and not through the data itself. Our connection to that line is that, because our observations are bundle-valued, the mixture parameters (μ_k, Σ_k) live in the fibre over the anchor m_k, so Θ is a bundle rather than a product, and the update must transport the fibre parameters as the

anchor moves (Lemma 1). A second-order trust-region variant of Bundle-GEM inherits our guarantees provided its step respects equation 12; Appendix I.2 reports the preconditioner and retraction ablation.

EM and coarsened data. EM and its asymptotics (Dempster et al., 1977; Wu, 1983; Louis, 1982; McLachlan & Krishnan, 2008; Balakrishnan et al., 2017; Dwivedi et al., 2020) underpin our ascent and information results; random-set and interval-censored likelihoods (Matheron, 1975; Molchanov, 2005; Heitjan & Rubin, 1991) are the $K = \mathbf{1}\{z \in Y\}$ case. **Intrinsic statistics.** Fréchet-mean asymptotics (Bhattacharya & Patrangenaru, 2003; Karcher, 1977), Riemannian tensor computing and learning (Pennec et al., 2006; Arsigny et al., 2006; Fletcher et al., 2004; Brooks et al., 2019; Chakraborty et al., 2022), and intrinsic Cramér–Rao bounds (Smith, 2005; Oller & Corcuera, 1995) develop inference on a single nonlinear space. We instead study sufficient statistics whose values occupy different fibres and must be compared through a connection. Recent Bernoulli results sharpen the statistical picture for non-Euclidean means and robust estimators (Eltzner, 2022; Yun & Park, 2023; Jung et al., 2025). **Bundle-valued statistics.** Closest to us in statistical theory, Das & Snášel (2026) establish dimension-free Hoeffding- and Bernstein-type concentration for transported empirical means, isolate a deterministic holonomy floor arising when minimizing geodesics are non-unique so that two transport rules disagree, and prove minimax lower bounds showing both the $n^{-1/2}$ term and that floor are unavoidable; they further give a median-of-means estimator for heavy tails and a CLT in the reference fibre. Their frequentist mean-estimation theory overlaps with our geometric motivation, and their holonomy term is the gauge-invariant object we recover in Theorem 5. Our distinct focus is likelihood inference under coarsening, a concrete transported mixture with its own optimization theory, and the bias of a reference identification substituted for transport, which has no counterpart there; which is reference-dependent and need not be an isometry. Pennec (2019) quantifies curvature effects on the empirical Fréchet mean on the base manifold; our reference-dependent fibre error and invariant loop result are different objects.

Sheaf and connection learning; Lie-group uncertainty; Fréchet regression. Sheaf neural networks (Hansen & Gebhart, 2020; Bodnar et al., 2022) learn restriction maps between stalks and are the closest architectural relative of Bundle-GEM, though they optimize a diffusion objective and learn the maps; filtering on matrix Lie groups (Barfoot & Furgale, 2014; Bourmaud et al., 2015) does the same covariance-transport bookkeeping as equation 12 in a recursive setting; and Fréchet regression (Petersen & Müller, 2019) is the natural comparison for our base-dependent means, with Theorem 6 a bundle-valued analogue of its local-averaging rates. Discrete connections on surfaces (Crane et al., 2010; Knöppel et al., 2013) underlie our torus construction and formalize the defect Proposition 2 prices. **Information geometry** (Amari, 2016) curves the parameter space; we curve the latent/observation space and its fibres.

K Ascent, convergence, identifiability, consistency, and rate

Proposition 4 (Local linear rate)**.** *Assume a C^2 EM map near an interior strict local maximum $\theta^\star$, a locally unique exact M-step, and $\mathcal{I}_{\mathrm{com}} \succ 0$, $\mathcal{I}_{\mathrm{obs}} = \mathcal{I}_{\mathrm{com}} - \mathcal{I}_{\mathrm{mis}} \succ 0$. The exact Bundle-EM operator converges linearly in the Riemannian distance with asymptotic rate $r = \rho\big(\mathcal{I}_{\mathrm{mis}}(\theta^\star)\,\mathcal{I}_{\mathrm{com}}(\theta^\star)^{-1}\big)$, the spectral radius of the bundle missing-information fraction.*

This is the classical "fraction of missing information governs EM's rate" result (Dempster et al., 1977; McLachlan & Krishnan, 2008) expressed in a normal chart. This chart is a computational convenience and does not introduce a reference choice of the kind analyzed in Section 4: because $\theta^\star$ is a critical point of $Q(\cdot \mid \theta^\star)$, its Hessian is coordinate-independent there (the connection term that otherwise distinguishes the ordinary from the covariant Hessian is proportional to the vanishing gradient), so $\mathcal{I}_{\mathrm{com}}(\theta^\star)$, $\mathcal{I}_{\mathrm{mis}}(\theta^\star)$, and hence the spectral radius r are the same in any chart. The geometry enters $\mathcal{I}_{\mathrm{mis}}$ through the intrinsic E-step, but, unlike the bias analyzed below, the rate is a standard missing-information quantity with no separate curvature term.

The transported retraction of §3 in full:

$$m_k^{(t+1)} = \mathrm{Exp}_{m_k^{(t)}}(\eta_t \xi_k), \qquad \mu_k^{(t+1)} = P_k\big(\mu_k^{(t)} + \eta_t \zeta_k\big), \qquad \Sigma_k^{(t+1)} = P_k\,\widetilde{\Sigma}_k\,P_k^{-1}, \quad \text{where } P_k := P_{m_k^{(t+1)} \leftarrow m_k^{(t)}}, \tag{12}$$

Table 11: Guarantees at a glance. Assumptions are stated in full in Appendix D (Assumption 1) and Appendix N (Assumption 2); this table is a map to them.

Result	Key assumptions	Guarantees
Thm. 1 (ascent)	positive likelihood, posterior-weighted integrability	monotone ascent of ℓ_{obs}
Thm. 2 (convergence)	Assumption 1: Karcher–Kendall ball, compact superlevel set, Lipschitz gradients, Armijo	finite limit, stationary limit points
Prop. 5 (identifiability)	known common anchor and base density; distinct Gaussian fibre components	identifiable mod label permutations
Thm. 3 (consistency)	identifiability + compact Θ + integrable envelope, ψ fixed	$\hat{\theta}_n \to \theta^\star$ mod labels, in probability
Prop. 4 (local rate)	locally unique exact M-step, C^2, $\mathcal{I}_{\mathrm{com}} \succ 0$, $\mathcal{I}_{\mathrm{obs}} \succ 0$ at an interior local maximum	linear rate $\rho(\mathcal{I}_{\mathrm{mis}}\mathcal{I}_{\mathrm{com}}^{-1})$
Thm. 4 (MSE identity)	isotropic noise, bounded linear reference	exact MSE gap, incl. transformed-noise term
Cor. 1 (crossover)	ambient reference, isotropic Σ_b, TS^2	leading threshold approximation $\kappa^\star$
Thm. 6 (radius tradeoff)	C^3 field, bounded moments of x, ε	risk bound; leading-bound optimizer $\rho^\star$
Prop. 2 (discretization)	small homotopy surface; $O(h)$ path tracking for mesh order	bias bound in enclosed area
Prop. 3 (misspecification)	fixed C^1 connection perturbation bounded by η	bias bound, centred vs. boundary sampling
Prop. 1 (topology)	continuous invertible maps, identity and cocycle	existence/non-existence of global baseline

Lemma 1 (Transported retraction on the parameter bundle)**.** *Use the connection to split a tangent vector at (m, μ, Σ) into a horizontal component ξ and vertical components (ζ, H). Let R^{Σ}_{Σ} be a first-order retraction on the positive-definite covariance cone. Then*

$$m(\eta) = \mathrm{Exp}_m(\eta\xi), \quad \mu(\eta) = P_{m(\eta)\leftarrow m}(\mu + \eta\zeta), \quad \Sigma(\eta) = P_{m(\eta)\leftarrow m} R^{\Sigma}_{\Sigma}(\eta H) P_{m\leftarrow m(\eta)}$$

defines a first-order retraction on the open parameter bundle (with ordinary feasible retractions for the other parameters). It does not preserve Q when the anchor changes. Ascent follows from evaluating the full objective along a gradient-related direction and applying line search. The proof is in Appendix N.

Theorem 1 (Monotone ascent)**.** *Assume $p_{\theta,\psi}(Y_i) \in (0, \infty)$ and $\log\{f_\theta(z)\mathcal{K}_\psi(Y_i \mid z)\}$ is integrable against the current posterior, with common posterior support and finite KL terms for the compared parameters. If the M-step satisfies $Q(\vartheta_{t+1} \mid \vartheta_t) \geq Q(\vartheta_t \mid \vartheta_t)$, then $\ell_{\mathrm{obs}}(\vartheta_{t+1}) \geq \ell_{\mathrm{obs}}(\vartheta_t)$, with equality iff the Q-increment is zero and $q_i^{(t+1)} = q_i^{(t)}$ a.e. for every i.*

Theorem 2 (Convergence to stationary points)**.** *Under Assumption 1, Bundle-GEM satisfies: (i) $\ell_{\mathrm{obs}}(\theta^{(t)}, \psi^{(t)})$ increases monotonically to a finite limit; (ii) every limit point is a Riemannian stationary point of ℓ_{obs}; and (iii) successive steps vanish.*

Consistency requires identifiability of the observed model. A normalized observation kernel need not preserve latent identifiability: a kernel independent of z, for example, discards all parameter information. The following proposition establishes only a fixed-anchor latent special case.

Proposition 5 (Identifiability of a fixed-anchor transported mixture, modulo labels)**.** *Fix a known anchor m, a known positive normalized base density $h(b)$, and a known connection/path rule on a normal region. Consider*

$$f_\theta(b, v) = h(b) \sum_{k=1}^{K} \pi_k \varphi(P_{m\leftarrow b} v; \mu_k, \Sigma_k).$$

For fixed K, positive weights, positive-definite covariances, and distinct Gaussian component pairs (μ_k, Σ_k), equality of latent densities implies equality of parameters modulo permutation. Moving

unknown anchors in equation 1 requires an additional identifiability argument; smoothness, convexity, and distinct modes alone do not prove linear independence of component laws.

To transfer this result to observations, additionally require that the observation operator $f \mapsto \int \mathcal{K}_{\psi^\star}(Y \mid z) f(z) d\mu_{\mathcal{E}}(z)$ is injective on the model family, or directly assume observed-model identifiability. Neither follows from normalization alone.

Theorem 3 (Consistency of the constrained likelihood target)**.** *Assume i.i.d. observations, a known and fixed observation parameter $\psi = \psi^\star$, a compact constrained parameter space Θ, continuity of $\log p_\theta(Y)$, an integrable uniform envelope, and a population maximizer unique modulo the $K!$ mixture-label permutations. Then any sequence of global observed maximizers $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell_{\mathrm{obs},n}(\theta)$ satisfies*

$$d_{\mathrm{perm}}(\hat{\theta}_n, \theta^\star) := \min_{\sigma \in \mathfrak{S}_K} d_\Theta(\hat{\theta}_n, \sigma\cdot\theta^\star) \xrightarrow{p} 0. \tag{13}$$

These are coordinate-free applications of classical EM and M-estimation theory (Wu, 1983; McLachlan & Krishnan, 2008; van der Vaart, 1998); the bundle-specific content is the realization of the sufficient statistics. Two scope notes: the consistency result concerns a constrained global maximizer, not convergence of a nonconvex GEM run to it; and Algorithm 1 updates ψ whereas Theorem 3 fixes it at $\psi^\star$, the joint case requiring kernel identifiability that we do not establish. Under the additional local regularity and positive-information assumptions of Proposition 4, Bundle-EM converges linearly at rate $\rho(\mathcal{I}_{\mathrm{mis}}\mathcal{I}_{\mathrm{com}}^{-1})$, the classical missing-information fraction (Proposition 4, Appendix N).

L Single-path expansion, aggregation bias, and the MSE identity

Lemma 2 (Single-path transport discrepancy)**.** *Let $\{\tau_{b_0\leftarrow b}\}_{b\in U}$ be a smooth family of linear maps $\mathcal{E}_b \to \mathcal{E}_{b_0}$ on a normal ball $U = B_\rho(b_0)$ with $\tau_{b_0\leftarrow b_0} = \mathrm{Id}$, and set $\Phi(x) := \tau_{b_0\leftarrow b} P^\nabla_{b\leftarrow b_0} \in \mathrm{End}\,\mathcal{E}_{b_0}$ for $x = \mathrm{Log}_{b_0}(b)$. Assume $D\Phi(0) = 0$ (first-order alignment). Then*

$$\Phi(x) = \mathrm{Id} + A_2(x) + R(x), \qquad A_2(x) = \tfrac{1}{2} D^2\Phi(0)[x,x], \qquad \|R(x)\| \le C\,\|\Phi\|_{C^3(U)}\,\|x\|^3. \tag{14}$$

(a) The curvature term is absent. *A_2 is a symmetric quadratic form in x, so any antisymmetric contribution drops out; in particular, when τ is the transport of a flat reference connection ∇^0 with $a := \nabla - \nabla^0$ and $a(b_0) = 0$, one has $A_2(x) = -\frac{1}{2}\partial_{(k}a_{j)}(b_0)x^j x^k$ and the antisymmetric part $F^\nabla_{jk} = \partial_j a_k - \partial_k a_j$ satisfies $F^\nabla_{jk}x^j x^k \equiv 0$.* (b) The coefficient is reference dependent, not canonical. *A_2 has no value independent of the chosen τ; Proposition 6 exhibits two standard choices on TS^2 whose A_2 differ in sign.* (c) Scope. *We state the lemma for a general smooth family rather than only for ∇^0-transport because an ambient-projection pairwise family need not be a cocycle. A local frame does define a flat connection on its chart; a global frame requires triviality (Proposition 1).*

Proposition 6 (Two standard references on TS^2, opposite-sign bias)**.** *Let $\mathcal{M} = S^2$ with $\kappa = 1/R^2$, $\mathcal{E} = TS^2$, ∇ Levi-Civita, and $x = \mathrm{Log}_{b_0}(b)$.*

- *(i)* Ambient projection $\tau^{\mathrm{amb}}_{b_0\leftarrow b} = \Pi_{T_{b_0}S^2}|_{T_bS^2}$ *gives* $\Phi_{\mathrm{amb}}(x) = I - \frac{1}{2}\kappa\,(x\otimes x) + O(\|x\|^4)$ *and hence* $\mathrm{Bias}^{\mathrm{amb}}_{\mathrm{ref}} = -\frac{1}{2}\kappa\,\Sigma_b v^\star + O(\rho^4)$.
- *(ii)* Normal-coordinate frame τ^{nc}, *which preserves coefficients in the frame $d(\mathrm{Exp}_{b_0})_x$, gives* $\Phi_{\mathrm{nc}}(x) = I + \frac{\kappa}{6}\big(\|x\|^2 I - x\otimes x\big) + O(\|x\|^4)$ *and hence* $\mathrm{Bias}^{\mathrm{nc}}_{\mathrm{ref}} = +\frac{\kappa}{6}\big(\,\mathrm{tr}\,\Sigma_b\, I - \Sigma_b\big)v^\star + O(\rho^4)$.

For isotropic $\Sigma_b = \sigma_b^2 I_2$ these are $-\frac{1}{2}\kappa\sigma_b^2 v^\star$ and $+\frac{1}{6}\kappa\sigma_b^2 v^\star$: opposite in sign and different by a factor of three in magnitude. *The released code includes a script that checks both expansions numerically.*

Proposition 7 (Reference-dependent aggregation bias)**.** *Let observations at $b_i = \mathrm{Exp}_{b_0}(x_i)$ carry $y_i = P_{b_i\leftarrow b_0} v^\star + \varepsilon_i$ with $\mathbb{E}[x_i] = 0$, $\mathrm{Cov}(x_i) = \Sigma_b$, $\|x_i\| \le \rho$, and $\mathbb{E}[\varepsilon_i \mid x_i] = 0$. The transport-aware estimator $\hat{v}_{\mathrm{bundle}} = \frac{1}{n}\sum_i P_{b_0\leftarrow b_i} y_i$ is unbiased. The reference-based estimator $\hat{v}_{\mathrm{prod}} = \frac{1}{n}\sum_i \tau_{b_0\leftarrow b_i} y_i$ has leading bias*

$$\mathrm{Bias}_{\mathrm{ref}} := \mathbb{E}[\hat{v}_{\mathrm{prod}}] - v^\star = -\tfrac{1}{2}\partial_{(k}a_{j)}(b_0)\Sigma_b^{jk} v^\star + O(\rho^3). \tag{15}$$

For the tangent bundle of a constant-curvature surface under the stated ambient reference $\tau_{b_0\leftarrow b} = \Pi_{T_{b_0}S^2}|_{T_bS^2}$, equivalently ambient averaging followed by projection onto $T_{b_0}S^2$, $\mathrm{Bias}_{\mathrm{ref}} = -\frac{1}{2}\kappa\Sigma_b v^\star + O(\rho^3)$ and therefore $\|\mathrm{Bias}_{\mathrm{ref}}\| = \frac{1}{2}|\kappa|\|\Sigma_b v^\star\| + O(\rho^3) \le \frac{1}{2}|\kappa|\|\Sigma_b\|_{\mathrm{op}}\|v^\star\| + O(\rho^3)$.

Theorem 4 (MSE decomposition under a linear reference)**.** *Assume i.i.d. base observations under Proposition 7, conditionally zero-mean noise with covariance $\sigma_v^2 I$ in each fibre, independent across observations, and parallel transport is isometric, and $\tau_{b_0\leftarrow b}$ is a bounded linear identification. Define $\Delta_{\mathrm{noise}} = \frac{\sigma_v^2}{n}\,\mathrm{tr}\{\mathbb{E}[\tau_{b_0\leftarrow b}\tau_{b_0\leftarrow b}^*] - I\}$. Then*

$$\mathrm{MSE}(\hat{v}_{\mathrm{prod}}) - \mathrm{MSE}(\hat{v}_{\mathrm{bundle}}) = \Delta_{\mathrm{noise}} + \frac{1}{n}\,\mathrm{tr}\,\mathrm{Cov}(\Phi(x)v^\star) + \|\mathrm{Bias}_{\mathrm{ref}}\|^2, \tag{16}$$

an exact identity. If τ is fibrewise isometric, $\Delta_{\mathrm{noise}} = 0$ and the gap is nonnegative. Under the isometric specialization,

$$\mathrm{MSE}_{\mathrm{prod}} - \mathrm{MSE}_{\mathrm{bundle}} = \frac{1}{n}\,\mathrm{tr}\,\mathrm{Cov}(A_2(x)v^\star) + \|\mathbb{E}[A_2(x)]v^\star\|^2 + O(\rho^5). \tag{17}$$

The gap is strict precisely when $\Phi(x)v^\star$ is nonconstant with positive probability, or its mean differs from $v^\star$; nonzero curvature alone is neither necessary nor sufficient for strictness. For a non-isometric reference such as ambient projection, the noise correction must be retained and the total MSE difference need not be nonnegative.

General fields and weighted learning. For $U(x) = P_{b_0\leftarrow b}V(b)$, $B = \mathbb{E}U - v^\star$, and $D = \mathbb{E}[(\Phi - I)U]$, the exact general-field identity is

$$\mathrm{MSE}_{\mathrm{prod}} - \mathrm{MSE}_{\mathrm{bundle}} = \Delta_{\mathrm{noise}} + \frac{\mathrm{tr}\,\mathrm{Cov}(\Phi U) - \mathrm{tr}\,\mathrm{Cov}(U)}{n} + 2\langle B, D\rangle + \|D\|^2. \tag{18}$$

It follows from the same conditional bias–variance calculation. Consequently, a field-bias ratio does not determine full risk: variance and bias alignment remain relevant. For fixed locations and deterministic normalized weights w_i, the radial model instead gives the conditional gap

$$\left\|\sum_i w_i(\Phi_i - I)v^\star\right\|^2 + \sigma_v^2\sum_i w_i^2\{\mathrm{tr}(\tau_i\tau_i^*) - r\}.$$

Posterior responsibilities or robust weights estimated from the same noisy features are not deterministic, noise-independent weights. The simple-mean identity cannot automatically be applied to the full nonlinear Bundle-GEM estimator; that extension needs its own risk analysis.

Lemma 3 (Covariant Taylor expansion along a geodesic)**.** *Along the geodesic $\gamma(s) = \mathrm{Exp}_{b_0}(sx)$,*

$$P_{b_0\leftarrow b}V(b) = V(b_0) + (\nabla_x V)(b_0) + \tfrac{1}{2}(\nabla_{x,x}^2 V)(b_0) + R_3(x), \qquad \|R_3(x)\| \le \tfrac{1}{6}L_3\|x\|^3. \tag{19}$$

M E-STEP FRÉCHET AND TRANSPORTED FIBRE STATISTICS

With latent component C_i, the E-step posterior $q_{ik}^{(t)}(b, v) = p_{\theta^{(t)},\psi^{(t)}}(b, v, C_i = k \mid Y_i)$ yields

$$N_k = \sum_i \int q_{ik}, \quad s_k(m_k) = \sum_i \int q_{ik}P_{m_k\leftarrow b}v, \quad T_k(m_k) = \sum_i \int q_{ik}ww^\top, \tag{20}$$

where $w = P_{m_k\leftarrow b}v$ and every integral is with respect to $\mu_{\mathcal{E}}$. For fixed m_k, the M-step has $\mu_k = s_k/N_k$, $\Sigma_k = T_k/N_k - \mu_k\mu_k^\top$, and $\pi_k = N_k/n$; m_k and β_k are updated by retraction ascent on the same Q-function. These formulas apply when the constraints are inactive. With only spectral bounds, diagonalize the empirical covariance and clip its eigenvalues to $[\lambda_{\min}, \lambda_{\max}]$. For lower-bounded weights with $0 < \pi_{\min} < 1/K$, solve $\pi_k = \max\{\pi_{\min}, N_k/\lambda\}$ with $\sum_k \pi_k = 1$; simply using N_k/n can violate feasibility. Other constraints require their own feasible maximizer or a sufficient-ascent step. Recompute $s_k(m), T_k(m)$ or differentiate their anchor dependence throughout an anchor line search while holding the E-step posterior fixed. Thus the transported moments determine the update rather than serving only as posterior summaries.

For component k of equation 1, a Fréchet base statistic and its transported fibre statistic are

$$\bar{b}_k = \arg\min_{b\in\mathcal{M}} \sum_i \int q_{ik}^{(t)}(z) d_{\mathcal{M}}(b, \pi z)^2 d\mu_{\mathcal{E}}(z), \qquad \bar{v}_k = \frac{1}{N_k} \sum_i \int q_{ik}^{(t)}(z) P_{\bar{b}_k \leftarrow \pi z} v(z) d\mu_{\mathcal{E}}(z). \tag{21}$$

a base Fréchet mean (Karcher, 1977; Afsari, 2011) and a transported fibre mean. For a decoupled base density with anchor-independent normalizer, $\bar{b}_k$ is the base M-step; for equation 1, it initializes a joint base/fibre update.

N Detailed mathematical proofs

Throughout, $(\mathcal{M}, g_{\mathcal{M}})$ is a complete smooth Riemannian d-manifold and $\pi : \mathcal{E} \to \mathcal{M}$ a smooth rank-r vector bundle with fibre metric $g^{\mathcal{E}}$ and metric-compatible connection ∇; $\mu_{\mathcal{E}}$ is the reference measure of §2. We write $\|\cdot\|$ for the relevant Riemannian norms and $C, C_1, \ldots$ for finite constants depending only on the quantities named in the hypotheses. All statements are made under Assumption 1 together with the following analytic conditions, which must be checked for the selected densities, kernels, parameter constraints, and observation design of §2.

Assumption 2 (Analytic regularity)**.** *(R1) Positivity and smoothness.* For every Y, the map $z \mapsto \mathcal{K}_\psi(Y \mid z) f_\theta(z)$ is strictly positive and $\mu_{\mathcal{E}}$-integrable with $0 < p_{\theta,\psi}(Y) < \infty$, and $(\theta, z) \mapsto \log f_\theta(z)$ and $(\psi, z) \mapsto \log \mathcal{K}_\psi(Y \mid z)$ are C^3. *(R2) Domination.* On each compact parameter neighborhood, the joint densities and their parameter derivatives through order three have integrable dominating functions sufficient for differentiation under the latent integral; posterior-weighted log densities and their derivatives are integrable. Also $\sup_\theta |\log p_\theta(Y)| \le H(Y)$ for a $P_{\theta^\star}$-integrable envelope H. Requiring the unweighted Gaussian log-density itself to have a Lebesgue-integrable envelope on an unbounded fibre would be incorrect. *(R3) Local nondegeneracy.* For the local exact-EM rate only, require $\mathcal{I}_{\mathrm{com}}(\theta^\star) \succ 0$ and $\mathcal{I}_{\mathrm{obs}}(\theta^\star) \succ 0$ at the specified interior local maximum. These conditions need not hold globally or at coincident mixture components. *(R4) Bounded geometry.* On the geodesic balls of Assumption 1(a), the quantities $\|F^\nabla\|, \|\nabla F^\nabla\|, \|R_{\mathcal{M}}\|$ and $1/\mathrm{inj}(\mathcal{M})$ are bounded, uniformly over the base points visited. *(R5) Moments.* Each posterior $q_i^{(t)}$ has $\mathbb{E}_{q_i}\|\delta Z\|^5 < \infty$, where, around the posterior mode $z_0 = (b_0, v_0)$, $\delta Z = (\delta b, \delta v)$ with $\delta b = \mathrm{Log}_{b_0}(\pi z)$ and $\delta v = P_{b_0 \leftarrow \pi z} v(z) - v_0$.

N.1 Second-order transport: why curvature does not appear on a single path

One might expect the connection curvature F^∇ to contribute a second-order term to the discrepancy between ∇-transport and the flat identification used by a product-space method along a single radial path. It does not, because the curvature is antisymmetric and the radial displacement product is symmetric, so their contraction vanishes. We make the correct statement precise and then give the invariant object, the loop holonomy.

Proof of Lemma 2. The difference $a = \nabla - \nabla^0$ is tensorial and C^2; write $a = a_j\, dx^j$, $a_j(b_0) = 0$. The relative transport is the value at $s = 1$ of $\Phi'(s) = -a(\gamma'(s))\Phi(s)$, $\Phi(0) = \mathrm{Id}$, along $\gamma(s) = \mathrm{Exp}_{b_0}(sx)$. Its Peano series is $\Phi(1) = \mathrm{Id} - \int_0^1 a(\gamma'(s))\, ds + \int_0^1\int_0^s a(\gamma'(s))a(\gamma'(u))\, du\, ds - \cdots$. In normal coordinates $\gamma'(s) = x + O(s\|x\|^2)$ and $a_j(\gamma(s)) = \partial_k a_j(b_0)\, sx^k + O(s^2\|x\|^2)$, so the first Peano term is $-\partial_k a_j(b_0) x^k x^j \int_0^1 s\, ds + O(\|x\|^3) = -\frac{1}{2}\partial_k a_j(b_0) x^j x^k + O(\|x\|^3)$, and the second is $O(\|a\|_{C^1}^2 \|x\|^4)$. Because $x^j x^k$ is symmetric, only the symmetric part $\partial_{(k} a_{j)}(b_0)$ of $\partial_k a_j(b_0)$ contributes; the antisymmetric part is exactly $-\frac{1}{2}F^\nabla_{jk}(b_0)$, and $F^\nabla_{jk} x^j x^k = 0$ identically. Collecting the $O(\|x\|^3)$ Taylor error of a and the $O(\|x\|^4)$ higher Peano terms into R gives the stated bound by (R4). For a general smooth family τ, equation 14 is the second-order Taylor expansion of $x \mapsto \Phi(x)$ with Lagrange remainder, and (a) follows because $D^2\Phi(0)$ is contracted with the symmetric $x \otimes x$. □

Remark 2 (The normal-coordinate coefficient is not canonical). It is tempting to identify A_2 with "the" tidal/sectional-curvature contraction obtained from the normal-coordinate reference. That identification is incorrect, and Proposition 6 is the counterexample: the normal-coordinate reference on TS^2 gives $A_2^{\mathrm{nc}}(x) = +\frac{\kappa}{6}(\|x\|^2 I - x \otimes x)$, whereas the ambient reference used in our sphere

experiment gives $A_2^{\mathrm{amb}}(x) = -\frac{1}{2}\kappa\,(x \otimes x)$. These differ in sign, in trace, and in whether they act isotropically. The constant $-\frac{1}{2}\kappa\Sigma_b v^\star$ reported in Proposition 7 belongs to the ambient reference; it is not the normal-coordinate coefficient, and it is not canonical.

Theorem 5 (Loop holonomy encodes invariant curvature information)**.** *Let S be a geodesic triangle (or piecewise-geodesic loop) at b_0 spanning an oriented area bivector $A^{jk} = -A^{kj}$, $|A| = O(\rho^2)$. The holonomy of ∇ around ∂S for a regular shrinking loop of diameter $O(\sqrt{|A|})$ is*

$$\mathrm{Hol}(\partial S) = \mathrm{Id} - \tfrac{1}{2}\, F^{\nabla}_{jk}(b_0)\, A^{jk} + O(|A|^{3/2}). \tag{22}$$

Both F^{∇}_{jk} and A^{jk} are antisymmetric, so their contraction need not vanish. The $O(|A|^{3/2})$ form assumes regular shrinking loops with diameter $O(\sqrt{|A|})$; for general small loops use $O(\rho\,\mathrm{area}(S) + \mathrm{area}(S)^2)$. For TS^2 this is a rotation by the spherical excess $\kappa|A|$.

Proof of Theorem 5. Parametrize S by a smooth homotopy $h : [0,1]^2 \to U$ with $h(0,\cdot) \equiv b_0$ and boundary ∂S. The non-abelian Stokes theorem for the transport of ∇ gives $\mathrm{Hol}(\partial S) = \mathcal{P}\exp\big(-\int_S \Omega\big)$, where Ω is the curvature 2-form transported into the frame at b_0; expanding the path-ordered exponential and using $\Omega = \frac{1}{2}F^{\nabla}_{jk}\,dx^j \wedge dx^k + O(\rho)$ over S yields $\mathrm{Hol}(\partial S) = \mathrm{Id} - \frac{1}{2}F^{\nabla}_{jk}(b_0)\int_S dx^j\wedge dx^k + O(|A|^{3/2}) = \mathrm{Id} - \frac{1}{2}F^{\nabla}_{jk}(b_0)A^{jk} + O(|A|^{3/2})$. For TS^2, $F^{\nabla}_{jk} = \kappa\,\epsilon_{jk}J$ with J the oriented quarter-turn, so $\frac{1}{2}F^{\nabla}_{jk}A^{jk} = \kappa|A|J$ under the stated orientation convention, a rotation by the enclosed solid angle (Gauss–Bonnet). Table 9 reports the measured rotation against κA. □

Remark 3. Lemma 2 corrects a tempting single-path curvature term (which vanishes) and Theorem 5 supplies the invariant object. Product-space aggregation uses ∇^0, so its per-observation fibre error relative to ∇ is $(\Phi(x) - \mathrm{Id})v = -\frac{1}{2}\partial_{(k}a_{j)}(b_0)x^j x^k v + O(\|x\|^3)$, the quantity averaged by the posterior in Proposition 7 and Theorem 4.

N.2 Proof of Theorem 1 (Monotone ascent)

Proof. By (R1) the exact posterior $q_i(z \mid Y_i;\theta,\psi) = f_\theta(z)\mathcal{K}_\psi(Y_i \mid z)/p_{\theta,\psi}(Y_i)$ is a well-defined probability density w.r.t. $\mu_{\mathcal{E}}$, and for every z in its support

$$\log p_{\theta,\psi}(Y_i) = \log\big(f_\theta(z)\mathcal{K}_\psi(Y_i \mid z)\big) - \log q_i(z \mid Y_i;\theta,\psi). \tag{23}$$

Both terms are $\mu_{\mathcal{E}}$-integrable against $q_i^{(t)}(z) := q_i(z \mid Y_i;\theta^{(t)},\psi^{(t)})$ by (R1)–(R2). Integrating against $q_i^{(t)}$ and summing over i gives the exact identity $\ell_{\mathrm{obs}}(\theta,\psi) = Q(\theta,\psi \mid \theta^{(t)},\psi^{(t)}) + H(\theta,\psi \mid \theta^{(t)},\psi^{(t)})$ with $H = -\sum_i \int q_i^{(t)} \log q_i(\cdot \mid Y_i;\theta,\psi)\,d\mu_{\mathcal{E}}$. Subtracting the same identity at $(\theta^{(t)},\psi^{(t)})$ and using $H(\theta,\psi \mid \cdot) - H(\theta^{(t)},\psi^{(t)} \mid \cdot) = \sum_i D_{\mathrm{KL}}\big(q_i^{(t)} \,\|\, q_i(\cdot \mid Y_i;\theta,\psi)\big)$,

$$\begin{aligned}\ell_{\mathrm{obs}}(\theta^{(t+1)},\psi^{(t+1)}) - \ell_{\mathrm{obs}}(\theta^{(t)},\psi^{(t)}) = {}& \underbrace{\big[Q(\theta^{(t+1)},\psi^{(t+1)} \mid \cdot) - Q(\theta^{(t)},\psi^{(t)} \mid \cdot)\big]}_{\geq 0 \text{ (M-step)}} \\ & + \underbrace{\textstyle\sum_i D_{\mathrm{KL}}\big(q_i^{(t)} \| q_i^{(t+1)}\big)}_{\geq 0}.\end{aligned} \tag{24}$$

Both bracketed terms are nonnegative; the second follows from Gibbs' inequality (finite by (R2)). Hence ℓ_{obs} is nondecreasing. Equality holds exactly when the Q-increment is zero and every KL term vanishes, equivalently $q_i^{(t+1)} = q_i^{(t)}$ $\mu_{\mathcal{E}}$-a.e. for all i. □

N.3 Proof of Theorem 2 (Convergence to stationary points)

Proof. Write $\vartheta = (\theta,\psi)$ and let $g_t = \mathrm{grad}\,\ell_{\mathrm{obs}}(\vartheta^{(t)})$. The exact E-step gives the Fisher identity $\mathrm{grad}_\vartheta\, Q(\vartheta \mid \vartheta^{(t)})|_{\vartheta=\vartheta^{(t)}} = g_t$. By Theorem 1, Assumption 1(d), and the uniform spectral bounds on G, there is $c_0 > 0$ such that

$$\ell_{\mathrm{obs}}(\vartheta^{(t+1)}) - \ell_{\mathrm{obs}}(\vartheta^{(t)}) \geq Q(\vartheta^{(t+1)} \mid \vartheta^{(t)}) - Q(\vartheta^{(t)} \mid \vartheta^{(t)}) \geq c_0\|g_t\|^2. \tag{25}$$

The iterates remain in the compact observed-likelihood superlevel set of Assumption 1(c), on which ℓ_{obs} is bounded above. Summing the preceding inequality yields $\sum_t \|g_t\|^2 < \infty$, hence $\|g_t\| \to 0$. Every accumulation point exists in the compact set and is stationary by continuity of the gradient. Finally, the Armijo-retraction step bound in Assumption 1(d) gives $d(\vartheta^{(t+1)}, \vartheta^{(t)}) \le C\|g_t\| \to 0$. □

N.4 Proof of Theorem 3 (Consistency of the constrained target)

Proof. Write $\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(Y_i)$ and $L(\theta) = \mathbb{E}_{\theta^\star}[\log p_\theta(Y)]$, finite by (R2). We first establish a uniform law of large numbers. By the theorem hypotheses Θ is compact; $\theta \mapsto \log p_\theta(y)$ is continuous for a.e. y; and by (R2) it is dominated by the integrable envelope H. Continuity, compactness, and a $P_{\theta^\star}$-integrable envelope give, by the uniform strong law for separable continuous processes (a Glivenko–Cantelli class via bracketing under the envelope H),

$$\sup_{\theta\in\Theta} \left|\bar{\ell}_n(\theta) - L(\theta)\right| \xrightarrow{\text{a.s.}} 0. \tag{26}$$

By hypothesis, the maximizer set is the finite orbit $[\theta^\star] = \{\sigma\cdot\theta^\star : \sigma \in \mathfrak{S}_K\}$. Fix $\epsilon > 0$; the set $K_\epsilon = \{\theta \in \Theta : d_{\mathrm{perm}}(\theta, \theta^\star) \ge \epsilon\}$ is compact and, by uniqueness modulo labels and continuity, $\delta := L(\theta^\star) - \sup_{\theta\in K_\epsilon} L(\theta) > 0$. On the a.s. event $\{\sup_\Theta |\bar{\ell}_n - L| < \delta/3\}$ (eventually), and using $\bar{\ell}_n(\hat{\theta}_n) \ge \bar{\ell}_n(\theta^\star)$,

$$\text{if } \hat{\theta}_n \in K_\epsilon: \quad L(\hat{\theta}_n) \ge \bar{\ell}_n(\hat{\theta}_n) - \tfrac{\delta}{3} \ge \bar{\ell}_n(\theta^\star) - \tfrac{\delta}{3} \ge L(\theta^\star) - \tfrac{2\delta}{3} > L(\theta^\star) - \delta = \sup_{K_\epsilon} L, \tag{27}$$

a contradiction. Hence $d_{\mathrm{perm}}(\hat{\theta}_n, \theta^\star) < \epsilon$ eventually a.s., a fortiori in probability. (The distance d_Θ is the Riemannian distance on compact Θ, equivalent to any chart metric.) □

N.5 Proofs of Proposition 7, Theorem 4, Proposition 6, and Corollary 1

Proof of Proposition 7. Write each observation as $y_i = P_{b_i\leftarrow b_0} v^\star + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i \mid x_i] = 0$. The transport-aware estimator averages $P_{b_0\leftarrow b_i} y_i = P_{b_0\leftarrow b_i} P_{b_i\leftarrow b_0} v^\star + P_{b_0\leftarrow b_i}\varepsilon_i = v^\star + P_{b_0\leftarrow b_i}\varepsilon_i$, since transport out and back along the same geodesic composes to the identity; hence $\mathbb{E}[\hat{v}_{\mathrm{bundle}}] = v^\star$ (unbiased). The product estimator averages $\tau_{b_0\leftarrow b_i} y_i = \tau_{b_0\leftarrow b_i} P_{b_i\leftarrow b_0} v^\star + \tau_{b_0\leftarrow b_i}\varepsilon_i$, and $\tau_{b_0\leftarrow b_i} P_{b_i\leftarrow b_0} = \Phi(x_i)$ is exactly the single-path discrepancy of Lemma 2. Taking the expectation over the base law ($\mathbb{E}x_i = 0$, $\mathrm{Cov}x_i = \Sigma_b$) and using equation 14,

$$\mathbb{E}[\hat{v}_{\mathrm{prod}}] - v^\star = \big(\mathbb{E}[\Phi(x)] - \mathrm{Id}\big)v^\star = -\tfrac{1}{2}\,\partial_{(k}a_{j)}(b_0)\,\Sigma_b^{jk}\,v^\star + O(\mathbb{E}\|x\|^3).$$

For the tangent bundle, the symmetric coefficient depends on the reference; for the ambient reference on S^2 it is $-\frac{1}{2}\kappa(x\otimes x)$, so $\mathbb{E}[\Phi] - \mathrm{Id} = -\frac{1}{2}\kappa\,\Sigma_b$ and $\|\mathrm{Bias}_{\mathrm{ref}}\| = \frac{1}{2}|\kappa|\|\Sigma_b v^\star\| + O(\rho^3) \le \frac{1}{2}|\kappa|\|\Sigma_b\|_{\mathrm{op}}\|v^\star\| + O(\rho^3)$. □

Proof of Theorem 4. Let $U = \Phi(x)v^\star$ and $B = \mathbb{E}[U] - v^\star = \mathrm{Bias}_{\mathrm{ref}}$. Independence across observations and the conditional noise moment assumptions gives the exact bias–variance identities

$$\mathrm{MSE}(\hat{v}_{\mathrm{prod}}) = \frac{\sigma_v^2}{n}\,\mathrm{tr}\,\mathbb{E}[\tau\tau^*] + \frac{1}{n}\,\mathrm{tr}\,\mathrm{Cov}(U) + \|B\|^2, \tag{28}$$

$$\mathrm{MSE}(\hat{v}_{\mathrm{bundle}}) = \frac{r\sigma_v^2}{n}. \tag{29}$$

Here $\tau = \tau_{b_0\leftarrow b}$ and τ^* is its metric adjoint. Subtraction proves the stated exact identity, including Δ_{noise}. If τ is isometric, that term vanishes, and the remaining gap is nonnegative. With $\Phi(x) = I + A_2(x) + R(x)$, $A_2(x) = O(\rho^2)$ and $R(x) = O(\rho^3)$. The cross terms between A_2 and R are $O(\rho^5)$, uniformly under (R4)–(R5), while $\mathbb{E}[A_2(x)]v^\star$ is the leading bias from Proposition 7. Hence

$$\mathrm{MSE}_{\mathrm{prod}} - \mathrm{MSE}_{\mathrm{bundle}} = \frac{1}{n}\,\mathrm{tr}\,\mathrm{Cov}(A_2(x)v^\star) + \|\mathbb{E}[A_2(x)]v^\star\|^2 + O(\rho^5).$$

Under the isometric specialization, the exact gap is strict if and only if U has nonzero variance or $B \neq 0$. □

Proof of Proposition 6. Work on S^2 of radius R, $\kappa = 1/R^2$, and let $s = \|x\|$, $\hat{u} = x/s$. Decompose $v = v_\parallel \hat{u} + v_\perp$ with $v_\perp \perp \hat{u}$. Parallel transport along the geodesic with initial velocity x carries $\hat{u}$ to the unit tangent $\hat{u}(s) = \cos(s/R)\hat{u} - \sin(s/R)\, b_0/R$ and fixes $v_\perp$ (which is orthogonal to the plane of the geodesic and lies in $T_{b_0} \cap T_b$).

(i) Ambient projection. $\Pi_{T_{b_0}} \hat{u}(s) = \cos(s/R)\hat{u}$ and $\Pi_{T_{b_0}} v_\perp = v_\perp$, so $\Phi_{\mathrm{amb}}(x)v = v - \big(1 - \cos(s/R)\big)v_\parallel \hat{u} = \big(I - (1-\cos(s/R))\,\hat{u}\otimes\hat{u}\big)v$. Since $1-\cos(s/R) = \frac{1}{2}\kappa s^2 + O(s^4)$ and $s^2\,\hat{u}\otimes\hat{u} = x\otimes x$, we get $\Phi_{\mathrm{amb}}(x) = I - \frac{1}{2}\kappa\,(x\otimes x) + O(\|x\|^4)$. Taking expectations with $\mathbb{E}x = 0$, $\mathrm{Cov}x = \Sigma_b$ gives $\mathbb{E}[\Phi_{\mathrm{amb}}] - I = -\frac{1}{2}\kappa\Sigma_b + O(\kappa^2\rho^4)$.

(ii) Normal-coordinate frame. The frame at b is $d(\mathrm{Exp}_{b_0})_x$ applied to a fixed basis of T_{b_0}. By the Jacobi-field computation on a constant-curvature surface, $d(\mathrm{Exp}_{b_0})_x$ is the identity on the radial direction and multiplies the transverse direction by $\sin(s/R)/(s/R)$. Reading the coefficients of the transported vector in that frame therefore rescales the transverse component by $(s/R)/\sin(s/R) = 1 + \frac{1}{6}\kappa s^2 + O(s^4)$ and leaves the radial component fixed, so $\Phi_{\mathrm{nc}}(x) = I + \frac{1}{6}\kappa s^2\big(I - \hat{u}\otimes\hat{u}\big) + O(\|x\|^4) = I + \frac{\kappa}{6}\big(\|x\|^2 I - x\otimes x\big) + O(\|x\|^4)$, whence $\mathbb{E}[\Phi_{\mathrm{nc}}] - I = \frac{\kappa}{6}(\mathrm{tr}\,\Sigma_b\, I - \Sigma_b) + O(\kappa^2\rho^4)$. For $\Sigma_b = \sigma_b^2 I_2$, (i) gives $-\frac{1}{2}\kappa\sigma_b^2 v^\star$ and (ii) gives $+\frac{1}{6}\kappa\sigma_b^2 v^\star$. A script in the released code checks both expansions numerically. □

Proof of Corollary 1. Let $s = \|x\|$, $u = x/s$, and $t = s/R$. Ambient projection gives $\Phi = I - (1 - \cos t)uu^\top$ and $\tau\tau^* = I - \sin^2(t)uu^\top$. Therefore

$$\Delta_{\mathrm{noise}} = -\frac{\sigma_v^2}{n}\mathbb{E}\sin^2(s/R) = -\frac{\sigma_v^2\kappa\,\mathrm{tr}\,\Sigma_b}{n} + O(\sigma_v^2\kappa^2\rho^4/n). \tag{30}$$

The bias is $-\frac{1}{2}\kappa\Sigma_b v^\star + O(\kappa^2\rho^4\|v^\star\|)$, and $\mathrm{tr}\,\mathrm{Cov}(\Phi v^\star) = O(\kappa^2\rho^4\|v^\star\|^2)$. Substitution gives equation 6, with

$$R_{\kappa,\rho,n} = O\big(\kappa^2\rho^4(\|v^\star\|^2 + \sigma_v^2)/n + \kappa^3\rho^6\|v^\star\|^2\big)\,.$$

This retains the finite-sample variance and higher-order bias remainders. Solving the displayed leading quadratic gives equation 7; it is not an exact if-and-only-if threshold. For a nondegenerate fixed bounded base law, the derivative of the exact gap at $\kappa = 0$ is $-\sigma_v^2\mathbb{E}s^2/n < 0$, proving a sufficiently small positive-curvature region where projection has lower risk. A crossover within the admissible region needs a positive gap elsewhere.

For a rotationally symmetric base law on $T_{b_0}S^2$, define $a = \mathbb{E}[1 - \cos(s/R)]$ and $b = \mathbb{E}[(1 - \cos(s/R))^2]$. Angular averaging gives the exact gap

$$\|v^\star\|^2\left[\frac{b}{2n} + \frac{n-1}{4n}a^2\right] - \frac{\sigma_v^2}{n}(2a - b). \tag{31}$$

Here $\mathbb{E}\Phi = (1 - a/2)I$, $\mathrm{tr}\,\mathrm{Cov}(\Phi v^\star) = (b/2 - a^2/4)\|v^\star\|^2$, and $\mathbb{E}\sin^2(s/R) = 2a - b$. This formula retains base-sampling variance and permits numerical evaluation for a fully specified radial law without asserting a universal single threshold. □

N.6 Proof of Proposition 5 (Identifiability of the transported mixture)

Proof. For almost every b with $h(b) > 0$, divide the equality of joint densities by $h(b)$ and apply the known isometry $w = P_{m\leftarrow b}v$. The resulting equality is that of two finite Gaussian mixtures on the fixed fibre $\mathcal{E}_m$. Finite Gaussian-mixture identifiability with distinct components gives equality up to permutation (Teicher, 1963; Yakowitz & Spragins, 1968). Since all component parameters lie in that same fixed fibre, no cross-anchor label matching is needed. This proof does not establish identifiability for unknown component-specific anchors. □

N.7 Proof of Lemma 1 (Transported retraction)

Proof. At zero step, the base point, mean, and covariance are unchanged. Parallel transport preserves their fibre membership and positive definiteness. Differentiation at zero, using the connection splitting, gives horizontal velocity ξ and covariant vertical velocities ζ, H because $DR_\Sigma^\Sigma(0)$ is the identity. These are the retraction axioms; no invariance of Q is required. The differential equals the

chosen input tangent vector, and equals a preconditioned gradient only when that vector was chosen accordingly.

A gauge change at a fixed anchor changes only coordinates and leaves Q unchanged. Moving the anchor changes the base distance, generally the normalizer, and the fibre transport to each datum. Geodesic convexity ensures unique pairwise geodesics but does not imply $P_{m'\leftarrow b} = P_{m'\leftarrow m}P_{m\leftarrow b}$: the discrepancy is triangle holonomy. Holding coordinates fixed in a smooth local trivialization can define a valid chart retraction with its corresponding coordinate derivative; it must not be confused with horizontal transport. □

N.8 Proof of Proposition 4 (Local linear rate)

Proof. Let M be the EM operator $\theta^{(t+1)} = M(\theta^{(t)})$ with an exact M-step, i.e. $M(\phi)$ solves $\nabla_\theta Q(\theta \mid \phi)|_{\theta=M(\phi)} = 0$, where $Q(\theta \mid \phi) = \sum_i \mathbb{E}_{q_i(\cdot;\phi)}[\log f_\theta(Z)\mathcal{K}(\cdot \mid Z)]$. By (R1)–(R3) the map $(\theta,\phi) \mapsto \nabla_\theta Q(\theta \mid \phi)$ is C^1 and $\nabla^2_{\theta\theta} Q(\theta^\star \mid \theta^\star) = -\mathcal{I}_{\mathrm{com}}(\theta^\star) \prec 0$ is invertible, so the implicit function theorem (in a normal chart at the fixed point $\theta^\star = M(\theta^\star)$) gives $M \in C^1$ near $\theta^\star$ with differential

$$dM(\theta^\star) = -\left[\nabla^2_{\theta\theta} Q\right]^{-1} \nabla^2_{\theta\phi} Q\big|_{(\theta^\star,\theta^\star)} = \mathcal{I}_{\mathrm{com}}(\theta^\star)^{-1}\, \mathcal{I}_{\mathrm{mis}}(\theta^\star), \tag{32}$$

using the standard EM identities $\nabla^2_{\theta\theta} Q(\theta^\star \mid \theta^\star) = -\mathcal{I}_{\mathrm{com}}$ and $\nabla^2_{\theta\phi} Q(\theta^\star \mid \theta^\star) = \mathcal{I}_{\mathrm{mis}}$ (the latter because varying ϕ moves the posterior, whose score covariance is $\mathcal{I}_{\mathrm{mis}}$). Thus $\rho(dM(\theta^\star)) = \rho(\mathcal{I}_{\mathrm{com}}^{-1}\mathcal{I}_{\mathrm{mis}}) = \rho(\mathcal{I}_{\mathrm{mis}}\mathcal{I}_{\mathrm{com}}^{-1}) < 1$ by (R3) and $\mathcal{I}_{\mathrm{obs}} = \mathcal{I}_{\mathrm{com}} - \mathcal{I}_{\mathrm{mis}} \succ 0$. Ostrowski's theorem for the fixed-point iteration $\theta^{(t+1)} = M(\theta^{(t)})$ then yields local linear convergence with asymptotic rate $r = \rho(\mathcal{I}_{\mathrm{mis}}\mathcal{I}_{\mathrm{com}}^{-1})$. The geometry enters $\mathcal{I}_{\mathrm{mis}}$ only through the intrinsic E-step moments; we do not attribute a separate curvature term to the rate. □

N.9 On a biased-estimator (curvature-corrected) bound

We do not claim a curvature-corrected Cramér–Rao theorem. The transport-aware estimator is unbiased under Proposition 7; under the isotropic rank-r model its MSE is $r\sigma_v^2/n$ (Theorem 4). The product estimator is biased, and a biased-estimator bound $\mathrm{Cov}[e] \succeq (I+\nabla_\theta \mathrm{Bias})\mathcal{I}_{\mathrm{obs}}^{-1}(I+\nabla_\theta \mathrm{Bias})^\top$ carries sensitivity terms $\nabla_\theta \mathrm{Bias}_{\mathrm{ref}}$. For comparing a biased and an unbiased estimator the exact MSE identity of Theorem 4 is the relevant quantity, and it is the one we use.

Remark 4 (On the perturbative scope). Proposition 7 is a second-order local expansion with an $O(\rho^3)$ remainder. Theorem 4 first gives an exact MSE identity and then an expansion with $O(\rho^5)$ remainder, controlled by (R4)–(R5). The loop-holonomy law (Theorem 5) is a leading-order gauge-covariant expansion with gauge-invariant conjugacy observables; on a round sphere the rotation-angle/area relation is exact modulo 2π.

O Proofs for Section 5

Propositions 2 and 3 are stated in full in Appendix H; we do not restate them here and prove them below alongside the remaining results of §5.

Throughout this appendix V is a C^3 section of $\mathcal{E}$ over $B_\rho(b_0)$. Use L_1, L_3 for full derivative bounds and $L_2 = \sup_{b,\|u\|=1} \|\nabla^2_{u,u} V(b)\|$ for the symmetric Hessian bound. Sampling and conditional noise assumptions are those of §5.

O.1 Covariant Taylor expansion and the radius tradeoff

Proof of Lemma 3. Set $u(s) := P_{b_0\leftarrow\gamma(s)} V(\gamma(s)) \in \mathcal{E}_{b_0}$, a curve in a fixed finite-dimensional vector space. Differentiating and using that parallel transport intertwines d/ds with the covariant derivative along γ, $u'(s) = P_{b_0\leftarrow\gamma(s)}\big(\nabla_{\gamma'(s)} V\big)$. Since γ is a geodesic, $\nabla_{\gamma'}\gamma' = 0$, so iterating gives $u^{(k)}(s) = P_{b_0\leftarrow\gamma(s)}\big(\nabla^k_{\gamma',\ldots,\gamma'} V\big)$ for $k \le 3$. Taylor's theorem with Lagrange remainder applied to u at $s = 0$, evaluated at $s = 1$ with $\gamma'(0) = x$, yields the statement with $\|R_3\| \le \frac{1}{6}\sup_s \|u'''(s)\| \le \frac{1}{6} L_3 \|x\|^3$, transport being an isometry. □

Theorem 6 (Bias–variance tradeoff in the aggregation radius)**.** *With $U(x) = P_{b_0 \leftarrow b} V(b)$, the model bias is*

$$B = \tfrac{1}{2}\Sigma_b^{jk}(\nabla^2_{jk}V)(b_0) + R, \qquad \|R\| \le \tfrac{1}{6}L_3\mathbb{E}\|x\|^3. \tag{33}$$

The exact transport-estimator risk is

$$\mathcal{R} = \|B\|^2 + \{r\sigma_v^2 + \operatorname{tr}\operatorname{Cov}(U(x))\}/n.$$

In particular $\operatorname{tr}\operatorname{Cov}(U) \le L_1^2\mathbb{E}\|x\|^2$ by integration along radial geodesics. For an idealized deterministic local sample count $n = n_0\omega_d\rho^d$ and $\operatorname{tr}\Sigma_b = c_d\rho^2$, this gives

$$\mathcal{R} \le \tfrac{1}{4}L_2^2c_d^2\rho^4 + \frac{r\sigma_v^2 + L_1^2c_d\rho^2}{n_0\omega_d\rho^d} + O(\rho^5).$$

When noise dominates the base-sampling variance and $L_2 > 0$, minimizing the leading noise–bias bound yields the stated $\rho^\star$ and an upper-bound rate of order $n_0^{-4/(d+4)}$ for fixed positive noise (Tsybakov, 2009). This is neither an exact finite-sample risk optimizer nor a minimax lower bound. A random local sample count requires conditioning and a rule for empty neighborhoods.

Proof. Average Lemma 3; centering cancels the linear term. The symmetric quadratic term gives equation 33, and $\|B\| \le L_2 \operatorname{tr}\Sigma_b/2 + O(L_3\rho^3)$. Conditional noise has zero cross-covariance with $U(x)$, so total variance gives the risk identity. The radial integral bounds $\|U(x) - V(b_0)\| \le L_1\|x\|$. Substitute the sample-count and dispersion relations and differentiate the leading function $L_2^2c_d^2\rho^4/4 + r\sigma_v^2/(n_0\omega_d\rho^d)$ to obtain $\rho^{d+4} = dr\sigma_v^2/(L_2^2c_d^2n_0\omega_d)$. For the reference estimator, use equation 18; its variance generally differs. □

Corollary 2 (A sufficient leading-bias condition)**.** *On TS^2 with ambient projection and isotropic $\Sigma_b = \sigma_b^2 I_2$, set $b = \frac{1}{2}\Sigma_b^{jk}(\nabla^2_{jk}V)(b_0)$ and $c = -\frac{1}{2}\kappa\Sigma_b v^\star$. The leading squared-bias difference is*

$$\|b + c\|^2 - \|b\|^2 = \tfrac{1}{4}\kappa^2\|\Sigma_b v^\star\|^2 - \kappa\langle\Sigma_b v^\star, b\rangle.$$

For $L_2 > 0$, $\Xi = |\kappa|\|v^\star\|/L_2 > 4$ is sufficient for this leading difference to be positive. For the finite-radius bias comparison, the leading margin must dominate the $O(\rho^5)$ remainder. A small Ξ provides no ranking: reference bias can cancel model bias when b is sufficiently aligned with $\kappa\Sigma_b v^\star$. Full MSE additionally obeys equation 18. If $L_2 = 0$ and the numerator is positive, interpret $\Xi = \infty$; if both vanish, compare the bias terms directly.

Proof. Expand the difference of squared norms. Since $\|b\| \le L_2 \operatorname{tr}\Sigma_b/2 = L_2\sigma_b^2$, $\|c\| > 2\|b\|$ follows from $|\kappa|\|v^\star\| > 4L_2$. Conversely, if $b = \alpha\kappa\Sigma_b v^\star$ with $\alpha > 1/4$, the displayed difference is negative. Alignment with $\kappa\Sigma_b v^\star$, rather than anti-alignment, produces this cancellation. □

O.2 Curvature as an obstruction

Proposition 8 (Parallel sections, radial extensions, and anchor dependence)**.** *Let $V^{b_0}(b) = P_{b \leftarrow b_0} v^\star$ on a normal ball.*

- *(i) If $\nabla V = 0$, then $F^\nabla(X, Y)V = 0$ for every X, Y. Thus a nonzero parallel tangent section cannot exist on an open surface patch with nonzero Gaussian curvature.*
- *(ii) The radial extension nevertheless exists smoothly. At its anchor, $\nabla V^{b_0} = 0$ and $\operatorname{Sym}\nabla^2 V^{b_0} = 0$, while $\nabla^2_{X,Y}V^{b_0} - \nabla^2_{Y,X}V^{b_0} = F^\nabla(X, Y)v^\star$. Curvature can force a nonzero antisymmetric Hessian without forcing the smoothing-bias contraction. On a shrinking ball, the symmetric Hessian of this smooth radial field tends to zero; there is no general conclusion $\Xi = O(1)$ for the symmetric-Hessian diagnostic.*
- *(iii) Set the value at a second anchor to $v_1 = P_{b_1 \leftarrow b_0} v^\star$. Comparing $P_{b \leftarrow b_1} v_1$ with $P_{b \leftarrow b_0} v^\star$ amounts to triangle holonomy, with the area-dependent expansion in Theorem 5 after returning to a common fibre.*
- *(iv) The transport estimator is exactly unbiased at the anchor b_0 under this radial model. Other anchors can introduce bias, but no universal nonzero lower bound or "every other anchor" claim follows; sampling symmetry and curvature action can produce cancellation.*

Proof. Part (i) is the commutator identity defining curvature. On a surface with $K \neq 0$, its action on tangent vectors is a nonzero multiple of a quarter-turn and has zero kernel. For (ii), the curve $P_{b_0 \leftarrow \gamma(s)} V^{b_0}(\gamma(s))$ is constant for every radial geodesic. Its first and second derivatives vanish; polarization gives zero symmetric covariant Hessian at b_0. The commutator gives the antisymmetric part. Smoothness gives the stated local continuity. Part (iii) follows by composing the two transports and their reverse into a loop, and part (iv) by transporting each observation back along the generating radial path and taking expectations. □

O.3 Path dependence and connection misspecification

Proof of Proposition 2. Compose $P^{\tilde{\gamma}}$ with the inverse of P^{γ} to obtain H in the target fibre. The small-surface holonomy expansion bounds $\|H - I\|$ by curvature times surface area, with the diameter/area remainder stated in the proposition. Jensen's inequality bounds $\|(\mathbb{E}H - I)v^\star\|$ by $\mathbb{E}\|H - I\|\|v^\star\|$. The ribbon hypothesis gives area $O(h\rho)$. The bound is independent of n; the signed mean error can still cancel across paths. □

Proof of Proposition 3. In a parallel local gauge, variation of transport gives a first-order term linear in $\delta_j(b_0)x^j$ and remainder $O((\|\nabla\delta\| + \|\delta\|^2)\rho^2)$. For centered sampling, the linear mean vanishes. The C^1 bound and $\eta \leq 1$ make the remainder $O(\eta\rho^2)$; without centering retain $O(\eta\rho)$. A C^0 bound does not control $\nabla\delta$, so it cannot justify the improved centered order. □

O.4 Topology

Remark 5 records a further case invisible to every local expansion here: on a non-simply-connected base a flat connection can still carry nontrivial holonomy, which no $O(\rho^2)$ bias law detects. The proof of Proposition 1 follows it.

Remark 5 (Topology determines the admissible baseline class; flat connections can still have global holonomy). Table 1 organizes regimes by curvature and discretization, but topology enters as a prior constraint on which baselines are even definable: a trivial bundle admits a global coordinate baseline (torus); a nontrivial one forces a choice between local gauges, a singular/patched frame, or a non-cocycle family such as ambient projection (TS^2). A further case is invisible to every local expansion in this paper: Theorem 5 concerns a loop that bounds a small oriented area, so curvature is the entire leading-order holonomy there; on a non-simply-connected base, a loop need not bound any area at all, and a connection can then be flat ($F^\nabla \equiv 0$ everywhere) while still carrying nontrivial holonomy around it. The standard example is the trivial complex line bundle over the circle, $\mathcal{E} = S^1 \times \mathbb{C}$, with connection $\nabla = d + i\alpha\, d\theta$ for constant α: $F^\nabla = d(i\alpha\, d\theta) = 0$ identically, yet the holonomy once around the circle is $e^{-2\pi i\alpha} \neq 1$ unless $\alpha \in \mathbb{Z}$. Such holonomy is a global invariant of the flat connection up to gauge (a representation of $\pi_1(\mathcal{M})$ in the structure group, sometimes called its monodromy), not a curvature effect, and none of our $O(\rho^2)$ bias laws would detect it. It does not arise in §6, whose discrete Levi-Civita connection has nonzero curvature on open regions and zero curvature on its top and bottom circles; a practitioner using a learned or discretized connection on a non-simply-connected base should check for it as a separate diagnostic from Table 1, by measuring holonomy around generators of $\pi_1(\mathcal{M})$ directly rather than inferring it from local curvature.

Proof of Proposition 1. ($\Leftarrow$) If $\mathcal{E} \cong \mathcal{M} \times \mathbb{R}^r$ with global trivialization $\Psi_b : \mathcal{E}_b \to \mathbb{R}^r$, set $\tau_{b' \leftarrow b} = \Psi_{b'}^{-1}\Psi_b$; this is continuous, satisfies $\tau_{b \leftarrow b} = \mathrm{Id}$, and is a cocycle. ($\Rightarrow$) Conversely, given such a family, fix b_0 and define $\Psi_b := \iota \circ \tau_{b_0 \leftarrow b}$ for a fixed linear isomorphism $\iota : \mathcal{E}_{b_0} \to \mathbb{R}^r$. The cocycle property makes Ψ well-defined and fibrewise invertible, and continuity of τ makes (π, Ψ) a global trivialization; hence $\mathcal{E}$ is trivial. For (i), ∂_u, ∂_v are globally defined, nonvanishing and independent on the embedded torus, so $T\mathbb{T}^2$ is trivial. For (ii), $\chi(S^2) = 2 \neq 0$, so by Poincaré–Hopf every continuous tangent field on S^2 vanishes somewhere, and no global frame exists; consequently, no global cocycle family τ exists on TS^2. This topological statement does not impose a universal quantitative frame-error law. For ambient projection on a round sphere, the singular values are 1 and $|\cos(d_{\mathcal{M}}(b, b_0)/R)|$; rank drops at $d_{\mathcal{M}}(b, b_0) = \pi R/2$. At the antipode the tangent planes coincide, so ambient projection is nonsingular, while the minimizing-geodesic transport is nonunique. These are distinct obstructions. □

P COMPUTATIONAL COMPLEXITY, PER OPERATION

Table 12 reports the cost of one Bundle-GEM iteration on n observations with rank-r fibres, d-dimensional base, P latent particles (or quadrature nodes) per observation and component, K mixture components, K_c inner Karcher iterations for the base Fréchet mean, and M M-step gradient/line-search evaluations. The general-posterior cost is dominated by P and K_c, not by the single additional parallel transport that distinguishes Bundle-GEM from a product-space iteration; dense vector-fibre transport costs $O(r^2)$ once its matrix is available. Base exponential, logarithm, path construction, and transport-ODE costs depend on geometry; SPD matrix functions generally require cubic cost in matrix order and are not covered by a universal $O(d)$ claim.

Table 12: Per-iteration complexity of Bundle-GEM. Costs are per observation unless noted; totals multiply by n. The extra transport is the only term absent from a product-space iteration.

Operation	Cost (per observation)	Notes
Posterior evaluation	$O(KP\,(c_{\mathcal{K}} + c_f))$	P particles per component
Base Fréchet/Karcher mean	$O(KK_cPd)$	K_c inner iters per component
Parallel transport	$O(KPr^2)$	one transport per particle and component
Fibre aggregation	$O(KPr^2)$	vectors and second moments
Monte-Carlo E-step	$O(KP)$ samples	shared with flat EM
M-step	$O(Mc_Q)$	c_Q: one complete Q/gradient evaluation
Extra vs. product-space	$O(KPr^2)$	transport per particle and component

Wall-clock. On the full Bundle-GEM task ($n = 400$, $K = 2$, exact E-step), one E- and M-step iteration takes 0.015 s with transport and 0.013 s with the ambient reference on one CPU core, so transport adds about 17%. Full runs take 1–3 s. Training the gauge-equivariant network of Appendix V.5 for 3000 steps takes about 3 minutes, against 35 s for the standard networks, on one core.

Q CURVED-MESH EXPERIMENT (TORUS TANGENT BUNDLE)

Construction. We mesh a torus $((a + c\cos v)\cos u, (a + c\cos v)\sin u, c\sin v)$ with $a = 1$, $c = 0.45$ on a 72×36 periodic grid ($N = 2592$ vertices), giving exact outward normals and exact Gaussian curvature $K = \cos v/(c(a+c\cos v))$, which is positive on the outer equator, zero at top/bottom, and negative (largest $|K|$) on the inner equator. The discrete Levi-Civita connection transports a tangent vector across an edge $p \to q$ by the minimal rotation taking n_p to n_q (Rodrigues). Out-and-back along the same discrete path is the exact identity, so the radial-model control is unbiased by construction; arbitrary fields need not be. Geodesic neighborhoods use Dijkstra distances on Euclidean edge lengths; paths for transport are recovered from Dijkstra predecessors. A global orthonormal frame (e_u, e_v) (the torus is parallelizable) defines the coordinate-frame baseline.

(A) Controlled ambient-reference error (implementation check). The field here is generated as radial transports of a single latent value, so by Proposition 8(iv) the transport-aware estimator is exactly unbiased by construction; this test confirms the implementation and the observed scaling in ρ and association with $|K|$, not the mechanism. At a query vertex q with latent unit tangent v_q, every vertex within geodesic radius ρ receives a transported copy $P_{q\to i}v_q$; the transport-aware estimate $\mathrm{mean}_i P_{i\to q}(P_{q\to i}v_q) = v_q$ is exact, and the ambient estimate $\mathrm{proj}_{T_q}\mathrm{mean}_i(P_{q\to i}v_q)$ carries an ambient-reference aggregation error (Table 13). The bias grows with ρ and monotonically with local $|K|$; the parallel-transport bias is 0 to machine precision throughout.

(B) Imputation. The ground-truth field is obtained by transporting the seed vector e_u at vertex 0 to every vertex along a breadth-first spanning tree of the mesh. It is corrupted by tangent Gaussian noise ($\sigma = 0.06$) and masked, and each masked vertex is reconstructed from the observed vertices within graph-geodesic radius ρ (at least three) by four rules: the ambient mean projected onto the

Table 13: Controlled ambient-reference aggregation error on the torus (noiseless; 120 query vertices). Left: mean ambient vs. parallel-transport error as the aggregation radius grows. Right: ambient error stratified by local $|K|$. Parallel transport is 0 in every cell by construction (Proposition 8(iv)), so these columns check the implementation and the reported scaling.

radius ρ	0.10	0.15	0.22	0.30	$\lvert K\rvert$	0.36	0.90	1.47	3.15
ambient bias	0.0017	0.0029	0.0038	0.0078	ambient bias	0.0012	0.0026	0.0064	0.0210
PT bias	0.0000	0.0000	0.0000	0.0000	PT bias	0.000	0.000	0.000	0.000

target tangent plane; the mean of (e_u, e_v) coordinates; one-shot transport of each neighbour along its Dijkstra path; and Bundle-GEM, which starts from the one-shot estimate and applies 40 sweeps of robust transport-aware propagation over the 1-ring, $W_m \leftarrow \sum_q w_{qm} R_{q\to m} W_q / \sum_q w_{qm}$ with Cauchy weights $w_{qm} = 1/(1 + (\|R_{q\to m}W_q - W_m\|/0.5\sigma)^2)$, which is the M-step of a Cauchy-noise model with the latent field held at observed values. The error is the mean Euclidean norm of the vector error over masked vertices for one mask draw (seed 0). Table 14 reports vector error as the window widens (mask 50%) and across masking rates (window $\rho = 0.22$). These are descriptive point estimates beyond the radial signal model. One-shot transport has the largest error at each reported setting. Bundle-GEM has lower error than the coordinate baseline at windows 0.10 and 0.18, but higher error at 0.28; across mask fractions it wins only at 20%. These are point estimates; Appendix V.4 gives intervals for a reimplementation. The spanning-tree field need not be smooth across non-tree edges, so it does not automatically satisfy the C^3 hypotheses. Appendix V.4 isolates the mechanism: the field is not smooth with respect to the connection across the branches of its generating tree, and local rules err where the generating and local paths enclose holonomy. Figure 3 shows the corresponding trends.

Extrinsic curvature qualification. For an embedded surface with shape operator S_{b_0}, ambient projection of a parallel vector along a geodesic gives

$$\Phi_{\mathrm{amb}}(x) = I - \tfrac{1}{2}(S_{b_0}x) \otimes (S_{b_0}x) + O(\|x\|^3).$$

To see this, differentiate an ambient representation of a parallel tangent vector twice; its first derivative is normal, while its second tangential derivative is $-\langle Sx, v\rangle Sx$. The sphere has $S = \pm R^{-1} I$, recovering the spherical coefficient. On a torus, $K = \det S$ does not determine $S\Sigma_b S^*$; even $K = 0$ need not remove ambient reference bias. Therefore, the $|K|$-stratified table reports an empirical association, not a general Gaussian-curvature bias law.

Table 14: Torus imputation, vector error (lower is better). Left: vs. aggregation window (mask 50%). Right: vs. mask fraction (window $\rho = 0.22$).

window ρ	0.10	0.18	0.28	mask	20%	40%	60%	80%
ambient	0.0772	0.0764	0.0856	ambient	0.0642	0.0740	0.0793	0.1010
coord-frame	0.0770	0.0732	0.0741	coord-frame	0.0585	0.0692	0.0752	0.0971
parallel transp.	0.0821	0.0924	0.1307	parallel transp.	0.0902	0.0967	0.1022	0.1236
Bundle-GEM	0.0717	0.0694	0.0790	Bundle-GEM	0.0565	0.0756	0.0773	0.1005

R GAUGE-EQUIVARIANT PIPELINE (ARCHITECTURE, EQUIVARIANCE, RESULTS)

Mesh and fields. The encoder study uses a coarser torus than Appendix Q: $a = 1$, $c = 0.55$ on a 32×16 periodic grid (512 vertices) with the four edge directions $(1,0), (0,1), (1,1), (1,-1)$, local frames (e_u, e_v), connection phases θ_{ji} from the minimal rotation of unit normals, and edge weights $w_{ij} = \exp(-\|p_i - p_j\|^2/0.1^2)$. Fields are covariantly smooth: starting from i.i.d. complex Gaussian values, 60 steps of transport-aware diffusion $z_i \leftarrow \frac{1}{2}z_i + \frac{1}{2}\sum_j w_{ij} e^{\mathrm{i}\theta_{ji}} z_j / \sum_j w_{ij}$ are applied, the

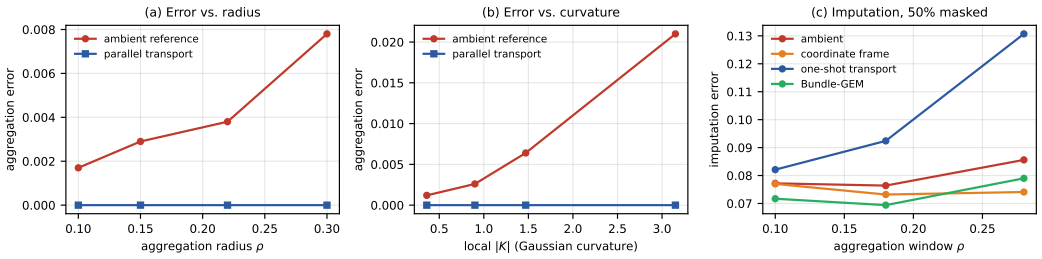

Figure 3: Torus study. (a) The ambient-reference aggregation error grows with the radius, while parallel transport is exact by construction. (b) The ambient error grows with the local Gaussian curvature $|K|$. The vertical axis in (a) and (b) is the reference-dependent error of Section 4, not holonomy, which is measured in Table 9. (c) Imputation error against the window at 50% masking; Appendix V.4 explains the poor result of one-shot transport on this field. A global coordinate frame exists because $T\mathbb{T}^2$ is trivial (Proposition 1).

result is normalized to unit mean modulus and multiplied by a random global phase. The training and test banks have 40 fields each, generated with different seeds.

Representation and layers. Each vertex carries tangent features written as complex numbers in its local frame. Table 15 lists the architecture and training settings. A layer first updates the invariant stream $s_i \leftarrow s_i + \mathrm{ReLU}(A[s_i, \bar{s}_i])$, with $\bar{s}_i$ the weighted neighbour mean, then forms the transported message $m_i = \sum_j w_{ij} e^{\mathrm{i}\theta_{ji}} z_j / \sum_j w_{ij}$ and updates $z_i \leftarrow \mathrm{nrelu}_b(W_s z_i + W_m m_i) \odot \sigma(G s_i)$, where W_s, W_m are complex-linear channel maps (they commute with the phase change of a gauge rotation), $\mathrm{nrelu}_b(z) = z\,\mathrm{relu}(|z| + b)/|z|$ acts on each channel with a gauge-invariant bias, and the sigmoid gate uses only invariant features. Messages propagate through all vertices, including masked ones. The head outputs a complex mean μ_i (one complex-linear map) and a gauge-invariant log-precision (a two-layer perceptron on s_i).

Table 15: Torus encoder: architecture and training.

Item	Setting
Inputs	observed value y_i (zero if masked) lifted to 20 complex channels; invariant input $(\mathrm{mask}_i, |y_i|)$ mapped to 12 channels
Layers	5 layers, 20 complex type-1 channels, 12 invariant channels, gates and norm nonlinearities as above
Output	complex mean μ_i; log-precision from a $12 \to 16 \to 1$ perceptron
Parameters	11,241 real parameters
Variants	Euclidean: $\theta_{ji} = 0$, $w_{ij} = 1$; manifold-only: $\theta_{ji} = 0$, geodesic w_{ij}; gauge-equivariant: both
Loss	$\frac{1}{2}\lvert\mu_i - z_i\rvert^2 + 0.1\,(\frac{1}{2}e^{\ell_i}\lvert\mu_i - z_i\rvert^2 - \ell_i)$ per vertex, weight 1 on masked and 0.3 on observed vertices
Training	Adam, learning rate 3×10^{-3}, 250 steps, batches of 4 fields, 50% masking, noise $\sigma = 0.15$, seed 0
Metric	mean Euclidean norm of the error on masked vertices, averaged over the 40 test fields
Refinement	40 fixed-point steps of the MAP objective with encoder prior precision $e^{\ell_i} \in [0.05, 50]$, data precision $1/\sigma^2$ at observed vertices, and a Cauchy-weighted transport-smoothness term ($\lambda = 1$, scale 0.5σ)
Noise shift	encoder trained at $\sigma = 0.15$; standalone Bundle-GEM uses the true test σ, starts from the encoder output with zero prior weight, and runs 60 robust steps

Exact gauge-equivariance. Under a per-vertex re-gauge (rotate features by $-g_i$ and set $\theta_{ji} \to \theta_{ji} - g_i + g_j$), a correct network's output rotates by $-g_i$. Measured relative error: gauge-equivariant encoder 1.2×10^{-7}, Euclidean encoder 0.60.

Four paradigms. All share the architecture and differ only in base weighting and the connection: *Euclidean* (no transport, uniform weights), *manifold-only* (no transport, geodesic weights), *gauge-equivariant* (transport + geodesic weights), and *gauge + Bundle-GEM* (encoder output refined by the robust transport-aware MAP step). Each is trained on 40 covariantly smooth (connection-harmonic) fields with 50% masking and Gaussian noise $\sigma = 0.15$, and evaluated on 40 held-out fields. Table 16 reports the masked-vector error and Figure 4 summarizes the comparisons; all settings are in Table 15, and the code is `encoder/exp_gauge.py` in the repository. The refinement is a MAP step with its own objective, not a maximum-likelihood GEM step.

Table 16: Torus tangent-field imputation, masked-vector error (lower is better; held-out test fields). The gauge-equivariant encoder has lower point estimates than the two non-transport encoders. Refinement changes the error by 0.0001 at 50% and 80% masking, which is negligible.

masked fraction	20%	50%	80%
Euclidean encoder	0.2008	0.2020	0.3631
Manifold-only encoder	0.1267	0.1657	0.3393
Gauge-equivariant encoder	**0.1143**	**0.1460**	**0.3050**
+ Bundle-GEM refinement	–	0.1459	0.3049

Refinement and distribution shift. Bundle-GEM refinement moves the field by 6–10% while changing the masked error by only 0.0001. The refinement is therefore active, but this does not show that the encoder has reached a GEM fixed point. Under noise shift, the trained encoder also has lower task error than the standalone model-based estimator (Table 17). The two procedures use different information, so this compares task error, not objectives.

Table 17: Noise distribution shift (train $\sigma = 0.15$, mask 50%): task error of the trained encoder and standalone Bundle-GEM given the true test σ. The two methods use different information.

test σ	0.15	0.30	0.45	0.60
gauge-equivariant encoder	0.1460	0.2173	0.2990	0.3849
Bundle-GEM (model)	0.2485	0.3362	0.4345	0.5391

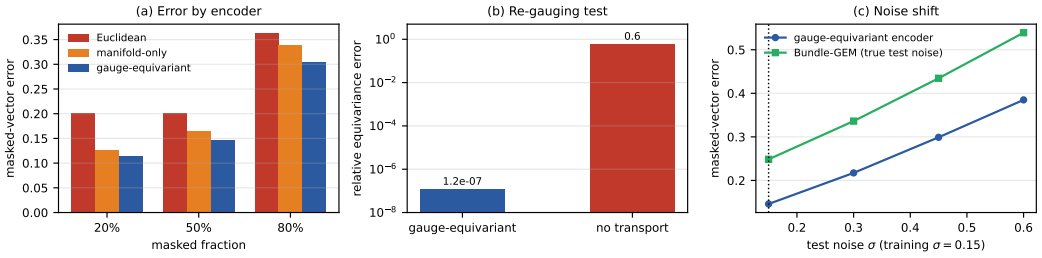

Figure 4: Gauge-equivariant pipeline. (a) The gauge-equivariant encoder improves over no-transport and base-only encoders at every masking rate. (b) Its numerical equivariance error is $\sim 10^{-7}$, versus ~ 0.6 without transport. (c) Noise-shift task errors; the encoder has lower task error than standalone Bundle-GEM at every test noise level.

S SPD STRUCTURE-TENSOR FIELDS: MANIFOLD-FIBRE EXTENSION

We compute the structure-tensor field $J(x) = G_\rho * (\nabla I \nabla I^\top) + \varepsilon I$ of a real grayscale image (van der Walt et al., 2014), a 2×2 SPD descriptor per pixel of the diffusion-tensor type (Weickert, 1998; Pennec et al., 2006). The base grid $\mathbb{R}^2$ is flat. This experiment compares intrinsic affine-invariant Fréchet aggregation with other matrix means; a flat base alone does not define a

Table 18: Real-image SPD structure-tensor reconstruction (72×72 grid, 40% censored). Bundle-GEM (intrinsic affine-invariant aggregation) vs. Product-EM (Euclidean). Lower error is better; the swelling ratio $\det(\hat{J})/\det(J^\star)$ should be 1 (Euclidean averaging inflates it). Stratifying by ground-truth anisotropy (FA quartiles) shows where intrinsic geometry helps.

Image	Metric	Overall	Q1 (low FA)	Q2	Q3	Q4 (high FA)
coffee	Bundle-GEM error ↓	**0.274**	0.155	0.238	0.286	0.416
	Product-EM error	0.337	0.209	0.350	0.399	0.392
	error reduction	**18.9%**	25.5%	31.9%	28.5%	−6.1%
	Product-EM swelling	1.154	1.135	1.218	1.207	1.047
	Bundle-GEM swelling	**1.035**	1.108	1.114	0.988	0.765
astronaut	Bundle / Product error	**0.451** / 0.565	overall reduction **20.1%**; Product swelling up to 1.45			
log-Euclidean	$\text{Bundle}_{\text{AI}}$ / LogEuc error	coffee $0.274/\mathbf{0.268}$, astronaut $0.451/\mathbf{0.444}$: log-Euclidean matches affine-invariant ($\leq 2.3\%$), same swelling 1.04				

Table 19: Ablations on the coffee SPD task (error reduction of Bundle-GEM over Product-EM; Product determinant swelling in parentheses). The intrinsic gain is robust to censoring and grows, then plateaus with neighborhood size.

Censoring	20%	40%	60%	Window	3×3	5×5	7×7
error reduction	19.8%	18.9%	15.4%	error reduction	8.9%	18.9%	18.4%
Product swelling	(1.152)	(1.154)	(1.146)	Product swelling	(1.051)	(1.154)	(1.244)

vector-bundle connection or turn SPD averaging into a Bundle-GEM E-step. We censor 40% of pixels and reconstruct each missing tensor by aggregating its observed neighbors (5×5 window, inverse-distance weights). The method labeled Bundle-GEM here uses the affine-invariant mean ($d(A, B) = \| \log(A^{-1/2} B A^{-1/2}) \|_F$); Product-EM uses the Euclidean mean. Errors are affine-invariant distances to ground truth over 1849 reconstructed tensors.

Table 18 and Figure 5 summarize the results. Intrinsic aggregation reduces reconstruction error by 18.9% (coffee) and 20.1% (astronaut) overall, and the Euclidean baseline shows determinant swelling that grows with anisotropy (median ratio up to 1.22 on coffee, 1.45 on astronaut), consistent with a known artifact of linear tensor averaging (Arsigny et al., 2006), whereas Bundle-GEM stays closer to unit volume overall and in the first three quartiles (in Q4 it under-swells, 0.765). The gain is largest in low-to-mid anisotropy quartiles (25–32%) and vanishes or slightly reverses in the extreme-anisotropy quartile (-6% on coffee), where local stationarity is least plausible. The two methods use identical weights and differ in the fibre mean. A log-Euclidean baseline matches the affine-invariant method within about 2.3% with the same low swelling (Table 18), confirming the gain comes from non-Euclidean aggregation rather than the affine-invariant metric specifically; details are in Appendix U.

Ablations. Table 19 varies the two design choices. The advantage is stable across missingness (15–20% over censoring rates 20–60%) and grows then plateaus with the aggregation window ($8.9\% \to 18.9\% \to 18.4\%$ from 3×3 to 7×7): larger neighborhoods pool more heterogeneous tensors, so respecting the fibre geometry matters more, and the Euclidean swelling worsens ($1.05 \to 1.24$). Because the base is flat and the fibre is not a vector space, this result does not test Proposition 7.

Baseline control. The arithmetic and intrinsic aggregation rules use the same weights. A comparison of two fully fitted EM models would also have to specify whether responsibilities are fixed or recomputed under each model. On TS^2, Product-EM replaces parallel transport by the specified ambient reference and therefore tests the general linear-reference identity in Theorem 4; the isometric nonnegative-gap corollary does not apply to ambient projection. On the SPD task, it replaces an intrinsic mean by an arithmetic mean; that comparison tests fibre geometry only.

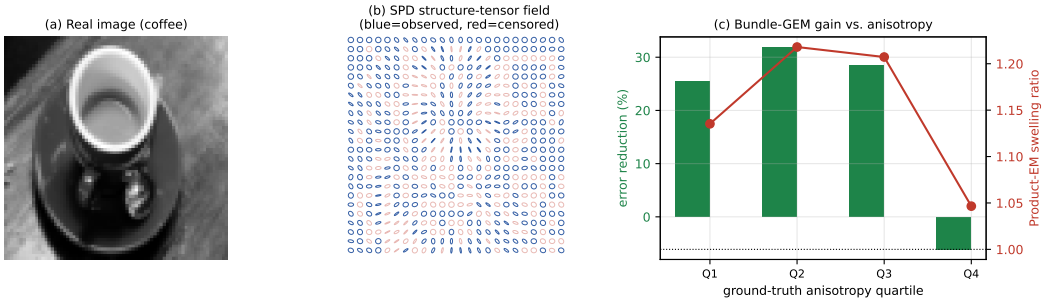

Figure 5: Real-data SPD reconstruction. (a) Source image. (b) Its SPD structure-tensor field (ellipses; blue observed, red censored). (c) Bundle-GEM error reduction (bars) and Product-EM determinant swelling (line) by ground-truth anisotropy quartile: intrinsic aggregation helps in the first three quartiles but slightly underperforms in Q4, while reducing volume inflation overall.

T Experimental details

TS^2 study. The anchor b_0 is the north pole of a sphere of radius $R = \kappa^{-1/2}$ and $v^\star$ is a unit vector in $T_{b_0}S^2$. Base points are $b_i = \mathrm{Exp}_{b_0}(x_i)$ with x_i uniform on the tangent disk of radius 0.6, so $\Sigma_b = 0.09\, I_2$ for every κ. Each trial draws $n = 40$ base points and observations $y_i = P_{b_i \leftarrow b_0} v^\star + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma_v^2 I_2)$, $\sigma_v = 0.15$, in an orthonormal frame of $T_{b_i}S^2$; we run 10^5 trials for each of the ten curvatures $\kappa \in \{0.02, 0.05, 0.10, 0.21, 0.44, 0.50, 0.83, 1.00, 1.24, 1.56\}$ (seed 0). Parallel transport along the minimizing geodesic is $P_{q \leftarrow p}(w) = w + (w \cdot \hat{u})\big((\cos\theta - 1)\hat{u} - \sin\theta\, p/R\big)$ with $\hat{u} = \mathrm{Log}_p(q)/\|\mathrm{Log}_p(q)\|$ and $\theta = \|\mathrm{Log}_p(q)\|/R$. The bias is the norm of the mean error over trials and the MSE the mean squared error. The exact gap equation 31 is evaluated by quadrature over the radial law, and $\kappa^\star$ by bisection. The robust run is described in Appendix F; its outlier law differs from the two-Gaussian instance of equation 1.

Mechanism checks (Table 9). Loop holonomy is measured for geodesic triangles on the unit sphere whose vertices lie at distance $s \in \{0.10, 0.15, 0.22, 0.32, 0.45\}$ from the north pole, $120°$ apart; the areas in the table correspond to $s = 0.10, 0.22, 0.45$. The flat-connection control uses $n = 40$ base points uniform on the tangent disk of radius 0.25, fibre noise 0.05 and 400 trials, and reports the mean norm of the difference between the reference and transport-aware averages.

SPD study. Grayscale conversions of two natural color photographs, the test images `coffee` and `astronaut` distributed with scikit-image (van der Walt et al., 2014) (`skimage.data`; `https://scikit-image.org/docs/stable/api/skimage.data.html`), are resized to 72×72; structure tensors use $\rho = 1.0$, $\varepsilon = 10^{-3}$, scaled to unit median trace. We censor 40% of the interior pixels whose full 5×5 window lies inside the image ($68 \times 68 = 4624$ pixels, hence 1849 censored tensors) and reconstruct each from its observed neighbours in that window with inverse-distance weights. The affine-invariant Fréchet mean is computed by Riemannian gradient descent (15 iterations) using the affine-invariant logarithm and exponential maps; the Euclidean baseline uses the arithmetic mean. Errors are affine-invariant distances to the ground truth, averaged over the reconstructed tensors; swelling is $\det(\hat{J})/\det(J^\star)$.

Code. All studies can be reproduced from the anonymized repository `https://anonymous.4open.science/r/transport-aware-bundles-73F3/`, which contains the sampling laws, random seeds, field generators and encoder configuration. Percentages in Table 18 are computed from unrounded errors, so recomputing them from the rounded means can differ in the last digit (e.g. 18.7% instead of 18.9% for `coffee`).

U Additional analyses and implementation guidance

We collect supplementary baselines and implementation guidance.

Log-Euclidean baseline (SPD). We also evaluate the log-Euclidean mean $\exp\big(\sum_i w_i \log A_i\big)$ as a fast geometric baseline (Appendix S, Table 18). It matches affine-invariant Bundle-GEM within about 2.3% (coffee 0.268 vs. 0.274; astronaut 0.444 vs. 0.451) and shares its low determinant swelling (≈ 1.04). The gain over Euclidean is therefore attributable to geometric log-domain aggregation, not the affine-invariant metric specifically; neither SPD metric is required by the vector-bundle theory, and on this flat-base task log-Euclidean is an equally good, cheaper surrogate.

Safeguards when the Karcher–Kendall ball condition fails. For wide, multimodal, or near-cut-locus posteriors we recommend: (i) *detection*: compare the posterior diameter to the curvature-dependent Karcher radius and the injectivity radius, and flag violations; (ii) a *trust-region E-step*: cap each Karcher iteration's step and restrict aggregation to a geodesic ball of safe radius; (iii) *damping/tempering* of the fibre-mean update and the posterior; (iv) *multi-anchor or mixture aggregation*: cluster the posterior into unimodal components (e.g. geodesic k-means on the base) and aggregate per component before combining, which avoids averaging across a cut locus; and (v) *robust weighting* (the IRLS reweighting already used by our Bundle-GEM step) to downweight conflicting samples. Our controlled studies use geodesic neighborhoods within a safe radius; the mixture and trust-region variants are the recommended route for multimodal supports.

Encoder uncertainty: calibration and a curvature-aware regularizer. The gauge-equivariant encoder is trained with a Gaussian-NLL term, a proper scoring rule (Gneiting & Raftery, 2007), so it emits a per-vertex posterior precision. On held-out torus fields, the reported uncertainty diagnostics are: the ratio of actual RMSE to predicted standard deviation is 1.06 (ideal 1), the rank correlation between predicted std and actual error is 0.34 (a descriptive association without an uncertainty interval), and a three-bin reliability check tracks closely (predicted std $0.086/0.105/0.150 \to$ actual RMSE $0.090/0.107/0.168$). The theory also suggests a *reference-risk diagnostic*: for the fixed ambient reference used on this torus, $|K| \times \text{dispersion} \times \|v\|$ is an empirical proxy for aggregation error. Regions with a larger proxy tend to receive lower precision, an empirical association rather than a general law. We also tested a *transport-consistency regularizer* $\lambda \sum_{(i,j)} \|z_i - e^{i\theta_{ji}} z_j\|^2$, a penalty that softly enforces the connection, on the no-transport encoder: a modest $\lambda = 0.3$ lowers its error ($0.165 \to 0.162$ at 50% masking, moving toward the gauge encoder's 0.146), while an overly large $\lambda = 1.0$ hurts (0.188). The penalty helps somewhat, but the gauge-equivariant architecture, which enforces equivariance by construction instead of pulling neighbouring features together, works better in this test. The benefit comes from the inductive bias of exact equivariance.

V Additional experiments

This appendix gives the protocols and full results of the experiments added to test the complete algorithm, the diagnostics, topology, the torus mechanism and real data.

V.1 Full Bundle-GEM on incomplete data

Model and data. The base is a sphere of radius R, with $\kappa = 1/R^2 \in \{0.25, 1, 2.25\}$. Two components have anchors at geodesic distance 0.45 on either side of the north pole ("pole") or of a point on the equator ("equator"), base bandwidth $\beta = 0.3$, fibre means $(0.8, 0.3)$ and $(-0.2, 0.9)$ in the anchor frames, covariances with standard deviations 0.15 and 0.07 along rotated axes, and equal weights. Base points are drawn exactly from $\exp(-d(b,m)^2/2\beta^2)$ with respect to the volume of the sphere, by rejection. Fibre values are generated by the transported model equation 1. A fraction $q \in \{0.2, 0.5, 0.8\}$ of the fibre values is missing, 30% of the rest is observed through one random linear component, and the remainder in full, all with Gaussian noise of known standard deviation $\sigma \in \{0.05, 0.1, 0.2\}$; $n = 400$.

Methods. All three EM methods fit a two-component mixture with the same base part and differ only in how the fibre at b is identified with the anchor fibre: parallel transport (Bundle-GEM), ambient projection onto the anchor tangent plane (with the Jacobian of the projection in the density), or the spherical-coordinate frame, which is global but singular at the poles. The observation model is linear-Gaussian in the fibre coordinates, so the E-step is exact for every method. The M-step updates (π, μ, Σ) in closed form with eigenvalues clipped to $[10^{-4}, 10]$, updates β_k by a one-dimensional

maximization using the exact normalizer, and moves each anchor by an Armijo step along the gradient of the full Q with the posterior held fixed, carrying (μ_k, Σ_k) with the method's own identification (the transported retraction for Bundle-GEM). Runs stop when the log-likelihood increment falls below 10^{-3} or after 150 iterations, from a common initialization by geodesic k-means on the base. One-shot transport imputes a missing fibre by the transported mean of fully observed neighbours within a radius chosen by leave-one-out on the observed points. The oracle uses the true parameters in the exact E-step.

Results. Table 20 lists all 20 settings. The observed log-likelihood never decreased in any run of any method. Besides the imputation error, we report the error of the fitted conditional mean field $b \mapsto \mathbb{E}[v \mid b]$, averaged over the true base law, which does not depend on the observation pattern: Bundle-GEM is between 0.022 and 0.051 everywhere, ambient EM is close at $\kappa = 0.25$, about 0.03 worse at $\kappa = 1$ and 0.16–0.21 worse at $\kappa = 2.25$, and the coordinate frame is about 0.4 worse whenever the data straddle a pole. The heavy tails of ambient EM at $\kappa = 2.25$ come from data near geodesic distance $\pi R/2$ from an anchor, where projection loses rank.

Table 20: Full Bundle-GEM study, all settings (20 seeds each). Left: RMSE of imputed missing fibres for Bundle-GEM (T), ambient-reference EM (A), coordinate-frame EM (C), one-shot transport (1S) and the true parameters (O). Right: error of the fitted mean field with two standard errors, and the mean number of iterations of Bundle-GEM.

data	κ	missing	σ	T	A	C	1S	O	field T	field A	field C	iters
pole	0.25	20%	0.10	0.305	0.306	0.570	0.320	0.304	0.023±0.003	0.027±0.003	0.421±0.003	64
pole	0.25	50%	0.10	0.306	0.307	0.572	0.323	0.304	0.026±0.005	0.030±0.005	0.421±0.004	97
pole	0.25	80%	0.10	0.307	0.308	0.579	0.358	0.303	0.040±0.007	0.043±0.007	0.429±0.007	142
pole	1.00	20%	0.10	0.320	0.335	0.567	0.335	0.320	0.024±0.002	0.059±0.003	0.400±0.003	59
pole	1.00	50%	0.10	0.316	0.326	0.566	0.337	0.315	0.027±0.003	0.063±0.004	0.401±0.004	96
pole	1.00	80%	0.10	0.314	0.356	0.569	0.357	0.310	0.042±0.007	0.069±0.005	0.409±0.006	146
pole	2.25	20%	0.10	0.345	2.45	0.530	0.363	0.344	0.029±0.003	0.204±0.014	0.346±0.004	64
pole	2.25	50%	0.10	0.357	1.97	0.543	0.383	0.353	0.034±0.004	0.193±0.020	0.348±0.004	103
pole	2.25	80%	0.10	0.357	1.67	0.547	0.430	0.351	0.051±0.009	0.235±0.085	0.356±0.007	143
equator	0.25	20%	0.10	0.329	0.329	0.329	0.344	0.327	0.022±0.001	0.025±0.002	0.023±0.001	66
equator	0.25	50%	0.10	0.322	0.322	0.322	0.339	0.318	0.027±0.004	0.031±0.004	0.028±0.004	98
equator	0.25	80%	0.10	0.323	0.324	0.324	0.353	0.318	0.045±0.007	0.047±0.007	0.045±0.007	142
equator	1.00	20%	0.10	0.318	0.329	0.320	0.338	0.317	0.023±0.002	0.055±0.002	0.036±0.002	66
equator	1.00	50%	0.10	0.320	0.330	0.322	0.348	0.319	0.027±0.003	0.058±0.002	0.039±0.003	97
equator	1.00	80%	0.10	0.324	0.333	0.326	0.361	0.318	0.048±0.008	0.072±0.007	0.055±0.007	141
equator	2.25	20%	0.10	0.319	1.74	0.332	0.342	0.316	0.023±0.002	0.233±0.069	0.064±0.002	67
equator	2.25	50%	0.10	0.324	2.14	0.337	0.360	0.321	0.029±0.004	0.186±0.018	0.068±0.003	99
equator	2.25	80%	0.10	0.329	1.83	0.341	0.387	0.323	0.047±0.009	0.212±0.030	0.080±0.007	143
pole	1.00	50%	0.05	0.316	0.343	0.566	0.335	0.315	0.025±0.003	0.061±0.004	0.401±0.004	38
pole	1.00	50%	0.20	0.317	0.326	0.566	0.341	0.315	0.034±0.004	0.066±0.004	0.401±0.004	148

V.2 Validating the diagnostics

Fields. Around the north pole b_0 of a sphere with curvature $\kappa \in \{0.1, 0.25, 0.5, 1, 1.5\}$ we set $V(b) = P_{b \leftarrow b_0}\big(v^\star + \frac{a}{2}\|x\|^2 e\big)$ with $x = \mathrm{Log}_{b_0}(b)$, $\|v^\star\| = 1$ and e a unit vector at angle $\theta \in \{0, \pi/2, \pi\}$ to $v^\star$. Along radial geodesics this field has symmetric covariant Hessian $a\, I \otimes e$ at b_0, so $L_2 = |a|$, $\mathrm{Bias}_{\mathrm{mod}} = a\sigma_b^2 e$ and $\Xi = \kappa/|a|$; we use $\Xi \in \{0.25, 0.5, 1, 2, 4, 8, 16\}$. Observations follow the protocol of the crossover study ($n = 40$, $\sigma_v = 0.15$, base points uniform on the tangent disk of radius 0.6), and each setting uses 5×10^4 paired trials.

Diagnostics. From an independent calibration sample of 8000 points of the same field (for example 200 training fields of 40 points), a least-squares fit of the transported observations on $(1, x, \frac{1}{2}xx^\top)$ gives $\hat{v}^\star$, the symmetric Hessian, a noise-corrected $\|\hat{v}^\star\|$ and $\widehat{\Xi}$. With the known geometry of the reference, the same fit gives a plug-in estimate of every term of the general-field identity equation 18 at the target sample size $n = 40$; the decision rule picks transport when this estimate is positive. With 2000 calibration points the estimate was too noisy to be useful.

Results. In all 30 settings with $\Xi > 4$, transport has the smaller squared bias (Corollary 2). In all 15 settings with $\Xi \leq 1$ and $\theta = 0$ the reference has the smaller bias, as the alignment condition

predicts, and in all 15 with $\Xi \le 1$ and $\theta = \pi$ transport has. The leading bias formula correlates with the measured bias difference at 0.79; the remainder is higher-order. The MSE ranking is resolved (difference larger than two standard errors) in 102 of the 105 settings, and transport is better in 72 of them. Table 21 compares decision rules on those 102 settings; the excess MSE is relative to the better of the two estimators. The estimate $\widehat{\Xi}$ is accurate for $\Xi \le 2$ (median ratio to the truth 0.91–0.99) but biased low for large Ξ (0.59 at 8, 0.38 at 16), because the noise in the fitted Hessian inflates the estimate of L_2 when the true Hessian is small.

Table 21: Choosing between transport and the ambient reference in the 105 controlled settings.

Decision rule	correct	mean excess MSE	worst excess MSE
always transport	70.6%	18.5%	388%
always ambient reference	29.4%	52.4%	447%
$\widehat{\Xi} > 4$	48.7%	20.4%	361%
$\widehat{\Xi} > 1$	69.6%	12.5%	388%
leading formulas with true parameters	92.2%	0.5%	17.9%
plug-in estimate of the exact identity	93.8%	0.1%	18.9%

V.3 A SURFACE WITHOUT A GLOBAL FRAME

We use an icosphere of radius 1 with 2562 vertices. Each field is $V = \operatorname{grad} f + n \times \operatorname{grad} h$ with f, h random quadratic polynomials restricted to the sphere, normalized to unit root-mean-square norm; such a field is smooth, not parallel, and vanishes somewhere, as it must. Tangent Gaussian noise with $\sigma = 0.05$ is added and 50% of the vertices are masked. Each masked value is the mean of the observed values within geodesic radius ρ, after identification by exact parallel transport, ambient projection, a single spherical-coordinate frame (singular at $\pm z$), a two-chart atlas (spherical frames with poles at $\pm z$ and $\pm x$; each target uses the chart whose poles are farther away), or the normal-coordinate frame of the target. Table 22 reports the RMSE over 20 fields, overall and for targets within colatitude 0.3 of the z-poles. Two standard errors of the paired differences to transport are at most 0.0008 for ambient projection, the atlas and the normal-coordinate frame, and 0.012 for the global frame.

Table 22: Genus-zero study: RMSE of reconstructed masked values (20 fields).

ρ	targets	transport	ambient	global frame	two-chart atlas	normal coordinates
0.15	all targets	0.0562	0.0568	0.0807	0.0573	0.0561
0.15	colatitude < 0.3	0.0573	0.0576	0.2430	0.0577	0.0572
0.25	all targets	0.0555	0.0595	0.0985	0.0575	0.0544
0.25	colatitude < 0.3	0.0532	0.0562	0.3372	0.0536	0.0523
0.35	all targets	0.0772	0.0880	0.1433	0.0814	0.0739
0.35	colatitude < 0.3	0.0717	0.0802	0.4713	0.0725	0.0691

V.4 MECHANISM OF THE TORUS IMPUTATION RESULTS

We reimplemented the torus imputation task of Appendix Q: the same mesh and discrete connection, triangulated with one diagonal per grid cell, a field obtained by transporting a random seed vector along the Dijkstra shortest-path tree from a random root, tangent noise with $\sigma = 0.06$, and 50% of the vertices masked. Each masked value is the mean of observed values within geodesic radius ρ after identification by transport along the Dijkstra path to the target (one-shot), transport along the tree paths that generated the field (neighbour to root to target), the global coordinate frame, or ambient projection. We also use a smooth field whose angle in the coordinate frame is $\sin u + \frac{1}{2}\cos v$. Table 23 reports the RMSE over 20 seeds; two standard errors are at most 0.016 for the tree field and 0.001 for the smooth field.

For the tree field, transport along the generating paths reaches the noise floor, and the three local rules have the same, much larger error. The one-shot error correlates at 0.95 with the mean holonomy of

the loops formed by the local path and the two tree paths (0.08–0.17 rad on average); targets with below-median holonomy have one-shot RMSE 0.04–0.07, those above it 0.26–0.29. The field is therefore not smooth with respect to the connection across the branches of the tree, and every local rule pays for this. On the smooth field, one-shot transport improves as the window grows and stays within 0.003 of the coordinate frame, for which this particular field is a natural reference. In this reimplementation one-shot transport is not singled out as the worst rule, unlike in Table 14, whose implementation differs in its tree and path construction; in both, the error of one-shot transport comes from the test field and not from transport itself.

Table 23: Torus mechanism test: RMSE of reconstructed masked values (20 seeds).

field	ρ	one-shot transport	tree-path transport	coordinate frame	ambient
spanning-tree field	0.10	0.208	0.066	0.207	0.207
spanning-tree field	0.18	0.184	0.038	0.184	0.185
spanning-tree field	0.28	0.195	0.023	0.195	0.199
smooth field	0.10	0.081	–	0.078	0.081
smooth field	0.18	0.054	–	0.051	0.057
smooth field	0.28	0.045	–	0.042	0.058

V.5 Real wind fields

The winds are the NCEP/NCAR Reanalysis 1 (Kalnay et al., 1996) long-term monthly means at 300 hPa, provided by the NOAA Physical Sciences Laboratory (`https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html`), in the files `uwnd_mean.nc` and `vwnd_mean.nc` distributed with the `windspharm` package (Dawson, 2016) (`https://github.com/ajdawson/windspharm`, directory `windspharm/examples/example_data`); the code downloads them automatically. Table 24 lists the protocol, Tables 25 and 26 the reconstruction results, Table 27 the diagnostics, and Figure 6 shows them graphically.

Table 24: Real-wind study: data and protocol.

Item	Setting
Data	12 monthly climatologies of (u, v) on a global 2.5° grid, converted to ambient vectors
Mesh	level-3 icosphere, 642 vertices; bilinear interpolation and projection onto tangent planes
Scaling	by the RMS of the training months (21.2 m/s)
Training set	8 months, 64 random rotations each, $V'(p) = QV(Q^\top p)$
Test set	March, June, September, December; natural orientation and 16 random rotations each ($4 + 64$ fields)
Corruption	50% of vertices masked; tangent Gaussian noise with 5% of the field RMS on observed vertices
Metric	RMSE on masked vertices relative to the field RMS; "near frame poles" means $\lvert z\rvert > 0.95$
Networks	four message-passing layers over the 6 nearest vertices; inputs include mask and position
gauge-equivariant	16 complex type-1 channels in local frames, Levi-Civita transport phases, norm nonlinearities gated by 16 invariant channels (Appendix R)
no transport	the same network with all phases set to zero
standard	32 real channels on east/north or on ambient components (output projected)
Training	Adam, 3000 steps, batch 8, learning rate 3×10^{-3} with linear decay over the last 30%, 3 seeds
Local averages	mean of observed vertices within geodesic radius 0.15 after identification
Diagnostics	targets: 642 vertices; data: level-4 icosphere vertices within radius 0.35 (median 79, half observed); predictions from the 8 training months; measurement on the 4 held-out months with 50 masks and noise draws each

Architectures. All four networks read, at every vertex i, the observed wind (zero if masked), the mask bit o_i and the position $p_i \in S^2$, and aggregate over the six nearest vertices $\mathcal{N}(i)$ with uniform weights. The *gauge-equivariant* network expresses the wind as a complex number y_i in a local frame (e_1, e_2) of $T_{p_i}S^2$ and computes the connection phase θ_{ji} as the angle, in the frame at p_i, of $e_1(p_j)$ after exact Levi-Civita transport along the great circle from p_j to p_i. With 16 complex channels z_i and 16 invariant channels h_i initialized by $z_i = y_i w_{\mathrm{in}}$ and $h_i = \mathrm{ReLU}(H_{\mathrm{in}}[o_i, |y_i|, p_i])$, each of the four layers computes

$$m_i = \tfrac{1}{6}\sum_{j\in\mathcal{N}(i)} e^{\mathrm{i}\theta_{ji}} z_j, \qquad u_i = \big(Az_i + Bm_i\big) \odot \sigma\big(Gh_i + G'|m_i|\big), \qquad z_i \leftarrow \mathrm{nrelu}_b(u_i),$$
$$h_i \leftarrow \mathrm{ReLU}\big(H_1 h_i + H_2 \bar{h}_i + H_3|u_i| + c\big), \qquad \bar{h}_i = \tfrac{1}{6}\sum_{j\in\mathcal{N}(i)} h_j,$$

with complex matrices A, B, real matrices G, G', H_1, H_2, H_3 and moduli taken channel-wise. The output is $\hat{y}_i = w_{\mathrm{out}}^{\top} z_i$, mapped back to a tangent vector through the local frame. The *no-transport* ablation sets $\theta_{ji} = 0$. The two *standard* networks use 32 real channels, $x_i = \mathrm{ReLU}(W_{\mathrm{in}}[c_i, o_i, p_i])$ and four layers $x_i \leftarrow \mathrm{ReLU}(W_1 x_i + W_2 \bar{x}_i + b)$, where c_i is the wind in east/north components or in ambient components; the output is a linear map to the same components, projected onto the tangent plane in the ambient case. The networks have 9488 (gauge and no transport), 8576 (east/north) and 8640 (ambient) real parameters. Weights are Gaussian with variance $1/\text{fan-in}$ (complex weights with independent real and imaginary parts), and complex parameters are updated with the conjugate gradient in Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The loss is the mean squared error of the ambient wind vector over masked vertices. The code is `realdata/gauge_net.py`.

Table 25: Networks on the held-out months: relative masked RMSE, mean over 3 seeds, $\pm$ two standard errors over test fields. Δ is the paired difference to the gauge-equivariant network on the 64 rotated fields; the gauge network is better on every field.

Network	random orientation	near frame poles	natural orientation	Δ (rotated)	equivariance error	training time
Gauge-equivariant	$\mathbf{0.152} \pm 0.007$	$\mathbf{0.150} \pm 0.016$	$\mathbf{0.144} \pm 0.023$	–	2×10^{-7}	185 s
Standard, ambient components	0.172 ± 0.007	0.171 ± 0.018	0.164 ± 0.021	$+0.020 \pm 0.002$	–	35 s
Standard, east/north components	0.189 ± 0.007	0.366 ± 0.036	0.160 ± 0.020	$+0.038 \pm 0.005$	–	36 s
Gauge network without transport	0.237 ± 0.009	0.214 ± 0.017	0.200 ± 0.022	$+0.085 \pm 0.007$	0.9–1.2	182 s

Table 26: Non-learned local averages (radius 0.15): relative masked RMSE on the same test fields.

	random orientation		natural orientation	
Identification	all	near poles	all	near poles
Parallel transport	0.317	0.320	0.319	0.177
Ambient projection	0.317	0.320	0.319	0.177
East/north frame	0.350	0.598	0.313	0.174

Table 27: Diagnostics on real winds: transport versus ambient averaging at radius 0.35. A vertex is resolved when the measured difference exceeds two standard errors.

Quantity	Value
$\widehat{\Xi}$: median (interquartile range), maximum	0.08 (0.05–0.12), 0.39
Resolved vertices	597 of 642
Transport better (resolved vertices)	398 (67%)
Correct choices: plug-in identity	97.7%
Correct choices: always transport	66.7%
Correct choices: $\widehat{\Xi} > 1$ or $\widehat{\Xi} > 4$	33.3% (never selects transport)
Correlation of plug-in prediction and measured difference	0.99

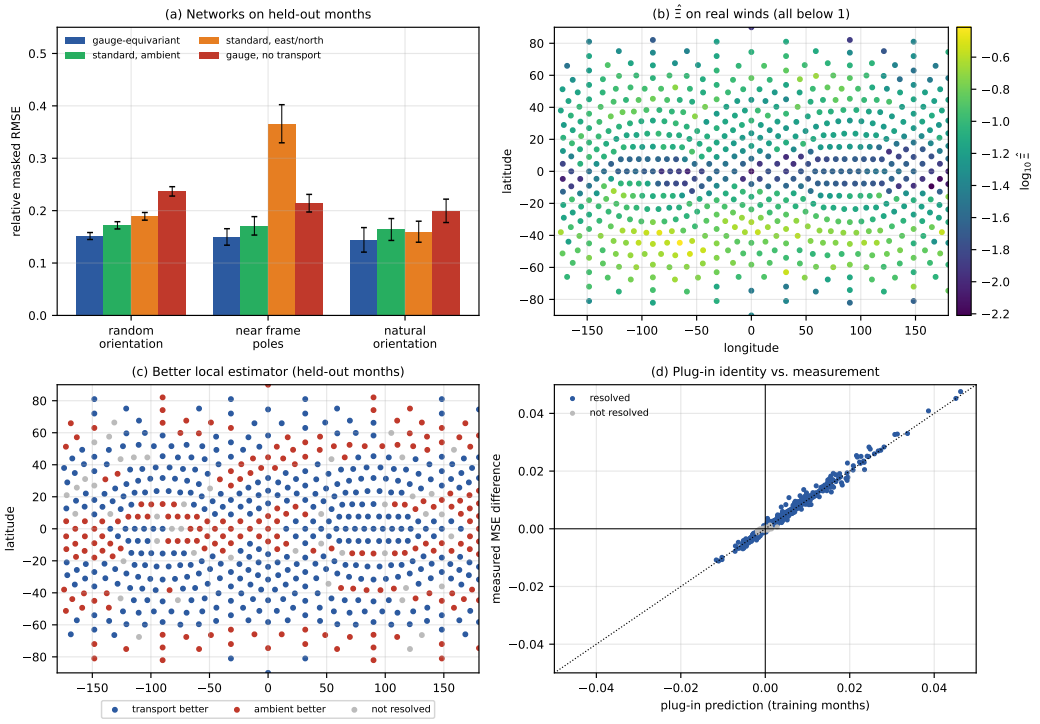

Figure 6: Real-wind study. (a) Masked RMSE of the four networks on held-out months (two standard errors over test fields). (b) $\widehat{\Xi}$ at each vertex, estimated from the training months: below 1 everywhere. (c) The better local estimator on the held-out months: transport and the ambient reference each win in large regions. (d) The plug-in estimate of the exact identity, computed from the training months, against the measured MSE difference on the held-out months.