REGULARIZED REGRESSION FOR HADAMARD MANIFOLD-VALUED DATA: FINITE-SAMPLE BOUNDS AND TANGENT-SPACE SPARSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Regression for Hadamard manifold-valued data requires predictions that account for intrinsic geometry while controlling model complexity. We represent regression curves as Riemannian exponential images of bounded tangent fields and combine an intrinsic loss with penalties on their coefficients or groups. For convex tangent models containing the ambient risk minimizer, we derive finite-sample oracle inequalities under geodesic strong convexity, Lipschitz continuity, a small-ball condition and localized complexity control. The bounds connect tangent-space error, integrated geodesic error and excess risk through an explicit curvature factor. Decomposable penalties express the regularization contribution in terms of subspace compatibility and approximation error, with a sparsity factor for exactly sparse representations. Simulations on the Poincaré ball and symmetric positive-definite matrices show a sparsity–prediction trade-off whose benefit diminishes for dense targets. An exploratory EEG covariance-interpolation study on 109 subjects gives lower mean held-out error than unpenalized fits, while constant and intrinsic regression baselines attain similar errors. These results identify the geometric and statistical conditions supporting tangent regularization and distinguish them from its observed numerical benefits.

1 INTRODUCTION

Covariance trajectories and hyperbolic representations are examples of data for which prediction takes place on a nonlinear space. Their geometry determines which distances and interpolations are meaningful, while finite observations require control of the predictor's complexity. Hadamard manifolds provide a useful setting for studying both requirements: geodesics are unique, the exponential map is globally invertible, and squared geodesic distance is convex along geodesics. Examples include hyperbolic space and symmetric positive-definite (SPD) matrices with the affine-invariant metric (Pennec, 2006; Bačák, 2014; Sala et al., 2018).

Intrinsic regression methods already address several aspects of this problem. Fréchet regression extends conditional-mean estimation to metric-valued responses (Petersen & Müller, 2019), geodesic regression fits a single geodesic (Fletcher, 2013), and total-variation regularization favors stepwise curves (Lin & Müller, 2021). Finite-sample bounds are also available for nonparametric metric-space regression (Schötz, 2022). Our question concerns a different combination: *how can a prescribed sparse tangent representation be combined with an intrinsic prediction loss, and which geometric and statistical conditions yield an oracle inequality for the resulting estimator?*

We anchor the representation at a reference point p and write a candidate curve as $\gamma = \mathsf{Exp}_p \circ v$. The tangent field v carries the coefficients or groups to be penalized, while the loss measures prediction on the manifold (Figure 1). This separates the representation from the prediction metric. It also creates two analytical difficulties. The exponential map distorts distances, and a restricted tangent model need not remain within itself after geodesic interpolation of its image curves. Our analysis builds on localized empirical-process and small-ball methods for regularized learning (Bartlett et al., 2005; Lecué & Mendelson, 2018; Alquier et al., 2019; Mendelson, 2015; 2018; 2021). Appendix A compares the settings and guarantees in detail.

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.

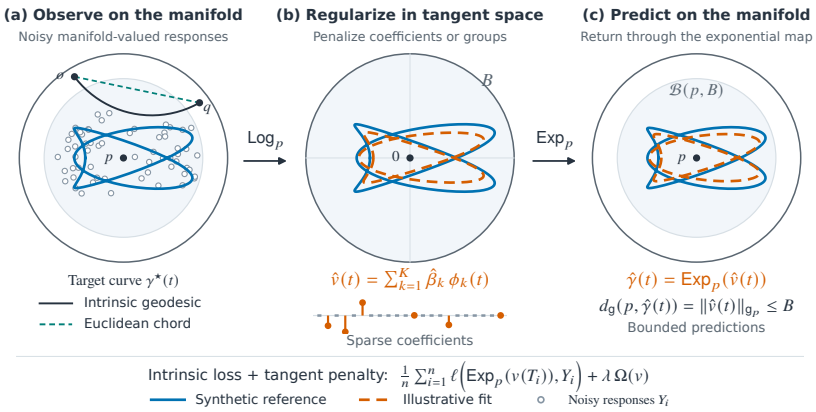


Figure 1: Illustrative schematic on the Poincaré disk. (a) A synthetic target curve and noisy responses. The geodesic and Euclidean chord between o and q illustrate the distinction between intrinsic and coordinate geometry. (b) Regularization acts on a fixed tangent representation while the loss is evaluated on the manifold. (c) The fitted field $\hat{v}$ maps to $\hat{\gamma} = \mathsf{Exp}_p \circ \hat{v}$, with $\hat{v}(t) = \sum_{k=1}^{K} \hat{\beta}_k \phi_k(t)$. Shading denotes the tangent B-ball and its image $\mathcal{B}(p, B)$. The curves and coefficients illustrate the construction and provide no empirical performance comparison or global optimality certificate.

Contributions. We summarize our main contributions in three parts:

(i) **a framework for intrinsic regression with representation penalties.** We connect regularized empirical risk minimization on tangent fields to intrinsic prediction through an explicit distance-comparison factor that transfers tangent-space control to geodesic error (Proposition 1).

(ii) **finite-sample analysis with separate geometric and statistical roles.** For convex models containing the ambient risk minimizer, a small-ball argument controls empirical tangent differences, and an auxiliary geodesic contraction class controls the first-order loss fluctuations. The resulting oracle inequality bounds tangent error, geodesic error and excess risk on the same event (Theorem 3). Decomposable penalties replace target-magnitude dependence in the regularization contribution by subspace compatibility and representation approximation error. The rate depends on the class used in the analysis. A realizable dictionary closed under geodesic interpolation admits an explicit K/n squared-distance bound, subject to its sample-size and constant requirements. The full localization radius remains in the distance bounds, so the sparsity-dependent penalty term does not establish an overall sparsity-adaptive rate.

(iii) **evidence for the sparsity–prediction trade-off.** Simulations and EEG covariance interpolation examine when regularization improves prediction and when tangent-loss, constant or other intrinsic estimators remain competitive.

We next define the estimator, state the assumptions and bounds, and evaluate the numerical trade-offs. The appendices give complete proofs, protocols and implementation limitations.

2 Tangent-space regularized regression

Let $(\mathcal{M}, \mathsf{g})$ be a finite-dimensional Hadamard manifold: complete, simply connected and of nonpositive sectional curvature. Write d_{g} for its geodesic distance and $\|\cdot\|_{\mathsf{g}_p}$ for the norm on the tangent space $\mathcal{T}_p\mathcal{M}$ for fixed $p \in \mathcal{M}$. The exponential map Exp_p is a diffeomorphism with inverse Log_p, and $\mathsf{d}_{\mathsf{g}}(p, \mathsf{Exp}_p v) = \|v\|_{\mathsf{g}_p}$ (Lee, 1997; Bridson & Haefliger, 1999). Additional geometric background is in Appendix B.

We observe n independent copies $Z_i = (T_i, Y_i)$ of a random pair with Borel law $\mathbb{P}$ on $[0, 1] \times \mathcal{M}$ and time marginal $\mathbb{P}_T$. Fix $p \in \mathcal{M}$ and $B > 0$ independently of the estimation sample. The ambient

tangent and curve classes are

$$\mathcal{V}_p^B = \{v : [0,1] \to \mathcal{T}_p\mathcal{M},\ \text{Borel},\ \|v(t)\|_{\mathsf{g}_p} \le B \quad \mathbb{P}_T\text{-a.e.}\} \text{ and } \Gamma_p^B = \{\mathsf{Exp}_p \circ v : v \in \mathcal{V}_p^B\}.$$

Viewed as equivalence classes modulo $\mathbb{P}_T$-almost-sure equality, the ambient tangent class $\mathcal{V}_p^B$ is convex and closed in $L_2(\mathbb{P}_T; \mathcal{T}_p\mathcal{M})$, and its curve image is geodesically convex. Population risks are invariant under changes on $\mathbb{P}_T$-null sets. Empirical evaluation uses fixed Borel representatives, subject to the standing regularity conditions below.

For a Borel loss $\ell(q, y)$, let $\ell_v(t, y) = \ell(\mathsf{Exp}_p(v(t)), y)$ and $\mathsf{R}(v) = \mathbb{E}\ell_v(T, Y)$. The ambient tangent and curve oracles are

$$v^\star \in \operatorname*{argmin}_{v \in \mathcal{V}_p^B} \mathsf{R}(v), \qquad \gamma^\star = \mathsf{Exp}_p \circ v^\star. \tag{1}$$

The assumptions below ensure finite risk and oracle existence and uniqueness up to $\mathbb{P}_T$-almost-sure equality (Appendix C.1). This is a bounded-class target in which equality with an unrestricted Bayes predictor is not assumed.

Estimation uses a fixed nonempty model $\mathcal{V} \subseteq \mathcal{V}_p^B$, and its curve image $\Gamma_\mathcal{V} = \{\mathsf{Exp}_p \circ v : v \in \mathcal{V}\}$. For a finite, nonnegative, convex penalty $\Omega : \mathcal{V} \to [0, \infty)$ and $\lambda \ge 0$, define

$$\hat{v} \equiv \hat{v}_n^\lambda \in \operatorname*{argmin}_{v \in \mathcal{V}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_v(T_i, Y_i) + \lambda\Omega(v) \right\}, \qquad \hat{\gamma} \equiv \hat{\gamma}_n^\lambda = \mathsf{Exp}_p \circ \hat{v}_n^\lambda. \tag{2}$$

We assume that, for every fixed $n \ge 1$ and $\lambda \ge 0$, the empirical criterion in (2) attains its minimum almost surely and admits a measurable global minimizer $\hat{v}_n^\lambda$, with jointly measurable predictions $(Z_1, \ldots, Z_n, t) \mapsto \hat{v}_n^\lambda(t)$ and measurable penalty $\Omega(\hat{v}_n^\lambda)$.

Standing regularity conditions. We impose the following conditions throughout the probabilistic analysis. The candidates in $\mathcal{V}$ and the chosen oracle $v^\star$ are fixed Borel representatives. There is a fixed $\mathbb{P}_T$-null set N, independent of the sample, such that

$$\|v(t)\|_{\mathsf{g}_p} \le B \quad \text{for all } v \in \mathcal{V} \cup \{v^\star\} \text{ and all } t \notin N.$$

The asserted convexity of the selected tangent model is understood on these representatives outside N. Tangent differences and geodesic contractions use the same representatives. Every localized empirical-process class used below, including the localized norm, cutoff, indicator, multiplier and ghost-sample classes, admits a countable reduction preserving its suprema almost surely. The corresponding localized Gaussian processes admit separable measurable versions.

For a fixed linearly independent dictionary $v = \sum_{k=1}^K \beta_k \phi_k$, the coefficient penalty is $\Omega(v) = \|\beta\|_1$. Group penalties select tangent directions, while cellwise and difference penalties favor zero cell values or few changes. These structures concern the prescribed representation: piecewise-constant tangent fields give piecewise-constant manifold curves. Appendix C.2 specifies the penalty domains and distinguishes norms from seminorms.

Sparsity of a manifold-valued curve is defined through its tangent representation, and the exponential map transfers it to γ. Sparsity is thus not a matter of "zeros on $\mathcal{M}$" but of a simple structure of the curve when viewed in $\mathcal{T}_p\mathcal{M}$. For instance, if $v(t) = 0$ on a time interval, then $\gamma(t) \equiv p$ there. If $v(t)$ is piecewise constant, γ is piecewise constant on the same cells; a nonconstant geodesic segment requires a different time parametrization. If only a few coordinates of $v(t)$ are nonzero, γ evolves in the image under Exp_p of a low-dimensional linear subspace $\mathcal{G} \subset \mathcal{T}_p\mathcal{M}$. The image curves γ are generated by tangent fields with many zeros, few active directions, or few change-points. The exponential image of a linear subspace need not be a geodesic submanifold; no such property is assumed.

We abbreviate $L_2(\mathbb{P}_T; \mathcal{T}_p\mathcal{M})$ to $L_2(\mathbb{P}_T)$ and assess estimation through the tangent-space error $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)}$, the integrated geodesic error defined by

$$\mathsf{D}_\mathsf{g}(\hat{\gamma}, \gamma^\star) = \left(\int_{[0,1]} \mathsf{d}_\mathsf{g}^2(\hat{\gamma}(t), \gamma^\star(t))\, \mathrm{d}\mathbb{P}_T(t) \right)^{1/2},$$

and the excess risk $\mathsf{R}(\hat{v}) - \mathsf{R}(v^\star)$. The tangent-space error and the integrated geodesic error are root-mean-square discrepancies over the time distribution $\mathbb{P}_T$, measured respectively in the fixed tangent space and on the manifold.

3 ASSUMPTIONS AND MAIN RESULTS

Full statements and proofs are provided in Appendices D–F. The model restrictions stated below form part of the hypotheses of the corresponding results.

3.1 ASSUMPTIONS

Assumption 1 (Bounded response set)**.** There exists a bounded, geodesically convex[1] Borel subset $\mathcal{Y} \subset \mathcal{M}$ such that $\mathbb{P}(Y \in \mathcal{Y}) = 1$.

Let $\mathcal{B}(p, B) = \{q : \mathsf{d}_{\mathsf{g}}(p, q) \le B\}$ (geodesically convex, Bridson & Haefliger, 1999, Prop. II.1.4) and let $\mathcal{D}_B = \mathrm{conv}(\mathcal{B}(p, B) \cup \mathcal{Y})$ be the geodesic convex hull of the prediction ball and the response set. Note that $\mathcal{D}_B$ is geodesically convex, bounded, and contains $\gamma(t)$ and Y for every $\gamma \in \Gamma_{\mathcal{V}}$.

Assumption 2 (Curvature lower bound on the prediction domain)**.** The sectional curvature of $\mathcal{M}$ restricted to $\mathcal{D}_B$ satisfies the two-sided bound $-\Lambda_B^2 \le \kappa(\mathcal{D}_B) \le 0$ for some finite constant $\Lambda_B \ge 0$.

The lower bound holds with $\Lambda_B = 0$ on $\mathbb{R}^d$, $\Lambda_B = 1$ on hyperbolic space $\mathbb{H}^d$ and $\Lambda_B = 1/\sqrt{2}$ globally on $\mathrm{SPD}(d)$ (Criscitiello & Boumal, 2023) (Appendix B).

Assumption 3 (Bounded In Expectation)**.** Assume that $\mathbb{E}_{(T,Y)\sim\mathbb{P}}|\ell(o, Y)| < \infty$ for some $o \in \mathcal{D}_B$.

Assumption 4 (Strongly geodesically convex loss)**.** For every $y \in \mathcal{Y}$, $\ell(\cdot, y)$ is continuously differentiable on an open neighborhood of $\mathcal{D}_B$, and there exists a constant $\mu > 0$, not depending on y, such that $\ell(\cdot, y) : \mathcal{D}_B \to \mathbb{R}$ is geodesically μ-strongly convex, i.e. for all $q_1, q_2 \in \mathcal{D}_B$, $\ell(q_2, y) \ge \ell(q_1, y) + \langle \mathsf{grad}\,\ell(\cdot, y)(q_1), \mathsf{Log}_{q_1}(q_2)\rangle_{\mathsf{g}_{q_1}} + \frac{\mu}{2}\mathsf{d}_{\mathsf{g}}^2(q_1, q_2)$.

Assumption 5 (Geodesic L-Lipschitz loss)**.** There exists a constant $L = L(B, \mathcal{Y}) < \infty$ such that for every $y \in \mathcal{Y}$ the function $\ell(\cdot, y) : \mathcal{D}_B \to \mathbb{R}$ is geodesically L-Lipschitz, i.e. for all $q_1, q_2 \in \mathcal{D}_B$, $|\ell(q_1, y) - \ell(q_2, y)| \le L\,\mathsf{d}_{\mathsf{g}}(q_1, q_2)$. The gradient is jointly Borel and satisfies $\|\mathsf{grad}\ell(\cdot, y)(q)\|_{\mathsf{g}_q} \le L$ on $\mathcal{D}_B$.

Assumption 3 makes the risk finite on the whole class. The squared geodesic loss $\ell(q, y) = \frac{1}{2}\mathsf{d}_{\mathsf{g}}^2(q, y)$, used in the main experiments, has $\mathsf{grad}\,\ell(\cdot, y)(q) = -\mathsf{Log}_q(y)$, is geodesically 1-strongly convex on *any* Hadamard manifold (the $\mathrm{CAT}(0)$ inequality), so $\mu = 1$, and is Lipschitz on $\mathcal{D}_B$ with $L \le \mathrm{diam}(\mathcal{D}_B)$ (Appendix D.3).

Assumption 6 (Convex model)**.** The fixed tangent class $\mathcal{V}$ is closed in $L_2(\mathbb{P}_T)$, convex and $v^\star \in \mathcal{V}$.

The last assumption is the small-ball condition of Mendelson (2015; 2021), on the difference class $\mathcal{H} := \mathcal{V} - \mathcal{V} = \{v - u : v, u \in \mathcal{V}\}$.

Assumption 7 (Small-ball condition)**.** There exists $\tau > 0$ such that $Q_{\mathcal{H}}(\tau) > 0$, where

$$Q_{\mathcal{H}}(\tau) = \inf_{h\in\mathcal{H}, \|h\|_{L_2(\mathbb{P}_T)}>0} \mathbb{P}\big[\|h(T)\|_{\mathsf{g}_p} \ge \tau\|h\|_{L_2(\mathbb{P}_T)}\big]\,.$$

If no difference has positive $L_2(\mathbb{P}_T)$ norm, set $Q_{\mathcal{H}}(\tau) = 1$.

3.2 NORM EQUIVALENCE AND EMPIRICAL LOWER BOUND

The estimator is penalized in $\mathcal{T}_p\mathcal{M}$ whereas the loss and the guarantees are stated through d_{g}. The hinge is the following comparison (Propositions G.3 and G.4), where the lower bound holds on any Hadamard manifold and the upper bound is a Jacobi-field comparison under Assumption 2.

Proposition 1 (Norm equivalence)**.** *Let $v, v^\star \in \mathcal{V}$ and set $\gamma = \mathsf{Exp}_p \circ v$, $\gamma^\star = \mathsf{Exp}_p \circ v^\star \in \Gamma_{\mathcal{V}}$. Under Assumption 2, there exists $\kappa_B = \sinh(\Lambda_B B)/(\Lambda_B B)$*

$$\|v - v^\star\|_{L_2(\mathbb{P}_T)}^2 \le \mathsf{D}_{\mathsf{g}}^2(\gamma, \gamma^\star) \le \kappa_B^2\,\|v - v^\star\|_{L_2(\mathbb{P}_T)}^2. \tag{3}$$

When $\Lambda_B = 0$, we set $\kappa_B = 1$, otherwise $\kappa_B > 1$ and $\kappa_B \sim e^{\Lambda_B B}/(2\Lambda_B B)$ for $\Lambda_B B \gg 1$. Appendix G gives the comparison proof.

[1] A subset of a Riemannian manifold is geodesically convex if any two of its points are joined by a unique shortest geodesic, and this geodesic is contained in the subset.

3.3 COMPLEXITY PARAMETERS

Our analysis builds on localized empirical-process and small-ball methods for regularized learning. We use localized Gaussian and Rademacher widths to control the empirical quadratic term, and a multiplier process to control first-order loss fluctuations (Bartlett et al., 2005; Lecué & Mendelson, 2018; Alquier et al., 2019; Mendelson, 2015).

Definition 1 (Gaussian width)**.** Let $\{G_f; f \in L_2(\mathbb{P}_T)\}$ be the canonical Gaussian process on the Hilbert space $L_2(\mathbb{P}_T)$, centered and isonormal, with covariance $\mathbb{E}[G_{f_1} G_{f_2}] = \langle f_1, f_2 \rangle_{L_2(\mathbb{P}_T)}$. For $\zeta_1 > 0$, the localized Gaussian fixed point of $\mathcal{H}$ is

$$\omega_G(\mathcal{H}, n, \zeta_1) = \inf\Big\{ r > 0 : \mathbb{E}\big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} G_h \big] \le \zeta_1 \, r\sqrt{n} \Big\},$$

where $r\mathcal{S}_{L_2(\mathbb{P}_T)} = \{h \in L_2(\mathbb{P}_T) : \|h\|_{L_2(\mathbb{P}_T)} = r\}$ denotes the sphere of radius $r > 0$ of $L_2(\mathbb{P}_T)$.

Definition 2 (Rademacher complexity)**.** Let $T_1, \dots, T_n$ be i.i.d. copies of $T \sim \mathbb{P}_T$, let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher variables (taking the values ± 1 with probability $1/2$, independent of the sample). For every $\zeta_2 > 0$, define

$$\omega_R(\mathcal{H}, n, \zeta_2) = \inf\Big\{ r > 0 : \mathbb{E}\Big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} \Big| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, \|h(T_i)\|_{\mathsf{g}_p} \Big| \Big] \le \zeta_2 r \Big\},$$

For both localized widths, the supremum over an empty sphere is defined to be zero. We set

$$\omega(\mathcal{H}, n, \zeta_1, \zeta_2) = \omega_Q(\zeta_1, \zeta_2) = \max\{\omega_G(\mathcal{H}, n, \zeta_1), \omega_R(\mathcal{H}, n, \zeta_2)\}.$$

The two widths measure the tangent difference class $\mathcal{H}$. For the multiplier process we use the geodesic contraction hull

$$\widetilde{\Gamma}_{\mathcal{V}} = \{\gamma_\theta, \gamma \in \Gamma_{\mathcal{V}}, \ \theta \in [0,1]\}, \ \text{where } \gamma_\theta(t) = \mathsf{Exp}_{\gamma^\star(t)}\big(\theta \mathsf{Log}_{\gamma^\star(t)} \gamma(t)\big).$$

The geodesic contraction is toward the oracle curve $\gamma^\star$, pointwise in time. The scalar θ is constant across time, and the hull is an auxiliary analysis class, that is stable under contraction, with $\mathsf{D}_{\mathsf{g}}(\gamma_\theta, \gamma^\star) = \theta \mathsf{D}_{\mathsf{g}}(\gamma, \gamma^\star)$ (see Lemma C.2).

Define

$$\phi_n^\ell(r) = \sup_{\{\gamma \in \widetilde{\Gamma}_{\mathcal{V}}; \mathsf{D}_{\mathsf{g}}(\gamma, \gamma^\star) \le r\}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \Big\langle \mathsf{grad}\, \ell(\cdot, Y_i)(\gamma^\star(T_i)), \mathsf{Log}_{\gamma^\star(T_i)}(\gamma(T_i)) \Big\rangle_{\mathsf{g}_{\gamma^\star(T_i)}} \Big|$$

For $\zeta_3 > 0$, $\delta \in (0,1)$, define,

$$\alpha_n^\ell(\zeta_3, \delta) = \max\Big\{ \inf\{r > 0 : \mathbb{P}[\phi_n^\ell(r) \le \zeta_3 r^2 \sqrt{n}] \ge 1 - \delta\}, \ \frac{L}{\sqrt{2}\, \zeta_3 \sqrt{n}} \Big\},$$

where L is the Lipschitz constant of Assumption 5. The radius $\alpha_n^\ell(\zeta_3, \delta)$ controls fluctuations of $\langle \mathsf{grad} \ell(\gamma^\star(T), Y), \mathsf{Log}_{\gamma^\star(T)} \gamma(T) \rangle$ over the contraction hull. For the positive-endpoint justification and the variance floor $r_0(\zeta_3) = L/(\sqrt{2}\zeta_3\sqrt{n})$, see Appendix D.5. A curve in this auxiliary hull need not be an admissible estimator.

3.4 THE ORACLE INEQUALITY

Write $\mathsf{R}_n \mathcal{L}_v = \mathsf{R}_n(v) - \mathsf{R}_n(v^\star)$ for the empirical excess risk. The probabilistic engine of the paper is the following lower bound, stated and proved as Theorem E.2 in Appendices E and I.

Theorem 2 (Lower bound on the empirical excess risk)**.** *Let Assumptions 1–7 hold. Fix $c_0 > 0$, $c_1 > 1$, $\delta \in (0,1)$, and choose*

$$0 < \eta < \min\Big\{1, \frac{\tau}{4\sqrt{2}c_0}, c_S^2\Big\}, \quad \zeta_1 = \frac{c_0 \eta^{3/2} Q_{\mathcal{H}}^2(\tau)}{256 c_1}, \quad \zeta_2 = \frac{c_0 \eta Q_{\mathcal{H}}(\tau)}{16\sqrt{2}}.$$

where $c_S > 0$ is a universal constant. Set $c_{\mathcal{A}}(\eta, \tau) = \min\{15/512, c_0^2 \eta^2 / \tau^2\}$. For a multiplier parameter $\zeta_3 > 0$ and a reference radius fixed before observing the sample, $\bar{r} > \omega_Q(\zeta_1, \zeta_2)$, define

$r^\star = \max\{\bar{r},\ \alpha_n^\ell(\zeta_3, \delta/4)\}$. *With probability at least* $1 - \delta - 5\exp\{-c_{\mathcal{A}}(\eta,\tau)Q^2_{\mathcal{H}}(\tau)n\}$, *the following inequality holds simultaneously for every* $v \in \mathcal{V}$ *satisfying* $\|v - v^\star\|_{L_2(\mathbb{P}_T)} > r^\star$*:*

$$\mathsf{R}_n\mathcal{L}_v \geq \left(\left(\frac{1}{4} - \frac{\sqrt{2}\,c_0\eta}{\tau}\right)\frac{Q_{\mathcal{H}}(\tau)\tau^2\mu}{8} - 4\zeta_3\kappa_B^2\right)\|v - v^\star\|^2_{L_2(\mathbb{P}_T)}. \tag{4}$$

Since $\hat{v}$ minimizes $\mathsf{R}_n(v) + \lambda\Omega(v)$, comparing with $v^\star$ gives the basic inequality $\mathsf{R}_n\mathcal{L}_{\hat{v}} \leq \lambda(\Omega(v^\star) - \Omega(\hat{v}))$, which combined with the lower bound (4) yields the main result.

Theorem 3 (Oracle inequality)**.** *Grant Assumptions 1–7, and make the standing choices of constants above. Set* $\varkappa_0 = (\frac{1}{4} - \sqrt{2}\,c_0\frac{\eta}{\tau})Q_{\mathcal{H}}(\tau)\frac{\tau^2\mu}{8} > 0$ *and choose* $\zeta_3 \in (0, \varkappa_0/(8\kappa_B^2)]$, $\varkappa = \varkappa_0 - 4\zeta_3\kappa_B^2$*. Then, for every* $\lambda \geq 0$*, any RERM estimator* $\hat{v}$ *of* (2) *satisfies*

$$\|\hat{v} - v^\star\|^2_{L_2(\mathbb{P}_T)} \leq (r^\star)^2 \ + \ \frac{\lambda}{\varkappa}\,\Omega(v^\star) \tag{5}$$

with probability at least $1 - \delta - 5\exp\big(-c_{\mathcal{A}}(\eta,\tau)Q^2_{\mathcal{H}}(\tau)n\big)$*. Moreover, on the same event, if in addition* $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} > r^\star$*, then also* $\varkappa\,\|\hat{v} - v^\star\|^2_{L_2(\mathbb{P}_T)} + \lambda\,\Omega(\hat{v}) \leq \lambda\,\Omega(v^\star)$*. Consequently, on the same event,*

$$\mathsf{D}^2_{\mathsf{g}}(\hat{\gamma}, \gamma^\star) \leq \kappa_B^2\Big((r^\star)^2 + \tfrac{\lambda}{\varkappa}\Omega(v^\star)\Big), \tag{6}$$

$$0 \leq \mathsf{R}(\hat{v}) - \mathsf{R}(v^\star) \leq L\,\kappa_B\left((r^\star)^2 + \tfrac{\lambda}{\varkappa}\Omega(v^\star)\right)^{1/2}. \tag{7}$$

Remark 1 (Rates and their scope). Theorem 3 separates the squared localization radius from the regularization cost $\lambda\Omega(v^\star)/\varkappa$. For a realizable bounded orthonormal dictionary (Assumption 8) whose curve image is also geodesically closed, Corollary F.9 bounds the localization term explicitly. With $n \geq K\max\{\zeta_1^{-2}, 32\zeta_2^{-2}\}$, $\zeta_3 = \varkappa_0/(8\kappa_B^2)$ and the reference radius specified in Corollary F.9,

$$\|\hat{v}_n^\lambda - v^\star\|^2_{L_2(\mathbb{P}_T)} \leq 8192\frac{L^2\kappa_B^6}{\delta^2\varkappa_0^2}\frac{K}{n} + \frac{\lambda\Omega(v^\star)}{\varkappa} \tag{8}$$

on the same confidence event. Corollary 4 refines the regularization contribution to $\lambda\sqrt{s}/\varkappa$ on the L_2 scale for an exactly s-sparse target. At $\lambda \asymp \sqrt{\log K/n}$, this gives a squared contribution of order $s\log K/n$, but the full localization term remains. Thus sparsity enters the bound without an overall sparsity-adaptive rate.

3.5 Decomposable penalties and dictionary guarantees

The term $\lambda\Omega(v^\star)/\varkappa$ involves the possibly large $\Omega(v^\star)$. Following Negahban et al. (2012) and Alquier et al. (2019), let $\mathcal{G}$ be a closed linear subspace of $L_2(\mathbb{P}_T)$ containing $\mathcal{V}$ on which the seminorm Ω is defined and finite, and v_M the orthogonal projection of v onto a closed subspace M. A pair $M \subseteq \bar{M} \subseteq \mathcal{G}$ of closed subspaces is a *decomposition pair* for Ω if $\Omega(v + w) = \Omega(v) + \Omega(w)$ for every $v \in M$, $w \in \bar{M}^\perp$, and $\Psi(\bar{M}) = \sup_{v\in\bar{M}\setminus\{0\}}\Omega(v)/\|v\|_{L_2(\mathbb{P}_T)}$ is the *subspace compatibility constant* (Definition F.1).

Theorem F.2 of the appendix then states: under the hypotheses of Theorem 3, for fixed $\lambda > 0$ and a decomposition pair $(M, \bar{M})$ with finite $\Psi(\bar{M})$, on the event $\mathcal{A}$ of (21),

$$\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} \leq \max\left\{r^\star,\ \ \frac{\lambda\,\Psi(\bar{M})}{\varkappa} + \sqrt{\frac{2\lambda\,\Omega(v^\star_{M^\perp})}{\varkappa}}\right\}, \tag{9}$$

whose second term simplifies to $\lambda\Psi(\bar{M})/\varkappa$ when $v^\star \in M$ exactly, while the maximum with $r^\star$ remains, and the geodesic and excess-risk bounds $\mathsf{D}^2_{\mathsf{g}}(\hat{\gamma}, \gamma^\star) \leq \kappa_B^2 r_\Omega^2$, $0 \leq \mathsf{R}(\hat{v}) - \mathsf{R}(v^\star) \leq L\kappa_B r_\Omega$ hold with r_Ω the right-hand side of (9). The proof inserts the decomposability inequality of Lemma F.1 into the basic inequality. We instantiate it for the dictionary regularization under one further assumption.

Assumption 8 (Orthonormal bounded dictionary)**.** Let $(\phi_k)_{k=1}^K \subset L_2(\mathbb{P}_T)$ functions satisfy $\mathbb{E}_{T\sim\mathbb{P}_T}\big[\langle\phi_j(T), \phi_k(T)\rangle_{\mathsf{g}_p}\big] = \delta_{jk}$, for $1 \leq j, k \leq K$, and $\sup_{k\leq K}\ \sup_{t\in[0,1]}\|\phi_k(t)\|_{\mathsf{g}_p} \leq M_\phi < \infty$, for a finite $M_\phi \geq 1$. Let $\mathcal{V}_p^{B,K} = \big\{\sum_{k=1}^K \beta_k\phi_k : \beta \in \mathbb{R}^K,\ \|\sum_k \beta_k\phi_k(t)\|_{\mathsf{g}_p} \leq B\ \forall t\big\}$, and take $E = \operatorname{span}\{\phi_k : k \leq K\}$, so $\mathcal{V} = \mathcal{V}_p^{B,K}$. The distance bound assumes $v^\star$ belongs to this model and uses its multiplier hull.

Orthonormality gives the isometry $\|v\|_{L_2(\mathbb{P}_T)} = \|\beta\|_2$ on $\mathcal{V}_p^{B,K}$ and $\Omega_{\text{lasso}}(v) = \|\beta\|_1$; if $v^\star$ is supported on S, $|S| = s$, the coordinate subspace $M = \bar{M} = \{\sum_{k\in S} \beta_k \phi_k\}$ is a decomposition pair with $\Psi(\bar{M}) = \sqrt{s}$. By the Paley–Zygmund inequality the small-ball condition holds on this class with $Q_{\mathcal{H}}(\tau) \geq (1-\tau^2)^2/C_4^4 \geq (1-\tau^2)^2/(KM_\phi^2) > 0$ for every $\tau \in (0,1)$, where $C_4 \leq (KM_\phi^2)^{1/4}$ is the L_4/L_2 equivalence constant of the dictionary span (Lemma F.3).

Corollary 4 (Sparse-Dictionary Oracle Inequality; Corollary F.4)**.** *Grant the hypotheses of Theorem 3 and Assumption 8 for the class $\mathcal{V}_p^{B,K}$ (so that $r^\star$ is the localization radius of that class), with $\Omega = \Omega_{\text{lasso}}$ and $v^\star$ exactly s-sparse in the dictionary $(\phi_k)_{k=1}^K$, and make the standing choices of Theorem 3. Then, for every $\lambda > 0$, on the event $\mathcal{A}$ of* (21)*, $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} \leq \max\{r^\star,\ \lambda\sqrt{s}/\varkappa\}$. In particular, with $\lambda = \lambda_n = 2LM_\phi\sqrt{2\log(2K)/n}$, the regularization term is $\frac{2LM_\phi\sqrt{s}}{\varkappa}\sqrt{\frac{2\log(2K)}{n}}$.*

The regularization term is of order $\sqrt{s\log K/n}$, as for the Euclidean LASSO (Alquier et al., 2019, Theorem 3.1), and depends on s rather than on $\|v^\star\|_{L_2(\mathbb{P}_T)}$ (Corollaries F.5 and F.6: Ω_{group}, Ω_{time}).

4 Experiments

We first isolate when a sparse tangent representation helps prediction, then evaluate covariance interpolation on EEG recordings. The comparisons use geodesic error throughout and compare penalties, representations and loss functions. The simulation slopes describe the computed fits; they do not certify the closure, multiplier-hull complexity or global optimization conditions of the distance theorem. Appendix J gives the protocols and full results and Appendix K reports structural comparisons, tuning limitations and solver cost.

4.1 Sparsity and prediction in controlled simulations

Design and comparisons. We generate curves on the Poincaré ball and $\mathrm{SPD}(6)$ by exponentiating tangent fields with s active Fourier coefficients. Each replication draws a fresh support and coefficient signs at fixed tangent L_2 norm a, i.i.d. uniform observation times, and isotropic tangent noise of root-mean-square magnitude σ at the generating curve. The base setting is $\mathbb{H}^2$, $n = 160$, $K = 22$, $s = 4$, $a = 0.5$, $\sigma = 0.25$ and $B = 6$. We vary one factor at a time over 22 settings, using 100 independent truth/data replications per setting. The estimators are geodesic ERM and ℓ_1-RERM, tangent least squares and LASSO, and an infeasible geodesic fit on the true support. Both regularized estimators select λ by four-fold cross-validation of held-out geodesic loss on the same grid. We report mean per-replication RMS geodesic error to the generating curve, with one sample standard deviation (SD) across replications; active coefficients satisfy $|\beta_k| > 10^{-3}$. The Fourier dictionary has $K = d(2n_{\text{freq}} + 1)$ coefficients on $\mathbb{H}^d$ and $K = 21(2n_{\text{freq}} + 1)$ on $\mathrm{SPD}(6)$. Fits use the accelerated proximal-gradient implementation in Appendix J.4.

Sample size and representation. Figure 2(a,b) compares sample sizes $n \in \{80, 160, 320, 640, 1280\}$ on $\mathbb{H}^2$ and $n \in \{80, 160, 320\}$ on $\mathrm{SPD}(6)$, with $K = 22$ and $K = 105$, respectively. The fitted mean-squared-error slopes are -0.98 and -1.00 for RERM versus -1.01 and -1.04 for ERM. Regularization lowers the error while leaving the fitted exponents similar: the mean-squared-error ratios RERM/ERM are 0.53–0.57 on $\mathbb{H}^2$ and 0.19–0.21 on $\mathrm{SPD}(6)$. Table 5 gives the individual values.

Adding inactive coordinates increases ERM error from 0.067 to 0.153 as K grows from 22 to 82, while RERM error changes from 0.048 to 0.058 (Figure 2(c)). Increasing the true support has the opposite effect: for $s = 2, 4, 8, 16$ with $K = 22$, the mean-squared-error ratios RERM/ERM are $0.33, 0.54, 0.73, 0.97$ (panel (d)). Prediction-tuned sparsity does not recover the support exactly: RERM retains 9.6 coordinates on average when $s = 4$. Tangent-space LASSO has lower mean error than geodesic RERM in all 22 settings, so these results support regularization rather than superiority of the intrinsic loss.

Figure 5 (Appendix K.1) compares Fourier RERM with TV-regularized Fréchet regression on sparse Fourier and four-segment piecewise-constant targets under the same observation mechanism and cross-validation rule. RERM has lower error than TV in all 50 replications of the former, whereas TV is better in 41 of 50 replications of the latter. The fitted curves show how a finite Fourier

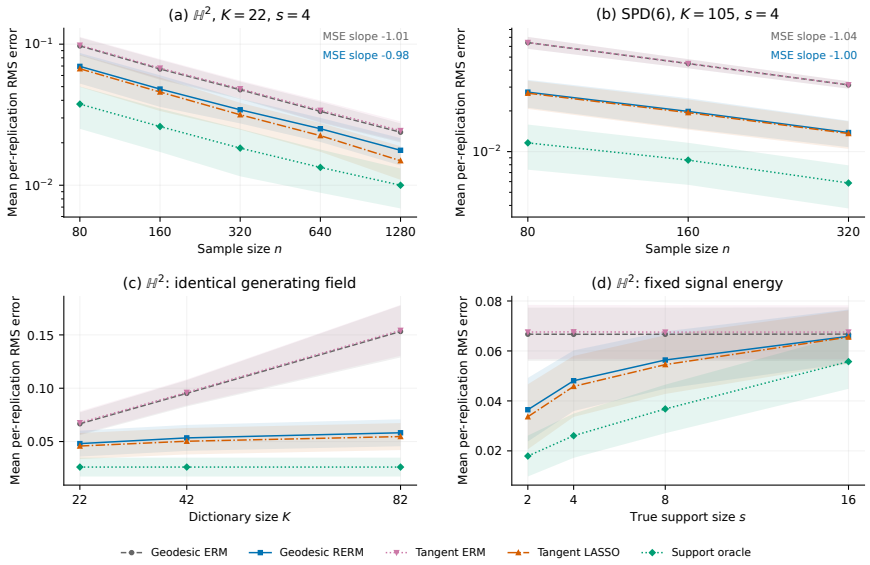


Figure 2: Sparsity and geodesic prediction with random generating curves: 100 paired truth/data replications per point; bands show ± one sample SD across replications. (a,b) Mean per-replication RMS geodesic error against sample size on $\mathbb{H}^2$ and $\mathrm{SPD}(6)$. Annotated slopes fit the *mean squared* error, not the plotted root error, and are descriptive. (c) The same $s = 4$ generating field embedded in larger dictionaries at $n = 160$. (d) Increasing support size at fixed signal energy on $\mathbb{H}^2$, $n = 160$. The support oracle uses information unavailable to the other estimators.

representation smooths across jumps, while TV represents separate levels. A curve can therefore have few changes without having few Fourier coefficients. This distinguishes coefficient sparsity from temporal piecewise constancy.

4.2 EEG COVARIANCE INTERPOLATION

Task and evaluation. We predict a missing central block of each covariance curve from the surrounding recording. The PhysioNet EEG Motor Movement/Imagery cohort (Goldberger et al., 2000; Schalk et al., 2004) supplies 109 subjects and 1 526 recordings. One-second windows on six sensorimotor channels give $\mathrm{SPD}(6)$ responses. The subjects are split $65/22/22$ for preprocessing, tuning and evaluation. Within every recording, the central 20% time block is held out; fitting windows that overlap its raw samples or a one-second buffer are purged. Each recording is fitted on its own remaining windows. Squared geodesic error is averaged over held-out windows, then over recordings within each subject, and finally over the 22 test subjects. We report means across those subjects, and this evaluates within-recording interpolation with tuning and evaluation on disjoint subjects.

Models and baselines. Fourier and cubic B-spline models each have 105 coefficients. The reference point is the Fréchet mean of each recording's fitting windows. Fourier intercepts are unpenalized, whereas all spline coefficients are penalized. Each model is compared with its unpenalized fit, the constant fitting-window Fréchet mean, and global and local-linear Fréchet regression. All use the same held-out blocks and the regularization parameter λ and local bandwidth are selected on validation subjects. The base-point-I Fourier fit is retained because it performed worse than the constant predictor on 21 of 22 test subjects. This comparison changes both the base point and intercept penalization, so it does not isolate their individual effects.

Results and interpretation. Table 1 shows lower mean error for both regularized models. The Fourier error falls from 3.296 to 2.564, a 22% aggregate reduction, with improvement on all 22 test subjects; the spline error falls from 2.883 to 2.560, improving on 21 of 22. The paired gains have mean ± SD 0.732 ± 0.624 and 0.323 ± 0.301, respectively. The selected representations are nearly

Table 1: EEG covariance interpolation: mean held-out squared geodesic error over 22 test subjects (308 recordings). Active: mean number of nonzero coefficients out of 105 for re-centred fits; the original fit uses the relative threshold in Appendix J.3. Validation selects $\lambda = 0.64$ for Fourier (upper grid endpoint), 0.16 for splines, and bandwidth $h = 0.6$ for local-linear regression. Fourier intercepts are unpenalized; all spline coefficients are penalized. Re-centred fits use the fitting-window Fréchet mean as base point. The final row retains the base-point-I specification as a sensitivity comparison. Bold and italics mark the lowest and second-lowest mean errors among the displayed methods, respectively.

Estimator	mean error $\pm$ SD	active
Fourier geodesic ERM (re-centred)	3.296 ± 1.304	105.0
Fourier geodesic RERM (re-centred)	2.564 ± 0.995	21.0
spline geodesic ERM (re-centred)	2.883 ± 1.126	105.0
spline geodesic RERM (re-centred)	$\mathbf{2.560 \pm 0.992}$	1.8
Fréchet mean of fitting windows (constant)	2.563 ± 0.995	—
global Fréchet regression	2.567 ± 1.000	—
local-linear Fréchet regression	$\textit{2.561} \pm \textit{0.997}$	—
Fourier geodesic RERM, base point I	2.781 ± 1.107	38.9

constant: Fourier fits have no non-constant coefficient in 306 of 308 test recordings, and spline fits retain 1.8 coefficients on average. Figure 4 gives the validation paths and per-subject summaries.

The constant, global and local-linear Fréchet estimators have mean errors close to the regularized fits. The close averages do not establish statistical equivalence or the absence of useful temporal structure; the SDs describe variation across subjects. The specification chronology cannot establish that all model choices preceded inspection of test results; these EEG comparisons are exploratory. Its estimated reference points, dependent windows and finite optimization fall outside the fixed-model distance guarantee, and $B = 40$ makes the curvature prefactors uninformative. Purging removes raw-sample overlap across the held-out boundary without making the remaining windows independent.

The active counts in Table 1 have different meanings for the two representations. Removing a nonconstant Fourier coefficient suppresses an oscillatory component, while the unpenalized intercept remains available. For the spline model, shrinking every coefficient to zero returns the fitting-window Fréchet mean through the exponential map. Thus 21 active Fourier coefficients can describe a constant curve, and zero active spline coefficients do not imply a zero covariance matrix. Counts are informative within a specified representation and penalization convention; they are not a common intrinsic measure of curve complexity across models.

5 Conclusion

Tangent representations provide a way to regularize manifold-valued regression while measuring prediction intrinsically. For fixed convex tangent models containing the ambient risk minimizer, the analysis yields finite-sample distance and excess-risk bounds under the stated geometric, small-ball and localized complexity assumptions. Decomposability refines the regularization term through subspace compatibility and representation approximation error; obtaining a fully sparsity-adaptive localization radius remains open. The experiments show why representation choice matters. Regularization helps sparse simulated targets, whereas TV can be preferable for piecewise-constant curves and tangent-loss estimators can outperform intrinsic fits. The exploratory EEG fits are nearly constant. These findings motivate evaluating each representation against simple intrinsic baselines.

Further theory should control multiplier complexity for practical dictionaries and accommodate estimated reference points and dependent observations. Transferring risk bounds to computed fits requires a quantified optimization gap and continuous feasibility control. Empirically, tasks with stronger temporal structure would help distinguish noise suppression from recovery of a changing manifold-valued signal.

REFERENCES

Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Annals of Statistics*, 47(4):2117–2144, 2019. doi: 10.1214/18-AOS1742.

Miroslav Bačák. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22 of *De Gruyter Series in Nonlinear Analysis and Applications*. De Gruyter, Berlin, 2014.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1999.

Christopher Criscitiello and Nicolas Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, 23(4):1433–1509, 2023. doi: 10.1007/s10208-022-09573-9.

Simone Fiori. Extended Hamiltonian learning on Riemannian manifolds: theoretical aspects. *IEEE Transactions on Neural Networks*, 22(5):687–700, 2011.

P. Thomas Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, 105(2):171–185, 2013. doi: 10.1007/s11263-012-0591-y.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12 (4):929–989, 1984. With discussion.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen C. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer-Verlag, Berlin, fifth edition, 2008.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. *Annals of Statistics*, 46(2):611–641, 2018. doi: 10.1214/17-AOS1562.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1991.

John M. Lee. *Riemannian Manifolds: An Introduction to Curvature*, volume 176 of *Graduate Texts in Mathematics*. Springer, 1997.

Zhenhua Lin and Hans-Georg Müller. Total variation regularized Fréchet regression for metric-space valued data. *Annals of Statistics*, 49(6):3510–3533, 2021. doi: 10.1214/21-AOS2095.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory (ALT 2016)*, volume 9925 of *Lecture Notes in Computer Science*, pp. 3–17. Springer, 2016. doi: 10.1007/978-3-319-46379-7_1.

Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21:1–21:25, 2015. doi: 10.1145/2699439.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1-2):459–502, 2018. doi: 10.1007/s00440-017-0784-y.

Shahar Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, 256(2): 147–167, 2021. doi: 10.4064/sm190420-21-11.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. doi: 10.1214/12-STS400.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. doi: 10.1007/s10851-006-6228-4.

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

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Christof Schötz. Nonparametric regression in nonstandard spaces. *Electronic Journal of Statistics*, 16(2), 2022. doi: 10.1214/22-EJS2056.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. Online optimization over Riemannian manifolds. *Journal of Machine Learning Research*, 24(84):1–67, 2023.

Reproducibility statement

The appendices contain the assumptions, complete proofs and retained experimental protocols. The numerical supplement contains stored fit records and scripts for rebuilding the figures and tables. This revision retains the existing experiments and corrects their mathematical and statistical presentation.

Ethics statement

The EEG experiment uses the publicly available, de-identified PhysioNet EEG Motor Movement/Imagery recordings (Goldberger et al., 2000; Schalk et al., 2004), distributed under the Open Data Commons Attribution License. No new human data were collected. The task is covariance interpolation and does not assess clinical utility. Raw recordings are not included in the manuscript source package.

AI use statement

Portions of the manuscript's text were refined with the assistance of a large language model (Claude Opus 5), and the code were developed with its assistance. All AI-assisted content and code were reviewed and verified by the authors, who remain solely responsible for any interpretations, results, and conclusions presented. The core methodology was developed without the use of any large language model.

SUPPLEMENTARY MATERIAL

The appendices first specify the model, loss and process assumptions, then give the oracle and dictionary bounds and their geometric and probabilistic proofs. The final appendices document the experimental protocols, complementary results, provenance requirements and limitations.

A RELATED WORK AND SCOPE OF COMPARISON

The proposed estimator combines intrinsic regression with a prescribed tangent representation. Table 2 separates this modeling choice from the assumptions and error criteria used in other regression guarantees.

Fréchet and geodesic regression. Global and local Fréchet regression extend weighted least squares to metric-valued responses (Petersen & Müller, 2019). Their consistency and convergence rates are asymptotic, with local theory for a scalar predictor. Schötz (2022) derives nonasymptotic bounds in a nonparametric setting. Geodesic regression by intrinsic least squares (Fletcher, 2013) provides a natural low-dimensional trajectory model. The present analysis instead concerns a regularized tangent representation and explicitly compares tangent and intrinsic errors.

Total-variation regularization. TV-regularized Fréchet regression favors stepwise trajectories and applies to Hadamard spaces, including nonsmooth examples beyond the manifolds considered here (Lin & Müller, 2021). Its fixed-design variation penalty and empirical design error differ from our representation penalty and random-time integrated error. The structural comparison in Appendix K.1 reflects this distinction: sparse Fourier and piecewise-constant truths favor different methods. The comparison does not rank the methods independently of the generating structure.

Table 2: Conceptual comparison of regression approaches. The rows describe different models and guarantees; they are not a ranking of predictive accuracy or a list of methods evaluated on every dataset.

Approach	Representation	Relevant scope
Fréchet regression (Petersen & Müller, 2019)	Weighted metric-space means	Asymptotic consistency and rates under objective-growth and regularity assumptions.
Nonparametric metric regression (Schötz, 2022)	Metric-valued regression	Finite-sample bounds in a nonparametric setting.
Geodesic regression (Fletcher, 2013)	A single geodesic	Intrinsic least squares with a fitted origin and velocity.
TV Fréchet regression (Lin & Müller, 2021)	Stepwise trajectories	Variation regularization, fixed-design error and Hadamard-space theory.
This paper	Exponential images of tangent fields	Finite-sample distance and excess-risk bounds for realizable convex tangent models; explicit dictionary rates under additional geodesic closure.

B GEOMETRIC BACKGROUND AND EXAMPLES

This appendix records the geometric conventions used in the estimator and the distance comparison. Detailed bounds for the exponential map are proved in Appendix G.

On a Riemannian manifold $(\mathcal{M}, \mathrm{g})$, the metric gives an inner product $\langle \cdot, \cdot \rangle_{\mathrm{g}_p}$ on each tangent space $\mathcal{T}_p\mathcal{M}$. The length of a piecewise continuously differentiable curve $\gamma : [0,1] \to \mathcal{M}$ is $\int_0^1 \|\dot\gamma(t)\|_{\mathrm{g}_{\gamma(t)}} \,\mathrm{d}t$, and the geodesic distance is the infimum of these lengths over curves with the prescribed endpoints. The exponential map sends an initial velocity $v \in \mathcal{T}_p\mathcal{M}$ to the time-one point of the geodesic starting at p with that velocity.

A finite-dimensional Hadamard manifold is complete, simply connected and has nonpositive sectional curvature. The Cartan–Hadamard theorem makes Exp_p a global diffeomorphism, with inverse Log_p. Any two points p, q are joined by a unique minimizing geodesic, parametrized on $[0, 1]$ as

$$\gamma(t) = \mathsf{Exp}_p(t\mathsf{Log}_p q), \qquad \mathsf{d}_\mathsf{g}(p, \gamma(t)) = t\,\mathsf{d}_\mathsf{g}(p, q).$$

Closed metric balls are geodesically convex. These facts justify the bounded prediction class and the pointwise contraction construction (Lee, 1997; Bridson & Haefliger, 1999).

Hyperbolic space. The Poincaré model of $\mathbb{H}^d$ is the open Euclidean unit ball with metric $\mathsf{g}_p = 4(1 - \|p\|^2)^{-2}\mathrm{Id}$ and sectional curvature -1. For $d = 2$ it is also called the Poincaré disk. Tangent coefficients in the experiments use an orthonormal frame for this metric, they should not be identified with unscaled Euclidean chart components at the origin.

Positive-definite matrices. For $P \in \mathrm{SPD}(d)$ and symmetric tangent matrices U, V, the affine-invariant metric is $\mathsf{g}_P(U, V) = \mathrm{tr}(P^{-1}UP^{-1}V)$. Its exponential and logarithm maps are

$$\mathsf{Exp}_P(U) = P^{1/2}\exp(P^{-1/2}UP^{-1/2})P^{1/2},$$
$$\mathsf{Log}_P(Q) = P^{1/2}\log(P^{-1/2}QP^{-1/2})P^{1/2}.$$

This is a Hadamard manifold (Pennec, 2006; Bačák, 2014). With the displayed metric normalization, the global sectional-curvature bound $-1/2 \le \kappa \le 0$ gives $\Lambda_B = 1/\sqrt{2}$ (Criscitiello & Boumal, 2023). The loss and distance formulas for both examples are given in Appendix D.

C Model classes and regularization

Let $(\mathcal{M}, \mathsf{g})$ be a finite-dimensional Hadamard manifold and $\mathcal{T} = [0, 1]$. The observations $Z_i = (T_i, Y_i)$ are i.i.d. with Borel law $\mathbb{P}$ and time marginal $\mathbb{P}_T$. The reference point p, radius $B > 0$, model and penalty are fixed independently of the observations used for estimation. A base point or model learned from those observations requires additional analysis; conditioning on it does not automatically preserve independence.

Ambient classes and selected tangent model. Define

$$\Gamma_p^B = \{\gamma : \mathcal{T} \to \mathcal{M} : \gamma \text{ measurable}, \ \mathsf{d}_\mathsf{g}(p, \gamma(t)) \le B \quad \mathbb{P}_T\text{-a.e.}\}, \tag{10}$$

and $\mathcal{V}_p^B = \{\mathsf{Log}_p \circ \gamma : \gamma \in \Gamma_p^B\}$. The identity $\|\mathsf{Log}_p(q)\|_{\mathsf{g}_p} = \mathsf{d}_\mathsf{g}(p, q)$ identifies $\mathcal{V}_p^B$ in $L_2(\mathbb{P}_T; \mathcal{T}_p\mathcal{M})$ with all fields bounded by B almost everywhere. It is convex and closed: convexity follows from the triangle inequality, and any L_2-convergent sequence has an almost-everywhere convergent subsequence preserving the bound. The ambient curve class is geodesically convex, since the prediction ball is geodesically convex and pointwise interpolation is continuous in its endpoints. Choose a nonempty tangent model $\mathcal{V} \subseteq \mathcal{V}_p^B$ and set

$$\Gamma_\mathcal{V} = \{\mathsf{Exp}_p \circ v : v \in \mathcal{V}\}, \qquad \mathcal{H} = \mathcal{V} - \mathcal{V}. \tag{11}$$

The full ambient class is used for population optimality; the restricted class is used for estimation and small-ball control. Using Assumption 6, $\mathcal{H}$ and, when $v^\star \in \mathcal{V}$, $\mathcal{V} - v^\star$ are star-shaped around zero. This supplies ordinary tangent rescaling; it does not assert geodesic closure of $\Gamma_\mathcal{V}$. Closedness of the ambient $\mathcal{V}_p^B$, rather than closedness of $\mathcal{V}$, is used for the ambient oracle.

Risk and ambient oracle. For the Borel loss ℓ, put $\ell_v(t, y) = \ell(\mathsf{Exp}_p(v(t)), y)$ and $\mathsf{R}(v) = \mathbb{E}\ell_v(T, Y) = \mathsf{R}(\mathsf{Exp}_p \circ v)$. Define

$$v^\star \in \operatorname*{argmin}_{v \in \mathcal{V}_p^B} \mathsf{R}(v), \qquad \gamma^\star = \mathsf{Exp}_p \circ v^\star. \tag{12}$$

The finite-risk, geometric and strongly convex loss assumptions yield existence and uniqueness up to $\mathbb{P}_T$-almost-sure equality (Lemma C.1). Fix one bounded Borel representative.

From Assumption 6, $v^\star \in \mathcal{V}$, this is a restriction on the pair consisting of the model and the data distribution. It makes the ambient oracle a feasible comparator for empirical minimization. It does not require knowledge of $v^\star$ to compute the estimator and does not follow from the dictionary, boundedness or the small-ball condition.

C.1 Existence and uniqueness of the ambient oracle

The oracle is defined over the full bounded ambient class. Its existence follows from a strongly convex midpoint argument and completeness; no small-ball condition is needed for this population result.

Remark C.1 (Properties of the ambient class). In $L_2(\mathbb{P}_T)$, $\mathcal{V}_p^B$ is nonempty, bounded, convex and closed, as shown in Appendix C. Its curve image is geodesically convex. No statistical small-ball assumption on this full class is used in the population argument.

Lemma C.1 (Ambient oracle existence and uniqueness)**.** *Grant Assumptions 1, 3, 2, 4 and 5. Then* R *attains its minimum on* $\mathcal{V}_p^B$*, uniquely up to* $\mathbb{P}_T$*-almost-sure equality. If Assumption 6 holds, this same oracle is the unique minimizer over* $\mathcal{V}$*, regardless of geodesic closure of* $\Gamma_{\mathcal{V}}$*.*

Proof. Write $\gamma_v = \mathsf{Exp}_p \circ v$. Integrability at $o \in \mathcal{D}_B$ and $|\ell(\mathsf{Exp}_p(v(T)), Y) - \ell(o, Y)| \le L(B + \mathsf{d}_{\mathsf{g}}(p, o))$ give uniformly finite risk above and below. Let $m = \inf_{v \in \mathcal{V}_p^B} \mathsf{R}(v) \in \mathbb{R}$. For $v, w \in \mathcal{V}_p^B$, the pointwise geodesic midpoint $\gamma_{1/2}$ belongs to Γ_p^B; hence $u = \mathsf{Log}_p \circ \gamma_{1/2} \in \mathcal{V}_p^B$. Strong convexity at the midpoint with the two endpoints cancels the linear terms:

$$\mathsf{R}(u) \le \tfrac{1}{2}\mathsf{R}(v) + \tfrac{1}{2}\mathsf{R}(w) - \tfrac{\mu}{8}\mathsf{D}_{\mathsf{g}}^2(\gamma_v, \gamma_w). \tag{13}$$

For a minimizing sequence (v_j) in the ambient class, the lower comparison gives

$$\frac{\mu}{8}\|v_j - v_k\|_{L_2}^2 \le \frac{\mu}{8}\mathsf{D}_{\mathsf{g}}^2(\gamma_{v_j}, \gamma_{v_k}) \le \tfrac{1}{2}(\mathsf{R}(v_j) - m) + \tfrac{1}{2}(\mathsf{R}(v_k) - m) \longrightarrow 0.$$

Completeness and closedness give a limit $v^\star \in \mathcal{V}_p^B$. A Borel representative exists: select a pointwise convergent subsequence on its full-measure Borel convergence set and set the field to zero elsewhere. The upper comparison and Cauchy–Schwarz imply

$$|\mathsf{R}(v) - \mathsf{R}(w)| \le L\mathsf{D}_{\mathsf{g}}(\gamma_v, \gamma_w) \le L\kappa_B\|v - w\|_{L_2},$$

so $\mathsf{R}(v^\star) = m$. Two distinct L_2 minimizers would have a midpoint of smaller risk by (13). If $v^\star \in \mathcal{V}$, the two infima are equal and ambient uniqueness implies uniqueness on $\mathcal{V}$.

□

Lemma C.2 (Structure of the contraction hull)**.** *For nonempty* $\mathcal{V} \subseteq \mathcal{V}_p^B$ *and* $\gamma^\star \in \Gamma_p^B$*,* $\gamma^\star \in \widetilde{\Gamma}_{\mathcal{V}}$ *and* $\Gamma_{\mathcal{V}} \subseteq \widetilde{\Gamma}_{\mathcal{V}} \subseteq \Gamma_p^B$*. The hull is stable under contraction toward* $\gamma^\star$*, and*

$$\mathsf{Log}_{\gamma^\star(t)}\gamma_\theta(t) = \theta\mathsf{Log}_{\gamma^\star(t)}\gamma(t), \qquad \mathsf{D}_{\mathsf{g}}(\gamma_\theta, \gamma^\star) = \theta\mathsf{D}_{\mathsf{g}}(\gamma, \gamma^\star).$$

The identities hold for the chosen representatives; ball containment is almost sure in time. The hull need not be geodesically convex between arbitrary pairs of its elements.

Proof. The choices $\theta = 0, 1$ give the first inclusions. Measurability follows from the continuous geodesic interpolation map, and convexity of $\mathcal{B}(p, B)$ gives the ambient bound. Uniqueness and constant speed of the geodesic from $\gamma^\star(t)$ give the logarithm identity and the pointwise distance scaling; squaring and integrating proves the D_{g} identity. Finally, $(\gamma_\theta)_{\theta'} = \gamma_{\theta\theta'}$, proving contraction stability. □

C.2 Penalties and their representation meaning

Sparse dictionary coefficients. Let $(\phi_k)_{k=1}^K$ be a fixed family of $\mathcal{T}_p\mathcal{M}$-valued basis functions on $[0, 1]$, and parameterize $v(t) = \sum_{k=1}^K \beta_k \phi_k(t)$, for $\beta \in \mathbb{R}^K$. A natural sparsity-inducing regularization is the ℓ_1-norm on the coefficients, $\Omega_{\text{lasso}}(v) = \sum_{k=1}^K |\beta_k|$. This encourages many coefficients β_k to vanish, so that v is represented using only a small subset of the basis functions. Consequently, the manifold curve

$$\gamma(t) = \mathsf{Exp}_p\Big(\sum_{k \in \Delta} \beta_k \phi_k(t)\Big)$$

is generated by a sparse set $\Delta \subset \{1, \dots, K\}$ of temporal "modes" around p. In geometric terms, γ is obtained by combining a few canonical deformation patterns encoded by the ϕ_k, leading to a low-complexity evolution on $\mathcal{M}$.

Time-local sparsity. Take the cell model of Proposition D.1, with m cells I_j of positive probability, and write $v(t) = a_j$ on I_j. Define $\Omega_{\mathrm{time}}(v) = \sum_{j=1}^{m} \|a_j\|_{\mathsf{g}_p}$. Zero cell coefficients give the constant prediction p on those cells. Any subsequent interpolation would define a different estimator and requires its own analysis.

Groups of tangent directions. Fix an orthonormal basis $\{e_1, \dots, e_d\}$ of $\mathcal{T}_p\mathcal{M}$ and write $v(t) = \sum_{k=1}^{d} v_k(t)\, e_k$, and $v_k \in L_2([0,1];\mathbb{R})$. A group-sparsity regularization across directions is given by $\Omega_{\mathrm{group}}(v) = \sum_{k=1}^{d} \|v_k\|_{L_2([0,1])}$. This encourages entire components $v_k(\cdot)$ to vanish. Hence, for many k we have $v_k(t) \equiv 0$ on $[0,1]$, and the tangent curve v takes values in a low-dimensional subspace $\mathrm{span}\{e_k : k \in \Delta\} \subset \mathcal{T}_p\mathcal{M}$ for some small index set Δ. The manifold curve $\gamma(t) = \mathsf{Exp}_p(v(t))$ then evolves in the image of a low-dimensional linear subspace through p, that is, in a sparse set of active geometric directions along the trajectory.

Sparse changes through discrete total variation. On ordered cells, set $\Omega_{\mathrm{TV}}(v) = \sum_{j=1}^{m-1} \|a_{j+1} - a_j\|_{\mathsf{g}_p}$. This encourages few changes in the tangent representation. Constant tangent segments map to constant manifold segments; the penalty does not imply nonconstant geodesic motion or change-point recovery.

Combined penalties. For fixed weights $\lambda_1, \lambda_2 \geq 0$, use $\Omega = \lambda_1\Omega_{\mathrm{time}} + \lambda_2\Omega_{\mathrm{TV}}$ on the same cell model. This is a seminorm and is a norm if $\lambda_1 > 0$. It encourages sparse cell values and sparse adjacent differences. The exponential map transfers these representation properties to a well-defined manifold-valued curve.

D Loss assumptions, small-ball control and localization

The loss and geometry assumptions specify a bounded domain on which the risk is finite. We then give a compatible model example and define the localized processes used in the oracle bound.

D.1 Geometry of the Prediction Domain

Geodesic convexity of the prediction domain $\mathcal{D}_B$. The ball $\mathcal{B}(p,B) = \{q \in \mathcal{M} : \mathsf{d}_{\mathsf{g}}(p,q) \leq B\}$, is geodesically convex on a Hadamard manifold (Bridson & Haefliger, 1999, Prop. II.1.4). Since $\mathcal{M}$ is uniquely geodesic, an arbitrary intersection of geodesically convex subsets of $\mathcal{M}$ is again geodesically convex. Consequently, $\mathcal{D}_B = \mathrm{conv}\big(\mathcal{B}(p,B) \cup \mathcal{Y}\big)$, the convex hull of the prediction ball and the response set, that is the intersection of all geodesically convex subsets containing $\mathcal{B}(p,B) \cup \mathcal{Y}$. It is bounded. Set $R = \max\{B, \sup_{y\in\mathcal{Y}} \mathsf{d}_{\mathsf{g}}(p,y)\} < \infty$. Since $\mathcal{B}(p,B) \cup \mathcal{Y} \subseteq \mathcal{B}(p,R)$ and the latter ball is geodesically convex, minimality of the hull gives

$$\mathcal{D}_B \subseteq \mathcal{B}(p,R), \qquad \mathrm{diam}(\mathcal{D}_B) \leq 2R < \infty. \tag{14}$$

No result below uses closedness or compactness of $\mathcal{D}_B$, only its geodesic convexity and the bound (14) are invoked.

Curvature Lower Bound on the Prediction Domain. The upper bound $\kappa(\mathcal{D}_B) \leq 0$ is automatic on a Hadamard manifold; a finite lower bound exists on the bounded set $\mathcal{D}_B$ by compactness of closed bounded balls and continuity of sectional curvature. We retain an explicit bound $-\Lambda_B^2$, satisfied by $\mathbb{R}^d$ with $\Lambda_B = 0$, by $\mathbb{H}^d$ with $\Lambda_B = 1$, and by $\mathrm{SPD}(d)$ with the affine-invariant metric *globally*, that is with no restriction to a bounded set, with $\Lambda_B = 1/\sqrt{2}$: the sectional curvature of that space lies in $[-1/2, 0]$ (Criscitiello & Boumal, 2023). The constant Λ_B depends on $\mathcal{D}_B$, hence on B and on $\mathcal{Y}$; the notation records only the dependence on B.

D.2 Assumptions on the Manifold Loss

Finite risk. By Bounded in expectation Assumption 3, $\mathbb{E}_{(T,Y)\sim\mathbb{P}}|\ell(o,Y)| < \infty$ for some $o \in \mathcal{D}_B$ Together with the Lipschitz assumption, Assumption 5 gives finite risk on the full ambient class. For $v \in \mathcal{V}_p^B$, almost surely,

$$|\ell(\mathsf{Exp}_p(v(T)), Y)| \leq |\ell(o,Y)| + L\big(\mathsf{d}_{\mathsf{g}}(o,p) + B\big).$$

The right-hand side is integrable and independent of v. Thus population and empirical excess risks are well defined.

Definition D.1 (Riemannian directional derivative)**.** For a continuously differentiable function $m : \mathcal{M} \to \mathbb{R}$, its differential at p is the linear map $F_p m = \mathrm{d}m_p : \mathcal{T}_p\mathcal{M} \to \mathbb{R}$ satisfying

$$F_p m(v) = \frac{\mathrm{d}}{\mathrm{d}t} m(\gamma(t))\bigg|_{t=0}, \qquad \gamma(0) = p, \quad \dot\gamma(0) = v.$$

The value is independent of the chosen C^1 curve with this initial velocity (Wang et al., 2023).

Definition D.2 (Riemannian gradient)**.** For each y, if $\ell(\cdot, y)$ is continuously differentiable, its gradient at q is the unique vector $\mathsf{grad}\ell(\cdot, y)(q) \in \mathcal{T}_q\mathcal{M}$ such that

$$\langle \mathsf{grad}\ell(\cdot, y)(q), w\rangle_{\mathsf{g}_q} = \mathrm{d}[\ell(\cdot, y)]_q(w), \qquad w \in \mathcal{T}_q\mathcal{M}.$$

The metric identifies the differential with this tangent vector (Fiori, 2011; Wang et al., 2023).

Geodesic μ-strong convexity of ℓ. Assumption 4 of geodesic μ-strong convexity reduces to geodesic convexity when $\mu = 0$. More generally, it can be viewed as a natural extension of the classical notion of convexity in Euclidean spaces, where the first-order lower-bound condition along straight lines is replaced by an analogous condition along geodesics on the manifold, with exponential maps playing the role of linear parametrizations. Because $\mathcal{D}_B$ is geodesically convex and bounded, Assumption 4 does not force superlinear growth of $\ell(\cdot, y)$ outside $\mathcal{D}_B$ and is therefore compatible with the Lipschitz Assumption 5.

We also require $\|\mathsf{grad}\ell(\cdot, y)(q)\|_{\mathsf{g}_q} \le L$ on $\mathcal{D}_B$, with the gradient jointly Borel in (q, y). The gradient bound implies the Lipschitz bound by integration along geodesics. We impose both explicitly, without relying on a converse at the boundary of the prediction domain. The constant L may depend on B, for the squared geodesic loss (see Appendix D.3), $L = \sup_{q\in\mathcal{D}_B,\, y\in\mathcal{Y}} \mathsf{d}_{\mathsf{g}}(q, y) \le \mathrm{diam}(\mathcal{D}_B)$, which is at most $2B$ when $\mathcal{Y} \subset \mathcal{B}(p, B)$.

D.3 Squared geodesic distance on standard manifolds

Squared distance satisfies both assumptions on the bounded domain.

Squared Geodesic Distance. $\ell(q, y) = \frac{1}{2}\mathsf{d}^2_{\mathsf{g}}(q, y)$ is Riemannian differentiable everywhere on $\mathcal{M}$, with $\mathsf{grad}\,\ell(\cdot, y)(q) = -\mathsf{Log}_q(y)$, and is geodesically 1-strongly convex on *any* Hadamard manifold. This is the CAT(0) comparison inequality (Bridson & Haefliger, 1999, Prop. II.1.4), not a property special to the Poincaré disk of Section J. Hence, Assumption 4 holds with $\mu = 1$ regardless of the ambient manifold. Its Lipschitz constant on $\mathcal{D}_B$ is $L = \sup_{q\in\mathcal{D}_B} \mathsf{d}_{\mathsf{g}}(q, y) \le \mathrm{diam}(\mathcal{D}_B)$, which is at most $2B$ when the response set $\mathcal{Y}$ of Assumption 1 is itself contained in $\mathcal{B}(p, B)$.

Augmented quadratic family. For a smooth, convex, nondecreasing scalar function $\varphi : [0, \infty) \to \mathbb{R}_+$ with $\varphi(0) = \varphi'(0) = 0$, the loss $\ell(q, y) = \frac{1}{2}\mathsf{d}^2_{\mathsf{g}}(q, y) + \varphi(\mathsf{d}_{\mathsf{g}}(q, y))$ remains geodesically 1-strongly convex. Convexity follows by composition with the convex distance function; $\varphi'(0) = 0$ ensures differentiability at $q = y$. A bounded derivative on the finite distance range gives the required Lipschitz and gradient bounds.

For squared distance, $\mu = 1$ on either experimental manifold. The formulas below specify the metric normalization and gradient used in the numerical method.

Hyperbolic space $\mathbb{H}^d$ (Poincaré ball model, $\kappa = -1$). Identifying $\mathbb{H}^d$ with the open unit ball $\{x \in \mathbb{R}^d : \|x\| < 1\}$ under the conformal metric $\mathsf{g}_x = \big(2/(1 - \|x\|^2)\big)^2 \mathrm{Id}$, the geodesic distance is

$$\mathsf{d}_{\mathsf{g}}(x, y) = \operatorname{arccosh}\left(1 + \frac{2\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}\right),$$

and at $p = 0$, $\mathsf{Log}_0(y) = 2\,\mathrm{artanh}(\|y\|)\, y/\|y\|$ in a g_0-orthonormal frame, with value zero at $y = 0$ by continuity. These are orthonormal-frame coordinates, not the unscaled Euclidean chart components. Thus $\mathsf{d}_{\mathsf{g}}(0, \mathsf{Exp}_0(u)) = \|u\|_{\mathsf{g}_0}$ exactly—the identity Lemma E.1 is built on. This is the manifold ($d = 2$) instantiated in Section 4's simulation (Figure 2), where $\Lambda_B = 1$, so $\kappa_B = \sinh(B)/B$ has the closed form of Proposition G.4.

$\mathbf{SPD}(d)$ with affine-invariant metric. With $\mathsf{g}_P(U,V) = \operatorname{tr}(P^{-1}UP^{-1}V)$ for symmetric tangent vectors U, V at P, the geodesic distance and the loss's gradient are

$$\mathsf{d}_{\mathsf{g}}(P,Q) = \big\| \log\big(P^{-1/2}QP^{-1/2}\big) \big\|_F,$$

and

$$\mathsf{grad}\, \ell(\cdot, Q)(P) = -\mathsf{Log}_P(Q) = -P^{1/2} \log\big(P^{-1/2}QP^{-1/2}\big) P^{1/2},$$

where $\log(\cdot)$ is the ordinary matrix logarithm; at the reference point $p = I$ these collapse to $\mathsf{d}_{\mathsf{g}}(I,Q) = \|\log Q\|_F$ and $\mathsf{grad}\, \ell(\cdot, Q)(I) = -\log Q$, exactly the identity the real-data pipeline of Section J is built on. Although its curvature is not constant, the affine-invariant SPD metric has the global bound $-1/2 \le \kappa \le 0$, so $\Lambda_B = 1/\sqrt{2}$ is available independently of the data (Section 2). The experiments use $d = 6$.

D.4 Small-Ball Condition

A single pair $(\tau, Q_{\mathcal{H}}(\tau))$ must serve *every* $h \in \mathcal{H}$ at once; positivity of $\mathbb{P}[\|h(T)\|_{\mathsf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}]$ for each fixed h is automatic for small τ and carries no information. The uniform version is restrictive, and in particular it *fails* for the class of all bounded measurable tangent fields as soon as $\mathbb{P}_T$ is nonatomic. Indeed, fix $\tau > 0$, a unit vector $u \in \mathcal{T}_p\mathcal{M}$ and $\varepsilon \in (0,1)$ with $\tau^2\varepsilon \le 1$, and use nonatomicity to pick a Borel set $A \subset \mathcal{T}$ with $\mathbb{P}_T(A) = \varepsilon$. For $h = c\,\mathbf{1}_A\, u$ with any $0 < c \le 2B$ one has $\|h\|_{L_2(\mathbb{P}_T)} = c\sqrt{\varepsilon}$ and $\|h(t)\|_{\mathsf{g}_p} = c\,\mathbf{1}_A(t)$, so that

$$\big\{\|h(T)\|_{\mathsf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}\big\} = \{T \in A\}, \qquad \mathbb{P}\big[\|h(T)\|_{\mathsf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}\big] = \varepsilon.$$

Letting $\varepsilon \downarrow 0$ gives $Q_{\mathcal{H}}(\tau) = 0$ for every $\tau > 0$. The amplitude c cancels, so the uniform bound B on the class does not help: the obstruction is concentration of the field on sets of small $\mathbb{P}_T$-measure, not its size. This counterexample concerns the full tangent representation of the ambient class Γ_p^B, not every selected $\Gamma_{\mathcal{V}} \subseteq \Gamma_p^B$.

The small-ball condition must be verified for the selected tangent class $\mathcal{V}$. Lemma F.3 below verifies it, with an explicit constant, for the finite-dimensional dictionary class of Assumption F.1, used in the dictionary specializations of Section F.

Proposition D.1 (Verified fixed-partition model)**.** *Fix a finite Borel partition $(I_j)_{j=1}^m$ of $[0,1]$, with boundary assignment fixed independently of the estimation sample, and $p_j = \mathbb{P}_T(I_j) > 0$, $p_{\min} = \min_j p_j$. Choose*

$$\Gamma_{\mathrm{part}} = \big\{\gamma : \gamma(t) = q_j \text{ for } t \in I_j,\ q_j \in \mathcal{B}(p,B)\big\} \text{ and } \mathcal{V}_{\mathrm{part}} = \Big\{ \sum_{j=1}^m u_j \mathbf{1}_{I_j} : \|u_j\|_{\mathsf{g}_p} \le B \Big\}.$$

For $\Gamma_{\mathcal{V}} = \Gamma_{\mathrm{part}}$ and $\mathcal{V} = \mathcal{V}_{\mathrm{part}}$, the small-ball condition holds with $\tau = 1$ and $Q_{\mathcal{H}}(1) \ge p_{\min}$. Under the finite-risk and strongly convex loss assumptions, the model-risk minimizer is unique and cellwise constant. Under the stated population loss and integrability assumptions, if, additionally, for each cell I_j there is a probability law ν_j such that

$$\mathcal{L}(Y \mid T = t) = \nu_j \quad \text{for } \mathbb{P}_T\text{-almost every } t \in I_j,$$

then the ambient oracle is cellwise constant and belongs to $\mathcal{V}_{\mathrm{part}}$ in tangent coordinates. Consequently, the realizability requirement of Assumption 6 holds.

Proof. The tangent coefficient set is a product of m radius-B balls, hence convex and compact. The squared L_2 norm is $\sum_j p_j \|u_j\|_{\mathsf{g}_p}^2$. With an orthonormal tangent basis (e_k), the fields $p_j^{-1/2}\mathbf{1}_{I_j} e_k$ form a bounded Borel L_2-orthonormal basis of dimension md. Geodesic interpolation remains constant on each cell and stays in the geodesically convex prediction ball, proving closure. For a nonzero difference with cell values h_j, at least one cell satisfies

$$\|h_j\|_{\mathsf{g}_p} \ge \Big(\sum_l p_l \|h_l\|_{\mathsf{g}_p}^2 \Big)^{1/2} = \|h\|_{L_2(\mathbb{P}_T)}.$$

The event $\{\|h(T)\|_{\mathsf{g}_p} \ge \|h\|_{L_2}\}$ therefore has probability at least $p_{\min}$. Finally, risk is $\sum_j p_j \mathbb{E}[\ell(\mathsf{Exp}_p(u_j), Y) \mid T \in I_j]$ and minimizes separately in each cell. Compactness and strong convexity give the unique model minimizer. Under the additional conditional-law hypothesis, let $r_j(q) = \mathbb{E}[\ell(q,Y) \mid T \in I_j]$. Its unique minimizer $q_j^\star \in \mathcal{B}(p,B)$ also minimizes conditional risk at almost every time in I_j. For any ambient curve, integrating $r_j(\gamma(t)) \ge r_j(q_j^\star)$ proves that the cellwise curve is the ambient oracle. □

D.5 Complexity parameters

Let $\mathscr{C} \subset L_2(\mathbb{P}_T; \mathcal{T}_p\mathcal{M})$ be star-shaped around 0 (i.e. $\theta f \in \mathscr{C}$ whenever $f \in \mathscr{C}$ and $\theta \in [0,1]$; Definition H.3) and let Φ be a functional on subsets of $L_2(\mathbb{P}_T; \mathcal{T}_p\mathcal{M})$ that is nonnegative, satisfies $\Phi(\varnothing) = 0$, and is positively homogeneous, in the sense that $\Phi(\theta A) = \theta\,\Phi(A)$ for $\theta > 0$ and $\Phi(A) \le \Phi(A')$ for $A \subset A'$.

Lemma D.2 (Upward Closure of the Fixed-Point Conditions)**.** *The function $r \mapsto \Phi(\mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)})/r$ is non-increasing on $(0,\infty)$. For $\zeta, c > 0$, the set $\{r > 0 : \Phi(\mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}) \le \zeta c r\}$ is empty or an interval unbounded above. When nonempty, the inequality holds for every radius strictly larger than its infimum.*

This applies to the Gaussian functional $\Phi(A) = \mathbb{E}\sup_{h\in A} G_h$, by linearity, and to

$$\Phi(A) = \mathbb{E}\sup_{h\in A}\left|\frac{1}{n}\sum_i \varepsilon_i \|h(T_i)\|_{\mathsf{g}_p}\right|.$$

The classes $\mathcal{H} = \mathcal{V} - \mathcal{V}$ and $\mathcal{H}_{v^\star} = \mathcal{V} - v^\star$ are convex and contain zero, so they are star-shaped.

Proof. For $0 < r \le r'$ and $h' \in \mathscr{C} \cap r'S(L_2)$, star-shapedness gives $(r/r')h' \in \mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, so $(r/r')\big(\mathscr{C} \cap r'S(L_2)\big) \subset \mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$ and, by homogeneity and monotonicity, $(r/r')\Phi(\mathscr{C} \cap r'S(L_2)) \le \Phi(\mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)})$, which is the claimed monotonicity of $\Phi(\mathscr{C} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)})/r$. If the inequality holds at some r_0 it therefore holds at every $r \ge r_0$; and if r exceeds the infimum ω of the set, there is $r_0 \in [\omega, r)$ in the set, so the inequality holds at r. □

Lemma D.3 (Localized Complexities are Finite)**.** *For every $n \ge 1$, $\zeta_1 > 0$ and $\zeta_2 > 0$, we have $\omega_R(\mathcal{H}, n, \zeta_2) \le 2B, \omega_G(\mathcal{H}, n, \zeta_1) \le 2B$ and hence $\omega(\mathcal{H}, n, \zeta_1, \zeta_2) \le 2B < \infty$.*

Proof. Every $h \in \mathcal{H} = \mathcal{V} - \mathcal{V}$ is of the form $h = v - u$ with $\|v(t)\|_{\mathsf{g}_p} \le B$ and $\|u(t)\|_{\mathsf{g}_p} \le B$ for $\mathbb{P}_T$-a.e. t, so $\|h(t)\|_{\mathsf{g}_p} \le 2B$ a.e., and therefore $\|h\|_{L_2(\mathbb{P}_T)} \le 2B$. Consequently, for every $r > 2B$ the index set $\mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)} = \{h \in \mathcal{H} : \|h\|_{L_2(\mathbb{P}_T)} = r\}$ is empty, both suprema in Definitions 1 and 2 equal 0, and both defining inequalities hold trivially (their right-hand sides $\zeta_2 r$ and $\zeta_1 r\sqrt{n}$ being positive). Hence each defining set contains $(2B, \infty)$ and each infimum is at most $2B$. □

Remark D.1 (What finite fixed points do and do not imply). The bound $\omega_Q \le 2B$ only reflects the diameter of the difference class. It does not ensure a nontrivial error bound or a positive confidence lower bound. On a bounded orthonormal dictionary, Proposition F.7 supplies a sufficient sample-size condition for $\omega_Q = 0$ using width upper bounds linear in the radius. The constraints need not make the widths exactly homogeneous. A statistical rate further requires the appropriate multiplier analysis on $\widetilde{\Gamma}_{\mathcal{V}}$.

The Gaussian and Rademacher widths control the design-dependent quadratic term. The first-order loss fluctuations require a different process, indexed by the contraction hull and using independent Rademacher signs:

$$\phi_n^\ell(r) = \sup_{\left\{\gamma\in\widetilde{\Gamma}_{\mathcal{V}};\mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star)\le r\right\}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\Big\langle \mathsf{grad}\,\ell(\cdot,Y_i)(\gamma^\star(T_i)), \mathsf{Log}_{\gamma^\star(T_i)}(\gamma(T_i))\Big\rangle_{\mathsf{g}_{\gamma^\star(T_i)}}\right| \tag{15}$$

which captures how the noise correlates with the localized deviations around $\gamma^\star(T_i)$. The radius in (15) is geodesic; it is compared with tangent-space radii through Lemma E.1. Define

$$\alpha_n^\ell(\zeta_3,\delta) = \max\left\{\inf\{r>0; \mathbb{P}[\phi_n^\ell(r) \le \zeta_3 r^2\sqrt{n}] \ge 1-\delta\},\ \frac{L}{\sqrt{2}\,\zeta_3\sqrt{n}}\right\},$$

where $\delta \in (0,1)$ and $\zeta_3 > 0$, and L is the Lipschitz constant of Assumption 5. Since $\mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star) \le 2B$ on $\widetilde{\Gamma}_{\mathcal{V}}$ and $\|\mathsf{grad}\,\ell\| \le L$, one has $\phi_n^\ell(r) \le 2LB\sqrt{n}$ for all r, so the event $\{\phi_n^\ell(r) \le \zeta_3 r^2\sqrt{n}\}$ is sure for $r \ge (2LB/\zeta_3)^{1/2}$.

Lemma D.4 (Multiplier monotonicity and positive-endpoint attainment)**.** *Fix $L > 0$, $\zeta_3 > 0$ and $\delta \in (0,1)$, and set $A_r = \{\phi_n^\ell(r) \le \zeta_3 r^2\sqrt{n}\}$ for $r > 0$. Then:*

(i) $r \mapsto \phi_n^\ell(r)/r$ *is non-increasing on* $(0,\infty)$*, pathwise; consequently* $r \mapsto \phi_n^\ell(r)$ *is continuous on* $(0,\infty)$ *and* $A_r \subseteq A_{r'}$ *whenever* $0 < r \le r'$*;*

(ii) the set $\{r > 0 : \mathbb{P}[A_r] \ge 1-\delta\}$ *is an interval unbounded above; if its infimum* a *is positive, the bound holds at* a*. If* $a = 0$*, it holds at every positive radius;*

(iii) consequently $\mathbb{P}\big[\phi_n^\ell(\alpha_n^\ell) \le \zeta_3(\alpha_n^\ell)^2\sqrt{n}\big] \ge 1-\delta$*, with* $\alpha_n^\ell = \alpha_n^\ell(\zeta_3,\delta)$ *as defined above.*

Proof. (i) Fix $0 < \theta \le 1$ and a curve $\gamma \in \widetilde{\Gamma}_{\mathcal{V}}$ with $\mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star) \le r$. Define its geodesic contraction towards $\gamma^\star$ by $\gamma_\theta(t) = \mathsf{Exp}_{\gamma^\star(t)}\big(\theta \mathsf{Log}_{\gamma^\star(t)}(\gamma(t))\big)$; the radius-$B$ Hadamard ball is geodesically convex, so the pointwise contracted predictions remain in that ball. Lemma C.2 keeps the contracted curve in the hull. By construction $\mathsf{Log}_{\gamma^\star(t)}(\gamma_\theta(t)) = \theta \mathsf{Log}_{\gamma^\star(t)}(\gamma(t))$ and $\mathsf{d}_{\mathsf{g}}(\gamma^\star(t),\gamma_\theta(t)) = \theta\, \mathsf{d}_{\mathsf{g}}(\gamma^\star(t),\gamma(t))$, so $\mathsf{D}_{\mathsf{g}}(\gamma_\theta,\gamma^\star) = \theta \mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star) \le \theta r$. Since the multiplier process (15) is linear in $\mathsf{Log}_{\gamma^\star(T_i)}(\gamma(T_i))$, its value at γ_θ is exactly θ times its value at γ. Taking suprema, $\phi_n^\ell(\theta r) \ge \theta\, \phi_n^\ell(r)$, i.e. $\phi_n^\ell(\theta r)/(\theta r) \ge \phi_n^\ell(r)/r$, which is the claimed monotonicity. Combining it with the (obvious) monotonicity of ϕ_n^ℓ itself gives, for $0 < r \le r'$,

$$\phi_n^\ell(r) \le \phi_n^\ell(r') \le \frac{r'}{r}\, \phi_n^\ell(r),$$

and letting $r' \downarrow r$ or $r \uparrow r'$ yields continuity. Finally, A_r may be rewritten as $\{\phi_n^\ell(r)/r \le \zeta_3 r\sqrt{n}\}$, whose left side is non-increasing in r and whose right side is strictly increasing; hence $A_r \subseteq A_{r'}$ for $r \le r'$.

(ii) By (i) the events A_r increase with r, so $r \mapsto \mathbb{P}[A_r]$ is non-decreasing and $\{r : \mathbb{P}[A_r] \ge 1-\delta\}$ is an interval unbounded above. Let a denote its infimum and suppose $a > 0$ (if $a = 0$, upward closure gives the event bound at every positive radius, without defining an event at zero). Take $r_k \downarrow a$ with $\mathbb{P}[A_{r_k}] \ge 1-\delta$. The A_{r_k} decrease, and by the continuity established in (i) one may pass to the limit in the inequality defining A_{r_k} to get $\bigcap_k A_{r_k} = A_a$. Continuity from above of $\mathbb{P}$ then gives $\mathbb{P}[A_a] = \lim_k \mathbb{P}[A_{r_k}] \ge 1-\delta$.

(iii) α_n^ℓ is the maximum of a and the floor $r_0(\zeta_3) > 0$. If $\alpha_n^\ell = a$ the claim is (ii); if $\alpha_n^\ell = r_0(\zeta_3) > a$ it follows from (ii) together with the upward closedness in (i). □

The floor $r_0(\zeta_3) = L/(\sqrt{2}\,\zeta_3\sqrt{n})$ ensures the variance threshold in the probability-symmetrization argument of Lemma H.10.

E Oracle bounds and interpretation

This appendix states the distance comparison and empirical lower bound, then proves the oracle inequality and its group-penalty consequence. The full geometric and probabilistic arguments are in Appendices G–I.

E.1 Distance comparison and empirical lower bound

The following comparison connects the $L_2(\mathbb{P}_T)$ distance between tangent fields to integrated geodesic distance. The lower bound is intrinsic to a Hadamard manifold; the upper bound uses the curvature lower bound on the prediction domain.

Proposition E.1 (Norm equivalence between Tangent and Geodesic Distances)**.** *Let* $v, v^\star \in \mathcal{V}_p^B$ *and set* $\gamma = \mathsf{Exp}_p \circ v$*,* $\gamma^\star = \mathsf{Exp}_p \circ v^\star \in \Gamma_p^B$*. Under Assumption 2, put* $\kappa_B = \sinh(\Lambda_B B)/(\Lambda_B B)$*, with continuous value* 1 *at* $\Lambda_B = 0$*. Then*

$$\|v - v^\star\|_{L_2(\mathbb{P}_T)}^2 \le \mathsf{D}_{\mathsf{g}}^2(\gamma,\gamma^\star) \le \kappa_B^2\, \|v - v^\star\|_{L_2(\mathbb{P}_T)}^2. \tag{16}$$

Identifying each field with its exponential image, for every $r > 0$*,*

$$\{\gamma \in \Gamma_p^B : \mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star) \le r\} \subset \{v \in \mathcal{V}_p^B : \|v - v^\star\|_{L_2(\mathbb{P}_T)} \le r\} \subset \{\gamma \in \Gamma_p^B : \mathsf{D}_{\mathsf{g}}(\gamma,\gamma^\star) \le r\,\kappa_B\}.$$

Proof. By Proposition G.3, $\|v(T) - v^\star(T)\|_{\mathsf{g}_p} \le \mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p v(T), \mathsf{Exp}_p v^\star(T))$ pointwise, so squaring and taking expectation over $T \sim \mathbb{P}_T$ gives $\|v - v^\star\|^2_{L_2(\mathbb{P}_T)} \le \mathbb{E}[\mathsf{d}^2_{\mathsf{g}}(\gamma(T), \gamma^\star(T))] = \mathsf{D}^2_{\mathsf{g}}(\gamma, \gamma^\star)$. Conversely, by Proposition G.4, $\mathsf{d}_{\mathsf{g}}(\gamma(T), \gamma^\star(T)) \le \kappa_B \|v(T) - v^\star(T)\|_{\mathsf{g}_p}$ pointwise, so $\mathsf{D}^2_{\mathsf{g}}(\gamma, \gamma^\star) \le \kappa_B^2 \|v - v^\star\|^2_{L_2(\mathbb{P}_T)}$. The displayed set inclusions follow immediately from (16). □

E.2 Empirical lower bound

Standing choices of constants. Let $\tau > 0$ and $Q_{\mathcal{H}}(\tau) > 0$ be the small-ball parameters of Assumption 7, μ and L the constants of Assumptions 4–5, κ_B the curvature constant of Proposition G.4, and $c_S > 0$ the universal Sudakov constant of Theorem H.6. One can choose free parameters $c_0 > 0$, $c_1 > 1$ (Lemma H.7; Remark H.2), a confidence level $\delta \in (0, 1)$, a multiplier-process parameter $\zeta_3 > 0$, and a localization parameter

$$\eta \in \big(0, \eta^\star(\tau)\big), \qquad \eta^\star(\tau) = \min\left\{1, \frac{\tau}{4\sqrt{2}\, c_0}, c_S^2\right\}, \tag{17}$$

These determine

$$\zeta_1 = \frac{c_0 \eta^{3/2} Q_H^2(\tau)}{256 c_1}, \quad \zeta_2 = \frac{c_0 \eta}{16\sqrt{2}} Q_H(\tau), \quad c_{\mathcal{A}}(\eta, \tau) = \min\left\{\frac{15}{512}, \frac{c_0^2 \eta^2}{\tau^2}\right\} > 0, \tag{18}$$

a reference radius $\bar{r}$, fixed in advance, with

$$\bar{r} > \omega_Q(\zeta_1, \zeta_2), \tag{19}$$

the localization radius

$$r^\star = \max\left\{\bar{r},\ \alpha_n^\ell(\zeta_3, \delta/4)\right\}, \tag{20}$$

and the event

$$\mathcal{A} \ = \mathcal{A}_{\mathrm{sb}}(\bar{r}) \cap \mathcal{A}_{\mathrm{mult}}, \tag{21}$$

where $\mathcal{A}_{\mathrm{sb}}(r)$ is the small-ball selection event of Lemma H.7 at radius r (probability at least $1 - 5\exp(-c_{\mathcal{A}}(\eta, \tau) Q_H^2(\tau) n)$) and $\mathcal{A}_{\mathrm{mult}}$ the multiplier-process event of Lemma H.10 (probability at least $1 - \delta$). Table 3 summarizes the status of each symbol.

Table 3: Reading guide to the constants of Theorems E.2 and E.3. *Free*: the analyst chooses it, subject to the stated constraint. *Determined*: a formula in quantities already fixed. *Model*: imposed by the loss, the manifold or the design law. *Universal*: an absolute constant, of which only positivity is used.

Symbol	Status; where fixed	Meaning
μ	model; Assumption 4	strong-convexity modulus of ℓ
L	model; Assumption 5	Lipschitz constant of ℓ
B, Λ_B	model; Assumption 2	radius and curvature bound of $\mathcal{D}_B$
$\tau, Q_H(\tau)$	model; Assumption 7	small-ball parameters of H
$\omega_G, \omega_R, \omega_Q$	model; Definitions 1 and 2	localized complexity fixed points
α_n^ℓ	model; (15)	multiplier fixed point, with floor r_0
c_S	universal; Theorem H.6	Sudakov constant
$c_0 > 0,\ c_1 > 1$	free; Remark H.2	complexity/net tuning
$\eta \in (0, \eta^\star(\tau))$	free; (17)	localization tuning
$\zeta_3 > 0$	free; Theorem E.3	multiplier-process tuning
$\delta \in (0, 1)$	free	confidence level
κ_B	determined; Proposition G.4	$\sinh(\Lambda_B B)/(\Lambda_B B) \ge 1$
ζ_1, ζ_2	determined; (18)	complexity multipliers
$\bar{r}$	free; (19)	any radius $> \omega_Q(\zeta_1, \zeta_2)$, fixed before the sample
$r^\star$	determined; (20)	$\max\{\bar{r}, \alpha_n^\ell(\zeta_3, \delta/4)\}$
$c_{\mathcal{A}}(\eta, \tau)$	determined; (43)	$\min\{15/512,\ c_0^2\eta^2/(\tau^2)\}$
$\varkappa_0, \varkappa$	determined; Theorem E.3	curvature/margin constants ($\neq$ the sectional curvature κ)

The analyst chooses $c_0, c_1, \eta, \zeta_3, \delta$ and the reference radius $\bar{r} > \omega_Q$. The constant c_S is universal; its role in the admissible range of η is explained in Remark H.1.

Theorem E.2 (Lower Bound on the Empirical Excess Risk)**.** *Grant Assumptions 1–7 , and make the standing choices* (17)–(21)*. Then the event $\mathcal{A}$ has probability at least $1 - \delta - 5\exp\big(-c_{\mathcal{A}}(\eta,\tau)\,Q_{\mathcal{H}}^2(\tau)\,n\big)$, and on $\mathcal{A}$, simultaneously for every $v \in \mathcal{V}$ with*

$$\|v - v^\star\|_{L_2(\mathbb{P}_T)} \;>\; r^\star, \tag{22}$$

one has

$$\mathsf{R}_n\mathcal{L}_v \;\geq\; \left[\Big(\frac{1}{4} - \sqrt{2}\,c_0\frac{\eta}{\tau}\Big)\,Q_H(\tau)\,\frac{\tau^2\mu}{8} \;-\; 4\zeta_3\kappa_B^2\right]\,\|v - v^\star\|_{L_2(\mathbb{P}_T)}^2. \tag{23}$$

In particular, if $\zeta_3 \leq \frac{1}{64\kappa_B^2}\big(\frac{1}{4} - \sqrt{2}\,c_0\frac{\eta}{\tau}\big)Q_H(\tau)\tau^2\mu$, the bracket is at least half its first term and $\mathsf{R}_n\mathcal{L}_v > 0$ for every such v.

The proof is given in Appendix I; it combines the concentration and empirical-process tools of Appendix H with the comparison estimates of Appendix G, in the order set out in Remark E.1.

Remark E.1 (Proof roadmap). Geodesic strong convexity decomposes the excess loss into a first-order term and a quadratic geodesic distance. The tangent small-ball argument, made uniform through the Gaussian and Rademacher fixed points, lower-bounds the empirical quadratic term (Lemmas H.7–H.8). Ambient optimality makes the population first-order term nonnegative, while the multiplier hull controls its empirical deviation (Lemmas I.1 and H.10). The radius $r^\star = \max\{\bar{r}, \alpha_n^\ell(\zeta_3, \delta/4)\}$ ensures that both empirical estimates apply. Appendix I assembles these steps without requiring geodesic closure of the estimator's model.

Remark E.2 (The floor does not determine the whole rate). The floor $r_0 = L/(\sqrt{2}\zeta_3\sqrt{n})$ is one lower bound on the multiplier radius. Neither that floor nor finiteness of ω_Q proves $r^\star = O(n^{-1/2})$. The widths on $\mathcal{H}$ and the multiplier radius on $\widetilde{\Gamma}_{\mathcal{V}}$ must both be controlled. Corollary F.9 does so under its explicit sample-size and closure assumptions.

E.3 Oracle inequality: proof of Theorem 3

We define the excess risk of $v \in \mathcal{V}$ by

$$\mathsf{R}(v) - \mathsf{R}(v^\star) = \mathbb{E}_{\mathbb{P}_{T,Y}}[(\ell_v - \ell_{v^\star})(T,Y)] = \int_{\mathcal{Z}} (\ell_v - \ell_{v^\star})(t,y)\mathrm{d}\mathbb{P}_{T,Y}(t,y) = \mathsf{R}(\mathcal{L}_v).$$

Since $\hat{v}$ minimizes $v \mapsto \mathsf{R}_n(\ell_v) + \lambda\Omega(v)$ over $\mathcal{V}$, Assumption 6 permits using $v = v^\star$ as comparator, giving $\mathsf{R}_n(\ell_{\hat{v}}) + \lambda\Omega(\hat{v}) \leq \mathsf{R}_n(\ell_{v^\star}) + \lambda\Omega(v^\star)$, that is, since $\mathcal{L}_{v^\star} \equiv 0$,

$$\mathsf{R}_n\mathcal{L}_{\hat{v}} \;\leq\; \lambda\big(\Omega(v^\star) - \Omega(\hat{v})\big). \tag{24}$$

We now assemble Theorem E.2 and (24) into the final result. Throughout, $\varkappa$ denotes the (assumed positive) curvature/margin constant of Theorem E.2, not to be confused with the sectional curvature κ of Assumption 2:

$$\varkappa = \left(\tfrac{1}{4} - \sqrt{2}\,c_0\tfrac{\eta}{\tau}\right)\,Q_H(\tau)\,\frac{\tau^2\mu}{8} - 4\zeta_3\kappa_B^2 > 0.$$

Theorem E.3. *Grant Assumptions 1–7, and make the standing choices* (17)–(21)*, so that $\varkappa_0 = (\frac{1}{4} - \sqrt{2}\,c_0\frac{\eta}{\tau})Q_H(\tau)\frac{\tau^2\mu}{8} > 0$; choose $\zeta_3 \in (0, \varkappa_0/(8\kappa_B^2)]$, so that $\varkappa = \varkappa_0 - 4\zeta_3\kappa_B^2 \geq \varkappa_0/2 > 0$. Then, for every $\lambda \geq 0$, on the event $\mathcal{A}$ of Theorem E.2 (probability at least $1 - \delta - 5\exp\big(-c_{\mathcal{A}}(\eta,\tau)Q_H^2(\tau)n\big)$), any RERM estimator $\hat{v}$ of* (2) *satisfies*

$$\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)}^2 \leq (r^\star)^2 \;+\; \frac{\lambda}{\varkappa}\,\Omega(v^\star). \tag{25}$$

Moreover, on the same event, if in addition $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} > r^\star$, then also

$$\varkappa\,\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)}^2 \;+\; \lambda\,\Omega(\hat{v}) \leq \lambda\,\Omega(v^\star). \tag{26}$$

Consequently, on the same event,

$$\mathsf{D}_{\mathsf{g}}^2(\hat{\gamma}, \gamma^\star) \leq \kappa_B^2\Big[(r^\star)^2 + \tfrac{\lambda}{\varkappa}\Omega(v^\star)\Big], \tag{27}$$

$$0 \leq \mathsf{R}(\hat{v}) - \mathsf{R}(v^\star) \leq L\,\kappa_B\,\Big[(r^\star)^2 + \tfrac{\lambda}{\varkappa}\Omega(v^\star)\Big]^{1/2}. \tag{28}$$

Proof. We work throughout on the event $\mathcal{A}$ of Theorem E.2, which has probability at least $1-\delta-5\exp(-c_{\mathcal{A}}(\eta,\tau)Q_H^2(\tau)n)$; the choice of ζ_3 guarantees that the bracket in (23) equals $\varkappa \geq \varkappa_0/2 > 0$. Since Theorem E.2 holds simultaneously for every v, it may be applied to the data-dependent $\hat{v}$.

Case 1: $\|\hat{v}-v^\star\|_{L_2(\mathbb{P}_T)} \leq r^\star$. Then (25) holds trivially since $\Omega(v^\star) \geq 0$ and $\lambda \geq 0$, $\varkappa > 0$.

Case 2: $\|\hat{v}-v^\star\|_{L_2(\mathbb{P}_T)} > r^\star$. By Theorem E.2 applied at $v=\hat{v}$,

$$\mathsf{R}_n\mathcal{L}_{\hat{v}} \;\geq\; \varkappa\,\|\hat{v}-v^\star\|^2_{L_2(\mathbb{P}_T)}.$$

By the basic optimality inequality (24) (valid deterministically, for every $\lambda \geq 0$), $\mathsf{R}_n\mathcal{L}_{\hat{v}} \leq \lambda(\Omega(v^\star)-\Omega(\hat{v}))$. Combining,

$$\varkappa\,\|\hat{v}-v^\star\|^2_{L_2(\mathbb{P}_T)} \leq \lambda\big(\Omega(v^\star)-\Omega(\hat{v})\big) \leq \lambda\,\Omega(v^\star), \tag{29}$$

using $\Omega(\hat{v}) \geq 0$ for the last step. This gives (25) (dividing by $\varkappa$), and (29) rearranged (without discarding $-\lambda\Omega(\hat{v})$) is exactly (26).

In either case, (25) holds, which is the claimed $L_2(\mathbb{P}_T)$-error oracle inequality. The geodesic-distance bound (27) follows from Lemma E.1 ($\mathsf{D}_{\mathsf{g}}^2(\hat{\gamma},\gamma^\star) \leq \kappa_B^2\|\hat{v}-v^\star\|^2_{L_2(\mathbb{P}_T)}$). Finally, for the excess-risk bound (28): the lower bound $\mathsf{R}(\hat{v}) \geq \mathsf{R}(v^\star)$ holds by definition of $v^\star$ as the global minimizer; for the upper bound, Assumption 5 together with Jensen's inequality gives

$$\mathsf{R}(\hat{v})-\mathsf{R}(v^\star) = \mathbb{E}[\ell(\hat{\gamma}(T),Y)-\ell(\gamma^\star(T),Y)] \leq L\,\mathbb{E}[\mathsf{d}_{\mathsf{g}}(\hat{\gamma}(T),\gamma^\star(T))] \leq L\,\mathsf{D}_{\mathsf{g}}(\hat{\gamma},\gamma^\star),$$

and combining with (27) gives (28). □

Interpretation. The bound equals $(r^\star)^2$ at $\lambda=0$ and is nondecreasing in λ; it does not prove an advantage over ERM. Such a comparison would need additional control, for example localization within penalty-constrained classes as in Lecué & Mendelson (2018); Alquier et al. (2019). Moreover, finiteness of the fixed points does not supply a rate. The dictionary bounds in Appendix F.5 quantify them under additional closure.

Remark E.3 (Constants and rate dependence). Increasing η raises the width thresholds ζ_1,ζ_2 and can reduce ω_Q, but lowers $\varkappa_0$. The latter also limits the permissible ζ_3, coupling the quadratic margin to the multiplier radius. For fixed other choices, α_n^ℓ is nonincreasing in ζ_3. Its ceiling $\zeta_3 = \varkappa_0/(8\kappa_B^2)$ therefore minimizes this radius, although a smaller value can improve the penalty prefactor $1/\varkappa$.

At the ceiling, the variance floor becomes

$$r_0 = \frac{4\sqrt{2}L\kappa_B^2}{\varkappa_0\sqrt{n}}, \qquad r_0^2 = \frac{32L^2\kappa_B^4}{\varkappa_0^2 n}.$$

Its squared-tangent, squared-geodesic and excess-risk contributions have factors κ_B^4, κ_B^6 and κ_B^3, respectively, with the latter on an $n^{-1/2}$ scale. These describe only the floor. The full dictionary radius in Corollary F.9 instead gives κ_B^6 and κ_B^8 in squared tangent and geodesic errors before the penalty terms. A slower component of $r^\star$ cannot be replaced by the floor's rate.

We next specialize the main theorem to one concrete regularizer.

Corollary E.4 (Group-Sparse Oracle Inequality)**.** *Fix an orthonormal basis $(e_1,\dots,e_d)$ of $\mathcal{T}_p\mathcal{M}$ and let*

$$\Omega_{\mathrm{group}}(v) = \sum_{k=1}^d \|v_k\|_{L_2([0,1])}, \qquad v(t) = \sum_{k=1}^d v_k(t)e_k,\; v_k \in L_2([0,1]),$$

on the ambient Hilbert space $\bigoplus_{k=1}^d L_2([0,1]) \cong L_2([0,1];\mathcal{T}_p\mathcal{M})$; this is a norm, so it is convex, and its subdifferential at v is

$$(\partial\Omega_{\mathrm{group}})_v = \Big\{g=(g_k)_{k=1}^d : g_k = \tfrac{v_k}{\|v_k\|_{L_2}} \text{ if } v_k \neq 0,\;\; \|g_k\|_{L_2} \leq 1 \text{ if } v_k = 0\Big\},$$

acting on $h=(h_k)_{k=1}^d$ by $g(h)=\sum_k\langle g_k,h_k\rangle_{L_2}$; in particular $\|g\|_{\Omega^*_{\mathrm{group}}} \leq 1$ for every $g \in (\partial\Omega_{\mathrm{group}})_v$ (the dual norm of Ω_{group} is the ℓ_∞-of-L_2 group norm). Assume that $\mathbb{P}_T$ has a density $f_T \geq c_- > 0$ on $[0,1]$, and set $c_{\mathbb{P}_T} = c_-^{-1/2}$, so that $\|u\|_{L_2([0,1])} \leq c_{\mathbb{P}_T}\|u\|_{L_2(\mathbb{P}_T)}$. If

$v^\star$ is supported on $S \subset \{1,\dots,d\}$, $|S| = s$ (i.e. $v_k^\star \equiv 0$ for $k \notin S$), then by Cauchy–Schwarz across the s active groups and the orthonormality of (e_k), $\Omega_{\mathrm{group}}(v^\star) = \sum_{k\in S} \|v_k^\star\|_{L_2([0,1])} \le \sqrt{s}\,\big(\sum_{k\in S} \|v_k^\star\|^2_{L_2([0,1])}\big)^{1/2} \le c_{\mathbb{P}_T}\sqrt{s}\,\|v^\star\|_{L_2(\mathbb{P}_T)}$. Under the hypotheses of Theorem E.3 with $\Omega = \Omega_{\mathrm{group}}$, for every $\lambda > 0$, on the same event,

$$\|\hat v - v^\star\|^2_{L_2(\mathbb{P}_T)} \le (r^\star)^2 + \frac{\lambda\, c_{\mathbb{P}_T}\sqrt{s}\,\|v^\star\|_{L_2(\mathbb{P}_T)}}{\varkappa}.$$

Proof. Immediate from (25) and the displayed bound on $\Omega_{\mathrm{group}}(v^\star)$. □

F Decomposable Regularization and Dictionary Bounds

Decomposability refines the regularization contribution in the oracle bound without changing its localization radius. We give the general inequality first, then instantiate it for coefficient, group and cellwise sparsity. The final subsection derives an explicit distance rate under additional geodesic closure and explains why that condition need not hold for Fourier models.

F.1 A General Oracle Inequality for Decomposable Regularizers

Write $\mathbb{H}$ for a closed linear subspace of $L_2(\mathbb{P}_T)$ containing $\mathcal{V}$ and hence $\mathcal{H} = \mathcal{V} - \mathcal{V}$ and on which the regularization Ω is defined and finite, with inner product $\langle u, w\rangle_{\mathbb{H}} = \mathbb{E}_{T\sim\mathbb{P}_T}[\langle u(T), w(T)\rangle_{\mathbf{g}_p}]$ inducing $\|\cdot\|_{L_2(\mathbb{P}_T)}$. For a closed linear subspace $M \subset \mathbb{H}$, u_M denotes the orthogonal projection of $u \in \mathbb{H}$ onto M, and $M^\perp$ its orthogonal complement. Nothing below uses any property of $\mathbb{H}$ beyond its being a Hilbert space containing $\hat v - v^\star$: for Ω_{group} one may take $\mathbb{H} = L_2(\mathbb{P}_T)$ under the density lower bound of Section F.3, for Ω_{lasso} the span of the dictionary $(\phi_k)_{k\le K}$, and for Ω_{time} the piecewise-constant class on the grid.

Definition F.1 (Decomposable Regularizer)**.** Following Negahban et al. (2012), let $\Omega : \mathbb{H} \to \mathbb{R}_+$ be a seminorm, i.e. nonnegative, absolutely homogeneous and subadditive (definiteness is not required). A pair of closed linear subspaces $M \subseteq \bar M \subseteq \mathbb{H}$ is a *decomposition pair* for Ω if

$$\Omega(u + w) = \Omega(u) + \Omega(w) \qquad \text{for every } u \in M,\ w \in \bar M^\perp.$$

The *subspace compatibility constant* of $\bar M$ is

$$\Psi(\bar M) = \sup_{u\in\bar M\setminus\{0\}} \frac{\Omega(u)}{\|u\|_{L_2(\mathbb{P}_T)}}.$$

We set $\Psi(\{0\}) = 0$ for the zero subspace.

Lemma F.1 (Decomposability Inequality)**.** *Let $(M, \bar M)$ be a decomposition pair for Ω and let $v^\star, \hat v \in \mathbb{H}$, $h = \hat v - v^\star$. Then*

$$\Omega(v^\star) - \Omega(\hat v) \le \Omega(h_{\bar M}) - \Omega(h_{\bar M^\perp}) + 2\,\Omega(v^\star_{M^\perp}).$$

In particular, if $v^\star \in M$ (exact membership in the model subspace), $\Omega(v^\star) - \Omega(\hat v) \le \Omega(h_{\bar M}) - \Omega(h_{\bar M^\perp}) \le \Omega(h_{\bar M})$.

Proof. Write $\hat v = (v^\star_M + h_{\bar M^\perp}) + (v^\star_{M^\perp} + h_{\bar M})$. Since $v^\star_M \in M$ and $h_{\bar M^\perp} \in \bar M^\perp$, decomposability and the triangle inequality give

$$\begin{aligned}\Omega(\hat v) &\ge \Omega(v^\star_M + h_{\bar M^\perp}) - \Omega(v^\star_{M^\perp} + h_{\bar M})\\ &\ge \Omega(v^\star_M) + \Omega(h_{\bar M^\perp}) - \Omega(v^\star_{M^\perp}) - \Omega(h_{\bar M}).\end{aligned}$$

Also $\Omega(v^\star) \le \Omega(v^\star_M) + \Omega(v^\star_{M^\perp})$. Subtracting the two bounds proves the result. If $v^\star \in M$, its $M^\perp$ component vanishes. □

Theorem F.2. *Grant the hypotheses of Theorem E.3, fix $\lambda > 0$, and suppose Ω admits a decomposition pair $(M, \bar M)$ (Definition F.1) with finite subspace compatibility constant $\Psi(\bar M)$. Then, on the event $\mathcal{A}$ of Theorem E.2,*

$$\|\hat v - v^\star\|_{L_2(\mathbb{P}_T)} \le \max\left\{ r^\star,\ \frac{\lambda\,\Psi(\bar M)}{\varkappa} + \sqrt{\frac{2\lambda\,\Omega(v^\star_{M^\perp})}{\varkappa}} \right\}. \tag{30}$$

In particular, if $v^\star \in M$ exactly, the second term simplifies to $\lambda\Psi(\bar{M})/\varkappa$. Consequently, writing $r_\Omega = \max\{r^\star,\ \lambda\Psi(\bar{M})/\varkappa + \sqrt{2\lambda\Omega(v^\star_{M^\perp})/\varkappa}\}$, the geodesic-distance and excess-risk bounds

$$\mathsf{D}_{\mathsf{g}}^2(\hat{\gamma}, \gamma^\star) \le \kappa_B^2\, r_\Omega^2, \qquad 0 \le \mathsf{R}(\hat{v}) - \mathsf{R}(v^\star) \le L\,\kappa_B\, r_\Omega$$

hold on the same event, exactly as in (27)–(28).

Proof. As in the proof of Theorem E.3, if $\|h\|_{L_2(\mathbb{P}_T)} \le r^\star$ (with $h = \hat{v} - v^\star$) the bound holds trivially. Otherwise, by (29),

$$\varkappa\|h\|^2_{L_2(\mathbb{P}_T)} \le \lambda\big(\Omega(v^\star) - \Omega(\hat{v})\big).$$

By Lemma F.1, $\Omega(v^\star) - \Omega(\hat{v}) \le \Omega(h_{\bar{M}}) - \Omega(h_{\bar{M}^\perp}) + 2\Omega(v^\star_{\bar{M}^\perp}) \le \Omega(h_{\bar{M}}) + 2\Omega(v^\star_{\bar{M}^\perp})$ (discarding $-\Omega(h_{\bar{M}^\perp}) \le 0$). By Definition F.1, $\Omega(h_{\bar{M}}) \le \Psi(\bar{M})\|h_{\bar{M}}\|_{L_2(\mathbb{P}_T)} \le \Psi(\bar{M})\|h\|_{L_2(\mathbb{P}_T)}$, the last step because orthogonal projection is a contraction in $\mathbb{H}$. Writing $x = \|h\|_{L_2(\mathbb{P}_T)}$, $\Psi = \Psi(\bar{M})$, $\Omega_0 = \Omega(v^\star_{\bar{M}^\perp})$, we have shown

$$\varkappa x^2 \le \lambda\Psi x + 2\lambda\Omega_0,$$

a quadratic inequality in $x \ge 0$ whose largest root is $x \le \dfrac{\lambda\Psi + \sqrt{(\lambda\Psi)^2 + 8\varkappa\lambda\Omega_0}}{2\varkappa} \le \dfrac{\lambda\Psi}{\varkappa} + \sqrt{\dfrac{2\lambda\Omega_0}{\varkappa}}$, the last inequality by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ applied to $a = (\lambda\Psi)^2$, $b = 8\varkappa\lambda\Omega_0$, together with $\sqrt{(\lambda\Psi)^2} = \lambda\Psi$ and $\sqrt{8\varkappa\lambda\Omega_0}/(2\varkappa) = \sqrt{2\lambda\Omega_0/\varkappa}$. This gives (30). The geodesic-distance and excess-risk bounds follow exactly as in the proof of Theorem E.3, via Lemma E.1 and Assumption 5. □

We now instantiate Theorem F.2 for three of the regularizations of Section C.2.

F.2 Sparse basis expansions and the coefficient penalty

Recall the coefficient regularization of Section C.2: a fixed dictionary $(\phi_k)_{k=1}^K \subset \mathcal{T}_p\mathcal{M}$-valued functions on $[0,1]$, $v(t) = \sum_{k=1}^K \beta_k\phi_k(t)$, $\Omega_{\text{lasso}}(v) = \|\beta\|_1$. To make its complexity explicitly computable, we introduce one new, clearly flagged assumption.

Assumption F.1 (Orthonormal Bounded Dictionary)**.** The fixed Borel functions $(\phi_k)_{k=1}^K \subset L_2(\mathbb{P}_T)$ satisfy $\mathbb{E}_{T\sim\mathbb{P}_T}\big[\langle\phi_j(T), \phi_k(T)\rangle_{\mathsf{g}_p}\big] = \delta_{jk}$, for $1 \le j, k \le K$, and

$$\sup_{k\le K}\ \sup_{t\in[0,1]} \|\phi_k(t)\|_{\mathsf{g}_p} \le M_\phi < \infty,$$

for a finite constant $M_\phi \ge 1$, whose dependence on K must be retained for a sequence of dictionaries. Let

$$\mathcal{V}_p^{B,K} = \Big\{ \sum_{k=1}^K \beta_k\phi_k : \beta \in \mathbb{R}^K,\ \|\sum_k \beta_k\phi_k(t)\|_{\mathsf{g}_p} \le B\ \forall t \Big\},$$

Set $\mathcal{V} = \mathcal{V}_p^{B,K}$ and $\Gamma_{\mathcal{V}} = \mathsf{Exp}_p \circ \mathcal{V}$. Under Assumption F.1, write $v_\beta = \sum_{k=1}^K \beta_k\phi_k$ and define the continuously feasible coefficient set by

$$\mathcal{B}_K = \Big\{ \beta \in \mathbb{R}^K : \Big\| \sum_{k=1}^K \beta_k\phi_k(t) \Big\|_{\mathsf{g}_p} \le B \text{ for every } t \in [0,1] \Big\}.$$

One has the exact isometry

$$\|v\|^2_{L_2(\mathbb{P}_T)} = \mathbb{E}\Big[\Big\|\sum_k \beta_k\phi_k(T)\Big\|^2_{\mathsf{g}_p}\Big] = \sum_{j,k}\beta_j\beta_k\,\mathbb{E}[\langle\phi_j(T),\phi_k(T)\rangle_{\mathsf{g}_p}] = \sum_{k=1}^K \beta_k^2 = \|\beta\|_2^2, \quad (31)$$

so $(\mathcal{V}_p^{B,K}, \|\cdot\|_{L_2(\mathbb{P}_T)})$ is isometric to $(\mathcal{B}_K, \|\cdot\|_2)$ via $v \leftrightarrow \beta$, and $\Omega_{\text{lasso}}(v) = \|\beta\|_1$.

Lemma F.3 (The small-ball condition holds on a bounded orthonormal dictionary class)**.** *Grant Assumption F.1 and let $\mathcal{H} = \mathcal{V}_p^{B,K} - \mathcal{V}_p^{B,K}$. Write $h_\beta = \sum_{k=1}^K \beta_k \phi_k$ and let*

$$C_4 = \sup_{\beta \in \mathbb{R}^K \setminus \{0\}} \frac{\|h_\beta\|_{L_4(\mathbb{P}_T)}}{\|h_\beta\|_{L_2(\mathbb{P}_T)}}, \qquad \|h\|_{L_4(\mathbb{P}_T)}^4 = \mathbb{E}\big[\|h(T)\|_{\mathbb{g}_p}^4\big],$$

be the L_4–L_2 equivalence constant of the dictionary span. Then $C_4^4 \le K M_\phi^2 < \infty$ and, for every $\tau \in (0,1)$,

$$Q_{\mathcal{H}}(\tau) \;\ge\; \frac{(1-\tau^2)^2}{C_4^4} \;\ge\; \frac{(1-\tau^2)^2}{K\, M_\phi^2} \;>\; 0. \tag{32}$$

In particular $\mathcal{V}_p^{B,K}$ satisfies Assumption 7.

Proof. Let $h \in \mathcal{H} \setminus \{0\}$. Since $\mathcal{V}_p^{B,K}$ is contained in the linear span of $(\phi_k)_{k \le K}$, so is $\mathcal{H}$, whence $h = h_\beta$ for some $\beta \neq 0$, and the isometry (31) gives $\|h\|_{L_2(\mathbb{P}_T)} = \|\beta\|_2 > 0$. Put $Z = \|h(T)\|_{\mathbb{g}_p}^2 \ge 0$, so that $\mathbb{E}Z = \|h\|_{L_2(\mathbb{P}_T)}^2$ and $\mathbb{E}Z^2 = \|h\|_{L_4(\mathbb{P}_T)}^4$, both finite because $\|h(t)\|_{\mathbb{g}_p} \le 2B$ for all t. The Paley–Zygmund inequality, applied with $\theta = \tau^2 \in (0,1)$, yields

$$\mathbb{P}\big[Z \ge \theta\, \mathbb{E}Z\big] \;\ge\; (1-\theta)^2\, \frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2} = (1-\tau^2)^2\, \frac{\|h\|_{L_2(\mathbb{P}_T)}^4}{\|h\|_{L_4(\mathbb{P}_T)}^4} \;\ge\; \frac{(1-\tau^2)^2}{C_4^4},$$

and the event $\{Z \ge \tau^2 \mathbb{E}Z\}$ is exactly $\{\|h(T)\|_{\mathbb{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}\}$. Taking the infimum over $h \in \mathcal{H} \setminus \{0\}$ gives the first inequality in (32).

For the second, Cauchy–Schwarz in $\mathbb{R}^K$ and the uniform bound of Assumption F.1 give, for every $t \in \mathcal{T}$,

$$\|h_\beta(t)\|_{\mathbb{g}_p} = \Big\| \sum_{k=1}^K \beta_k \phi_k(t) \Big\|_{\mathbb{g}_p} \le \|\beta\|_2 \Big(\sum_{k=1}^K \|\phi_k(t)\|_{\mathbb{g}_p}^2 \Big)^{1/2} \le \sqrt{K}\, M_\phi\, \|\beta\|_2,$$

so that $\mathbb{E}\|h_\beta(T)\|_{\mathbb{g}_p}^4 \;\le\; \big(\sup_t \|h_\beta(t)\|_{\mathbb{g}_p}^2\big)\, \mathbb{E}\|h_\beta(T)\|_{\mathbb{g}_p}^2 \;\le\; K M_\phi^2 \|\beta\|_2^4 \;=\; K M_\phi^2 \|h_\beta\|_{L_2(\mathbb{P}_T)}^4$, i.e. $C_4^4 \le K M_\phi^2$. □

Decomposability and Compatibility. Suppose $v^\star = \sum_{k \in S} \beta_k^\star \phi_k$ is exactly s-sparse, $S \subset \{1, \dots, K\}$, $|S| = s$. Let $M = \bar{M} = \{\sum_k \beta_k \phi_k : \beta_k = 0 \; \forall k \notin S\}$, an s-dimensional coordinate subspace under the isometry (31). For $u = \sum_{k \in S} u_k \phi_k \in M$ and $w = \sum_{k \notin S} w_k \phi_k \in \bar{M}^\perp$, $\|\beta(u) + \beta(w)\|_1 = \sum_{k \in S} |u_k| + \sum_{k \notin S} |w_k| = \Omega_{\text{lasso}}(u) + \Omega_{\text{lasso}}(w)$ (disjoint supports), so $(M, \bar{M})$ is a decomposition pair. By Cauchy–Schwarz, for $u \in \bar{M}$, $\Omega_{\text{lasso}}(u) = \sum_{k \in S} |u_k| \le \sqrt{s} \big(\sum_{k \in S} u_k^2 \big)^{1/2} = \sqrt{s}\, \|u\|_{L_2(\mathbb{P}_T)}$, with equality at $u_k \equiv \text{const}$ on S, so

$$\Psi(\bar{M}) = \sqrt{s}. \tag{33}$$

Corollary F.4 (Sparse-Dictionary Oracle Inequality)**.** *Grant Assumptions 1–7 and F.1, applied to the class $\mathcal{V}_p^{B,K}$ (so that $r^\star$ is the localization radius of that class), with $\Omega = \Omega_{\text{lasso}}$ and $v^\star$ exactly s-sparse in the dictionary $(\phi_k)_{k=1}^K$, and make the standing choices of Theorem E.3. Then, for every $\lambda > 0$, on the event $\mathcal{A}$ of Theorem E.2,*

$$\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} \le \max\left\{ r^\star,\; \frac{\lambda \sqrt{s}}{\varkappa} \right\}.$$

In particular, with $\lambda = \lambda_n = 2LM_\phi \sqrt{2\log(2K)/n}$, the regularization term is $\frac{2LM_\phi\sqrt{s}}{\varkappa} \sqrt{\frac{2\log(2K)}{n}}$.

Proof. Immediate from Theorem F.2 with $\Psi(\bar{M}) = \sqrt{s}$ (33) and $\Omega(v^\star_{M^\perp}) = 0$ (exact sparsity). □

With the choice $\lambda_n \asymp \sqrt{\log K / n}$ customary for the LASSO, the regularization term is of order $\sqrt{s \log K / n}$ up to the constants $\varkappa$, L and M_ϕ, which is the order of the rate obtained by Alquier et al. (2019, Theorem 3.1) for the logistic LASSO in the Euclidean case, and it depends on s rather than on $\|v^\star\|_{L_2(\mathbb{P}_T)}$, in contrast with the bound $\big(\lambda\sqrt{s}\|v^\star\|_{L_2(\mathbb{P}_T)}/\varkappa\big)^{1/2}$ obtained by plugging $\Omega_{\text{lasso}}(v^\star) \le \sqrt{s}\|v^\star\|_{L_2(\mathbb{P}_T)}$ into (25) and taking square roots.

F.3 GROUP-SPARSE REGULARIZATION TERM

For this group specialization take a closed linear representation space $\mathbb{H} \subseteq L_2(\mathbb{P}_T)$ containing $\mathcal{V}$, invariant under each tangent-coordinate projection. Thus the coordinate groups form orthogonal subspaces of $\mathbb{H}$. Recall $\Omega_{\mathrm{group}}(v) = \sum_{k=1}^{d} \|v_k\|_{L_2([0,1])}$, $v(t) = \sum_{k=1}^{d} v_k(t) e_k$, of Corollary E.4, and suppose $\mathbb{P}_T$ has a density f_T with respect to Lebesgue measure on $[0,1]$ satisfying $f_T \geq c_- > 0$ (making the informal "density bounded away from 0" proviso of Corollary E.4 precise). Since $(e_k)_{k=1}^{d}$ is orthonormal in $\mathcal{T}_p\mathcal{M}$, $\|v(t)\|_{\mathbf{g}_p}^2 = \sum_k v_k(t)^2$ pointwise, so *exactly* (no approximation),

$$\|v\|_{L_2(\mathbb{P}_T)}^2 = \mathbb{E}_T\Big[\sum_{k=1}^{d} v_k(T)^2\Big] = \sum_{k=1}^{d} \|v_k\|_{L_2(\mathbb{P}_T)}^2.$$

Let $v^\star$ be supported on $S \subset \{1,\dots,d\}$, $|S| = s$ (i.e. $v_k^\star \equiv 0$ for $k \notin S$), and let $M = \bar{M} = \{v : v_k \equiv 0 \; \forall k \notin S\}$: decomposability again holds exactly (disjoint groups). For $u \in \bar{M}$, using $f_T \geq c_-$,

$$\begin{aligned}\Omega_{\mathrm{group}}(u) &= \sum_{k\in S} \|u_k\|_{L_2([0,1])} \leq \frac{1}{\sqrt{c_-}} \sum_{k\in S} \|u_k\|_{L_2(\mathbb{P}_T)} \\ &\leq \frac{\sqrt{s}}{\sqrt{c_-}} \Big(\sum_{k\in S} \|u_k\|_{L_2(\mathbb{P}_T)}^2\Big)^{1/2} = \frac{\sqrt{s}}{\sqrt{c_-}} \|u\|_{L_2(\mathbb{P}_T)},\end{aligned}$$

so $\Psi(\bar{M}) \leq \sqrt{s}\, c_{\mathbb{P}_T}$ with $c_{\mathbb{P}_T} = 1/\sqrt{c_-}$, exactly the constant already used informally in Corollary E.4; only this upper bound is used below.

Corollary F.5 (Sharpened Group-Sparse Oracle Inequality)**.** *Grant the hypotheses of Theorem E.3, assume the representation space $\mathbb{H}$ is invariant under tangent-coordinate projections, and take $\Omega = \Omega_{\mathrm{group}}$, $v^\star$ exactly s-group-sparse, and $\mathbb{P}_T$ having density $f_T \geq c_- > 0$ on $[0,1]$. Then, for every $\lambda > 0$, on the event $\mathcal{A}$ of Theorem E.2,*

$$\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} \leq \max\Big\{r^\star,\ \frac{\lambda\sqrt{s}}{\varkappa\sqrt{c_-}}\Big\}.$$

Proof. Immediate from Theorem F.2 with $\Psi(\bar{M}) \leq \sqrt{s}/\sqrt{c_-}$ and $\Omega(v^\star_{M^\perp}) = 0$. □

Squaring, $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)}^2 \leq \max\{(r^\star)^2,\ \lambda^2 s/(\varkappa^2 c_-)\}$. Compared with Corollary E.4, whose regularization term $\lambda c_{\mathbb{P}_T}\sqrt{s}\|v^\star\|_{L_2(\mathbb{P}_T)}/\varkappa$ is linear in λ and depends on $\|v^\star\|_{L_2(\mathbb{P}_T)}$, the regularization term here is quadratic in λ and depends only on $(s, \varkappa, c_-)$; it is smaller whenever $\lambda\sqrt{s}/(\varkappa\sqrt{c_-}) < \|v^\star\|_{L_2(\mathbb{P}_T)}$, at the cost of the exact-sparsity hypothesis $v^\star_{M^\perp} = 0$.

F.4 TIME-LOCAL GRID REGULARIZATIONS

Recall regularization (ii): a grid $0 = t_1 < \cdots < t_J = 1$ and $\Omega_{\mathrm{time}}(v) = \sum_{j=1}^{J-1} \|v_j\|_{\mathbf{g}_p}$. For this specialization choose E to be the full cellwise-constant vector space of Proposition D.1; we take v to be the piecewise-constant interpolant $v(t) = v_j$ for $t \in I_j$ (as already anticipated informally in Section C.2), write $\mathcal{V}_p^{B,J} = \mathcal{V}$ for the class of such interpolants in this choice of model, and assume each grid cell carries a nonvanishing share of the design mass.

Assumption F.2 (Near-Uniform Grid Design)**.** The grid has $t_J = 1$, so that the cells are $I_j = [t_j, t_{j+1})$ for $j = 1,\dots,J-2$ together with the closed last cell $I_{J-1} = [t_{J-1}, 1]$; a piecewise-constant v takes the value v_j on I_j, and Ω_{time}, M and $\bar{M}$ are indexed by $j \in \{1,\dots,J-1\}$ accordingly. There exists a constant $c_{\mathrm{grid}} > 0$, not depending on J, such that $\mathbb{P}_T(I_j) \geq c_{\mathrm{grid}}/J$ for every one of these $J-1$ cells.

Under Assumption F.2, for a piecewise-constant v, $\|v\|_{L_2(\mathbb{P}_T)}^2 = \sum_{j=1}^{J-1} \|v_j\|_{\mathbf{g}_p}^2\, \mathbb{P}_T(I_j) \geq (c_{\mathrm{grid}}/J)\sum_{j=1}^{J-1} \|v_j\|_{\mathbf{g}_p}^2$. Let $v^\star$ be supported on $J_0 \subset \{1,\dots,J-1\}$, $|J_0| = s$ time-grid cells, $M = \bar{M} = \{v : v_j \equiv 0 \; \forall j \notin J_0\}$ (again exactly decomposable, disjoint index blocks). For $u \in \bar{M}$,

$$\Omega_{\mathrm{time}}(u) = \sum_{j\in J_0} \|u_j\|_{\mathbf{g}_p} \leq \sqrt{s}\Big(\sum_{j\in J_0} \|u_j\|_{\mathbf{g}_p}^2\Big)^{1/2} \leq \sqrt{s}\,\sqrt{J/c_{\mathrm{grid}}}\,\|u\|_{L_2(\mathbb{P}_T)},$$

so $\Psi(\bar{M}) \le \sqrt{sJ/c_{\mathrm{grid}}}$, which is all Theorem F.2 requires.

Corollary F.6 (Time-Local Sparse Oracle Inequality)**.** *Grant the hypotheses of Theorem E.3 applied to the piecewise-constant class $\mathcal{V}_p^{B,J}$ introduced above (so that $r^\star$ is the localization radius of that class), with $\Omega = \Omega_{\mathrm{time}}$, $v^\star$ exactly supported on s grid cells, and Assumption F.2. Then, for every $\lambda > 0$, on the event $\mathcal{A}$ of Theorem E.2,*

$$\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)} \le \max\Big\{r^\star,\ \frac{\lambda}{\varkappa}\sqrt{\frac{sJ}{c_{\mathrm{grid}}}}\Big\}.$$

Proof. Immediate from Theorem F.2 with $\Psi(\bar{M}) \le \sqrt{sJ/c_{\mathrm{grid}}}$ and $\Omega(v^\star_{M^\perp}) = 0$. □

The compatibility upper bound depends on resolution and the cell probabilities. For a specified active set J_0 with $p_j = \mathbb{P}_T(I_j) > 0$, the exact compatibility constant is $(\sum_{j\in J_0} p_j^{-1})^{1/2}$, by weighted Cauchy–Schwarz, with equality for parallel cell vectors of lengths proportional to p_j^{-1}. Proposition D.1 verifies the structural and small-ball requirements; realizability requires its additional conditional-law hypothesis or another justification, as well as the loss and sampling assumptions. The compatibility constant does not force observed error to grow with J. No change-point-adaptive TV rate is claimed. A TV-specific compatibility argument would be additional theory.

F.5 A Global Rate for the Dictionary Class

The next two propositions bound processes indexed by the dictionary itself. The small-ball widths use this class, whereas the distance theorem's multiplier radius uses its contraction hull. The explicit rate corollary assumes realizability and optional geodesic closure, so the hull equals the dictionary. Without closure, the general distance theorem remains available under realizability, but a rate requires a separate hull bound.

Proposition F.7 (Vanishing dictionary fixed points)**.** *Grant Assumption F.1 and let $\mathcal{H} = \mathcal{V}_p^{B,K} - \mathcal{V}_p^{B,K}$. Then*

$$n \ge K/\zeta_1^2 \implies \omega_G(\mathcal{H}, n, \zeta_1) = 0, \qquad n \ge 32K/\zeta_2^2 \implies \omega_R(\mathcal{H}, n, \zeta_2) = 0,$$

and hence $\omega_Q(\zeta_1, \zeta_2) = 0$ as soon as $n \ge K\max\{\zeta_1^{-2}, 32\zeta_2^{-2}\}$.

Proof. Write $\Phi_i : \mathbb{R}^K \to \mathcal{T}_p\mathcal{M}$ for the linear map $\beta \mapsto \sum_k \beta_k \phi_k(T_i)$ and $\hat{\Sigma} = \frac{1}{n}\sum_i \Phi_i^*\Phi_i$, a positive semidefinite $K \times K$ matrix. Orthonormality of the dictionary gives the exact identity

$$\mathbb{E}\big[\operatorname{tr}\hat{\Sigma}\big] = \sum_{k=1}^K \mathbb{E}\big\|\phi_k(T)\big\|_{\mathbf{g}_p}^2 = \sum_{k=1}^K \|\phi_k\|_{L_2(\mathbb{P}_T)}^2 = K. \tag{34}$$

Gaussian part. By the isometry (31), $h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$ corresponds to $\beta \in \mathbb{R}^K$ with $\|\beta\|_2 = r$, and in an orthonormal basis the isonormal process satisfies $G_h = \langle g, \beta\rangle$ for a standard Gaussian vector g in $\mathbb{R}^K$. Hence

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}\cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} G_h\Big] \le \mathbb{E}\sup_{\|\beta\|_2 = r}\langle g, \beta\rangle = r\,\mathbb{E}\|g\|_2 \le r\sqrt{K},$$

by Jensen. The defining inequality $\mathbb{E}[\sup] \le \zeta_1 r\sqrt{n}$ therefore holds *for every* $r > 0$ as soon as $\sqrt{K} \le \zeta_1\sqrt{n}$; the defining set is then all of $(0,\infty)$ and its infimum is 0.

Rademacher part. For $h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, write $h = \sum_k \beta_k\phi_k$, so $\|\beta\|_2 = r$ and $h(T_i) = \Phi_i\beta$. We enlarge this coefficient set to $\{\delta : \|\delta\|_2 \le 2r\}$; the factor two is conservative and keeps the stated sufficient sample-size condition. The map $u \mapsto \|u\|_{\mathbf{g}_p}$ is 1-Lipschitz and vanishes at 0, so the vector-contraction inequality of Maurer (2016, Theorem 3 and Corollary 4) gives, conditionally on the sample, with independent coordinate-sign vectors $\xi_i \in \{-1,1\}^d$,

$$\mathbb{E}_\varepsilon \sup_{\|\delta\|_2\le 2r} \frac{1}{n}\sum_i \varepsilon_i \|\Phi_i\delta\|_{\mathbf{g}_p} \le \frac{\sqrt{2}}{n}\mathbb{E}_\xi \sup_{\|\delta\|_2\le 2r}\Big\langle \sum_i \Phi_i^*\xi_i, \delta\Big\rangle = \frac{2\sqrt{2}r}{n}\mathbb{E}_\xi\Big\|\sum_i \Phi_i^*\xi_i\Big\|_2.$$

By Jensen and independence, $\mathbb{E}_\xi \| \sum_i \Phi_i^* \xi_i \|_2 \le (\sum_i \|\Phi_i\|_F^2)^{1/2} = (n \operatorname{tr} \hat{\Sigma})^{1/2}$, so the right-hand side is at most $2\sqrt{2}\, r\, (\operatorname{tr} \hat{\Sigma}/n)^{1/2}$. The supremum of the absolute value costs a further factor 2, the process being symmetric in ε; taking the expectation over the sample and using Jensen with (34),

$$\mathbb{E}\Big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} \Big| \frac{1}{n} \sum_i \varepsilon_i \|h(T_i)\|_{\mathbf{g}_p} \Big| \Big] \;\le\; 4\sqrt{2}\, r \sqrt{K/n}.$$

This is at most $\zeta_2 r$ for every $r > 0$ as soon as $4\sqrt{2}\sqrt{K/n} \le \zeta_2$, i.e. $n \ge 32K/\zeta_2^2$, and the infimum is again 0. □

The sufficient sample-size conditions follow from upper bounds linear in r, uniformly over positive radii. Exact homogeneity of the constrained localized widths is not required and is not asserted. These fixed points concern the dictionary difference class only.

Proposition F.8 (Restricted dictionary multiplier radius)**.** *Grant Assumption F.1, the geometric comparison hypotheses, and the measurable uniform gradient bound L. Fix $\beta^\star \in \mathcal{B}_K$ and put $\gamma_\beta = \mathsf{Exp}_p \circ v_\beta$. Define*

$$S_r = \{\beta \in \mathcal{B}_K : \mathsf{D}_{\mathbf{g}}(\gamma_\beta, \gamma_{\beta^\star}) \le r\}, \quad \phi_{n,K}^{\ell}(r) = \sup_{\beta \in S_r} \Big| \frac{1}{\sqrt{n}} \sum_i \varepsilon_i a_i(\beta) \Big|,$$

where $a_i(\beta) = \langle Z_i, \mathsf{Log}_{\gamma_{\beta^\star}(T_i)} \gamma_\beta(T_i) \rangle$ and $Z_i = \mathsf{grad}\ell(\cdot, Y_i)(\gamma_{\beta^\star}(T_i))$. Let $\alpha_{n,K}^{\ell}(\zeta_3, \delta)$ be defined as in (15)*'s subsequent radius definition, but with this restricted process and the same floor. For $\zeta_3 > 0$ and $0 < \delta < 1$,*

$$\alpha_{n,K}^{\ell}(\zeta_3, \delta) \le \max\left\{ \frac{2\sqrt{2}\, L\kappa_B}{\delta \zeta_3} \sqrt{\frac{K}{n}},\; r_0(\zeta_3) \right\}.$$

We write $c_D = 2\sqrt{2}$ for this numerical constant.

Proof. Write $A = L\kappa_B$ and $\Phi_i b = \sum_k b_k \phi_k(T_i)$. Nonexpansion of Log and the upper exponential comparison give

$$|a_i(\beta) - a_i(\beta')| \le A \|\Phi_i(\beta - \beta')\|_{\mathbf{g}_p}, \qquad a_i(\beta^\star) = 0.$$

The lower comparison and orthonormality imply $\|\beta - \beta^\star\|_2 \le r$ for $\beta \in S_r$. The pointwise constraints make $\mathcal{B}_K$ closed, while orthonormality and the radius bound make it bounded. The upper comparison makes the population distance continuous in coefficients. Thus S_r is compact; sampled coefficient functions are continuous, so their suprema equal suprema over a countable dense subset.

Condition on the observations. Applying the vector-contraction theorem (Maurer, 2016, Theorem 3) to the displayed increment bound, with independent coordinate signs ξ_i, gives

$$\begin{aligned} \mathbb{E}_\varepsilon \sup_{\beta \in S_r} \sum_i \varepsilon_i a_i(\beta) &\le \sqrt{2} A\, \mathbb{E}_\xi \sup_{\beta \in S_r} \Big\langle \sum_i \Phi_i^* \xi_i, \beta - \beta^\star \Big\rangle \\ &\le \sqrt{2} A r\, \mathbb{E}_\xi \Big\| \sum_i \Phi_i^* \xi_i \Big\|_2 \le \sqrt{2} A r \Big(\sum_i \|\Phi_i\|_F^2 \Big)^{1/2}. \end{aligned}$$

The last inequality follows by Cauchy–Schwarz: expansion of the squared norm eliminates cross terms between independent signs, leaving $\sum_i \|\Phi_i\|_F^2 = n \operatorname{tr} \hat{\Sigma}$. The comparator belongs to S_r, so the one-sided suprema are nonnegative. Bounding the absolute supremum by the sum of the positive and negative suprema costs a factor two. Divide by $\sqrt{n}$, average over the observations, and use Jensen and (34) to obtain

$$\mathbb{E}\phi_{n,K}^{\ell}(r) \le 2\sqrt{2} A r \sqrt{K}.$$

Markov's inequality therefore yields

$$\mathbb{P}\{\phi_{n,K}^{\ell}(r) > \zeta_3 r^2 \sqrt{n}\} \le \frac{2\sqrt{2} A \sqrt{K}}{\zeta_3 r \sqrt{n}}.$$

For $r \ge 2\sqrt{2} A \sqrt{K/n}/(\delta \zeta_3)$ this is at most δ, proving the radius upper bound. This bounds the dictionary-indexed process, not automatically the hull process. □

Assumption F.3 (Geodesic closure for the explicit dictionary rate)**.** For every $\gamma_0, \gamma_1 \in \Gamma_{\mathcal{V}}$ and every $a \in [0,1]$, the curve $t \mapsto \mathsf{Exp}_{\gamma_0(t)}\big(a\mathsf{Log}_{\gamma_0(t)}\gamma_1(t)\big)$ belongs to $\Gamma_{\mathcal{V}}$.

This additional assumption is used for the explicit dictionary rate. It is not required by the general oracle inequality and is not automatic for a Fourier dictionary.

Corollary F.9 (Global rate for geodesically closed dictionaries)**.** *Grant all hypotheses of Theorem E.3, including ambient-oracle realizability in the dictionary, and additionally Assumptions F.1 and F.3, with $\mathcal{V} = \mathcal{V}_p^{B,K}$. Then $\widetilde{\Gamma}_{\mathcal{V}} = \Gamma_{\mathcal{V}}$ and its multiplier process is the restricted dictionary process. These conditions hold for the cell model under the additional conditional-law hypothesis of Proposition D.1 and the stipulated loss and sampling assumptions. Then take $L > 0$, $K \geq 1$, $\zeta_3 = \varkappa_0/(8\kappa_B^2)$ and $n \geq K\max\{\zeta_1^{-2}, 32\zeta_2^{-2}\}$. The following bounds hold with probability at least $1 - \delta - 5\exp(-c_{\mathcal{A}}(\eta,\tau)Q_{\mathcal{H}}^2(\tau)n)$, with $C = 8192$:*

$$\big\|\hat{v}_n^{\lambda} - v^{\star}\big\|_{L_2(\mathbb{P}_T)}^2 \leq C\frac{L^2\kappa_B^6}{\delta^2\varkappa_0^2}\frac{K}{n} + \frac{\lambda\Omega(v^{\star})}{\varkappa},$$

$$\mathsf{D}_{\mathsf{g}}^2(\hat{\gamma}, \gamma^{\star}) \leq \kappa_B^2 \times (\textit{the same bound}).$$

For $\lambda = 0$, this conditional bound has order K/n for a fixed model. A growing-K assertion additionally requires uniform control of $L, \kappa_B, \varkappa_0, Q_{\mathcal{H}}(\tau), \tau$, tuning choices and the parent probability bound. It is not a sparsity-adaptive or optimality assertion.

Proof. For the dictionary process, Proposition F.7 gives $\omega_Q = 0$. Choose $\bar{r} = r_0(\zeta_3)/2 > 0$. Under the stated assumptions, its radius is then $r^{\star} = \alpha_{n,K}^{\ell}(\zeta_3, \delta/4)$. The floor is $4\sqrt{2}L\kappa_B^2/(\varkappa_0\sqrt{n})$, and Proposition F.8 bounds the other term by

$$\frac{32c_D L\kappa_B^3\sqrt{K}}{\delta\varkappa_0\sqrt{n}}.$$

The factor 32 includes the substitution $\delta \mapsto \delta/4$. With $c_D = 2\sqrt{2}$, this term dominates the floor for $K \geq 1$, $\delta < 1$ and $\kappa_B \geq 1$. Squaring gives the admissible coefficient 8192 for the displayed radius contribution. Substitution into the conditional parent bounds gives the result. Ambient optimality supplies the first-order condition; realizability makes the oracle feasible. The additional closure assumption identifies the hull with the dictionary and justifies use of Proposition F.8. The calculation does not supply the same rate for an enlarged hull. □

Remark F.1 (The global rate is not sparsity-adaptive). Under the stated realizability and additional closure assumptions, combining Corollary F.9 with Corollary F.4 gives

$$\|\hat{v} - v^{\star}\|_{L_2(\mathbb{P}_T)} \leq \max\{O(\sqrt{K/n}),\ \lambda\sqrt{s}/\varkappa\}.$$

At the level $\lambda_n \asymp \sqrt{\log K/n}$ the second term is $O(\sqrt{s\log K/n})$. The first can therefore dominate when K is large relative to $s\log K$, provided the hidden constants are comparable, which is the regime the sparsity regularization is designed for. The global rate is thus the parametric rate of the *full* K-dimensional class, and the sparsity level enters the bound of Corollary F.4 without improving this rate—exactly the gap already noted after that corollary. Closing it would require localizing Definitions 1, 2 and the multiplier radius to Ω-balls, as in Lecué & Mendelson (2018), which we do not do here.

Remark F.2 (Fourier models need not be geodesically closed). In the hyperboloid model of $\mathbb{H}^2$ at $p = (1,0,0)$, take $v_{\pm}(t) = (\pm a\cos(2\pi Nt), b)$, with $a, b > 0$ and $a^2 + b^2 < B^2$. Both fields lie in a Fourier dictionary of maximum frequency N. Write $r(t) = \sqrt{a^2\cos^2(2\pi Nt) + b^2}$. The exponential formula $\mathsf{Exp}_p(u) = (\cosh\|u\|, \sinh(\|u\|)u/\|u\|)$ shows that the geodesic midpoint of their exponential images has logarithm $(0, m(t))$ at p, where

$$m(t) = \operatorname{artanh}\left(\frac{b}{r(t)}\tanh r(t)\right).$$

Indeed, the midpoint is the Lorentz-normalized sum of the two images. The function m has period $1/(2N)$ and is nonconstant: it equals b when the cosine vanishes and is smaller otherwise, since $\tanh r/r$ is strictly decreasing for $r > 0$. A trigonometric polynomial of maximum frequency N with this period must be constant. Thus the midpoint leaves the dictionary. This proves the failure of automatic closure; it does not prevent the ambient-oracle theorem from applying under realizability with its hull radius. It does prevent replacing that radius by the restricted dictionary bound solely on the basis of dictionary membership.

F.6 SUMMARY

Table 4 collects the compatibility constant $\Psi(\bar{M})$ and the resulting regularization term of (30) for each regularization treated above, in the format of Table 1 of Alquier et al. (2019); the full bound is the maximum of this term and of the localization radius $r^\star$.

Table 4: Compatibility bounds and regularization terms for the regularizations of Sections F.2–F.4, in the format of Table 1 of Alquier et al. (2019). Each bound is $\max\{r^\star, \text{regularization term}\}$ with $\varkappa$ the curvature/margin constant of Theorem E.3; the LASSO row uses $\lambda = \lambda_n$ of Corollary F.4.

Regularization	upper bound on $\Psi(M)$	regularization term of (30)
Ω_{lasso} (dictionary, Corollary F.4)	$\sqrt{s}$	$\dfrac{2LM_\phi\sqrt{s}}{\varkappa}\sqrt{\dfrac{2\log(2K)}{n}}$
Ω_{group} (sharpened, Corollary F.5)	$\sqrt{s}/\sqrt{c_-}$	$\dfrac{\lambda\sqrt{s}}{\varkappa\sqrt{c_-}}$
Ω_{time} (grid, Corollary F.6)	$\sqrt{sJ/c_{\text{grid}}}$	$\dfrac{\lambda}{\varkappa}\sqrt{\dfrac{sJ}{c_{\text{grid}}}}$

All three regularization terms are of order $\lambda\sqrt{s}/\varkappa$, up to a logarithmic factor (LASSO), the design factor $1/\sqrt{c_-}$ (group) or the grid-resolution factor $\sqrt{J/c_{\text{grid}}}$ (time-local). They replace the norm-dependent term of Theorem E.3 and Corollary E.4 by one governed by the sparsity level; the bounds of this section are on $\|\hat{v} - v^\star\|_{L_2(\mathbb{P}_T)}$ and those of Theorem E.3 on its square, so the two must be put on the same scale before being compared.

G COMPARISON ESTIMATES BETWEEN TANGENT-SPACE AND GEODESIC DISTANCE

The estimator is represented in a fixed tangent space, whereas its loss measures geodesic prediction. This appendix proves the two pointwise comparisons used in Lemma E.1. Nonpositive curvature gives the lower bound; a curvature lower bound on the prediction ball gives the upper bound. We first state the differential comparison and then integrate it along appropriate paths.

Let $\mathcal{M}$ be a Hadamard manifold and let $\kappa_{\min} \le 0 \le \kappa_{\max}$ be real numbers. In Lemma G.1 below the bounds $\kappa_{\min} \le \kappa \le \kappa_{\max}$ are required only for the sectional curvatures of the 2-planes containing $\dot{c}(r)$, at the points of the geodesic segment $c([0, s])$ along which the lemma is applied, and each half of its conclusion uses only the corresponding one-sided bound. For abbreviation, we define for $\rho \in \mathbb{R}$

$$s_\rho(t) = \begin{cases} \dfrac{1}{\sqrt{\rho}}\sin(\sqrt{\rho}\,t), & \text{if } \rho > 0, \\ t, & \text{if } \rho = 0, \\ \dfrac{1}{\sqrt{-\rho}}\sinh(\sqrt{-\rho}\,t), & \text{if } \rho < 0. \end{cases}$$

This function solves the scalar Jacobi equation for constant sectional curvature ρ, $\ddot{f}(t) + \rho f(t) = 0$, with initial values $f(0) = 0,\ \dot{f}(0) = 1$.

Lemma G.1 (See Corollaries 4.2.2 and 4.6.1 in Jost, 2008)**.** *Let $c : [0, 1] \to \mathcal{M}$, $c(s) = \mathsf{Exp}_p(s\dot{c}(0))$, be a geodesic. For $v \in \mathcal{T}_p\mathcal{M}$, let J be the Jacobi field with $J(0) = 0$ and $\dot{J}(0) = v$, where the dot on a vector field denotes covariant differentiation along c. For every $s \in [0, 1]$,*

$$J(s) = (\mathrm{D}_{s\dot{c}(0)}\mathsf{Exp}_p)(sv).$$

Furthermore, let $\langle v, \dot{c}(0)\rangle_{\mathsf{g}_p} = 0$. Then, provided that $s\|\dot{c}(0)\|_{\mathsf{g}_p} \le \frac{\pi}{\sqrt{\kappa_{\max}}}$ when $\kappa_{\max} > 0$, we have the following estimate for the derivative of the exponential map:

$$\|v\|_{\mathsf{g}_p}\,\frac{s_{\kappa_{\max}}(s\|\dot{c}(0)\|_{\mathsf{g}_p})}{s\|\dot{c}(0)\|_{\mathsf{g}_p}} \le \big\|(\mathrm{D}_{s\dot{c}(0)}\mathsf{Exp}_p)v\big\|_{\mathsf{g}_{\mathsf{Exp}_p(s\dot{c}(0))}} \le \|v\|_{\mathsf{g}_p}\,\frac{s_{\kappa_{\min}}(s\|\dot{c}(0)\|_{\mathsf{g}_p})}{s\|\dot{c}(0)\|_{\mathsf{g}_p}}. \tag{35}$$

If v is radial, then

$$(\mathrm{D}_{s\dot{c}(0)}\mathsf{Exp}_p)v = P^c_{0\to s}v, \qquad \|P^c_{0\to s}v\|_{\mathsf{g}_{c(s)}} = \|v\|_{\mathsf{g}_p},$$

where $P^c_{0\to s}$ denotes parallel transport along c. The vectors before and after differentiation belong to different tangent spaces. At $s = 0$ or $\dot{c}(0) = 0$, read the displayed ratios by continuity as 1 and use $D_0\mathsf{Exp}_p = \mathrm{Id}$.

Lemma G.2 (Gauss's lemma, in the form used below)**.** *Let $x \in \mathcal{T}_p\mathcal{M} \setminus \{0\}$ and $y \in \mathcal{T}_p\mathcal{M}$. Then*

$$\big\langle (\mathrm{D}_x\mathsf{Exp}_p)(x),\, (\mathrm{D}_x\mathsf{Exp}_p)(y)\big\rangle_{\mathsf{g}_{\mathsf{Exp}_p x}} = \langle x, y\rangle_{\mathsf{g}_p}.$$

In particular, taking $y = x$ gives $\|(\mathrm{D}_x\mathsf{Exp}_p)(x)\|_{\mathsf{g}_{\mathsf{Exp}_p x}} = \|x\|_{\mathsf{g}_p}$ (the radial direction is preserved), and if $\langle x, y\rangle_{\mathsf{g}_p} = 0$ then $(\mathrm{D}_x\mathsf{Exp}_p)(x)$ and $(\mathrm{D}_x\mathsf{Exp}_p)(y)$ are orthogonal in $\mathcal{T}_{\mathsf{Exp}_p x}\mathcal{M}$: the decomposition of a tangent vector into radial and orthogonal parts is carried by $\mathrm{D}_x\mathsf{Exp}_p$ into an orthogonal decomposition.

Proposition G.3. *(Corollary 4.8.1 in Jost, 2008) The exponential map is distance nondecreasing, i.e.*

$$\|v - w\|_{\mathsf{g}_p} \le \mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p v, \mathsf{Exp}_p w) \quad \textit{for all } v, w \in \mathcal{T}_p\mathcal{M}$$

The lower bound needs no curvature lower bound. To control expansion in the opposite direction, we use Assumption 2 along radial geodesics contained in the prediction ball.

Proposition G.4 (Upper Lipschitz bound for Exp_p on $\mathcal{B}(p, B)$)**.** *Under Assumption 2, for all $v, w \in \mathcal{V}$ and $\mathbb{P}_T$-a.e. $t \in [0, 1]$,*

$$\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p(v(t)), \mathsf{Exp}_p(w(t))) \;\le\; \kappa_B\, \|v(t) - w(t)\|_{\mathsf{g}_p},$$

where $\kappa_B = \frac{\sinh(\Lambda_B B)}{\Lambda_B B}$, interpreted as $\kappa_B = 1$ when $\Lambda_B = 0$ (its limit as $\Lambda_B \downarrow 0$). In particular $\kappa_B > 1$ strictly whenever $\Lambda_B > 0$.

Proof. Fix $t \in [0, 1]$ and write $u = v(t), w_0 = w(t) \in \{x \in \mathcal{T}_p\mathcal{M} : \|x\|_{\mathsf{g}_p} \le B\}$; we must bound $\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w_0)$. Let $\sigma : [0, 1] \to \mathcal{T}_p\mathcal{M}$ be the straight segment $\sigma(s) = w_0 + s(u - w_0)$, so that $\sigma(0) = w_0$, $\sigma(1) = u$, and $\gamma(s) = \mathsf{Exp}_p(\sigma(s))$ is a smooth curve from $\mathsf{Exp}_p(w_0)$ to $\mathsf{Exp}_p(u)$. Hence,

$$\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w_0) \le \mathcal{L}(\gamma) = \int_0^1 \big\|\mathrm{D}_{\sigma(s)}\mathsf{Exp}_p(\dot\sigma(s))\big\|_{\mathsf{g}_{\mathsf{Exp}_p(\sigma(s))}}\, \mathrm{d}s.$$

Decompose $\dot\sigma(s) = u - w_0$ into its component along the radial direction $\sigma(s)$ (when $\sigma(s) \neq 0$) and an orthogonal component, exactly as in the proof of Proposition G.3: writing $\dot\sigma(s) = a(s)\sigma(s) + \eta(s)$ with $\langle \eta(s), \sigma(s)\rangle_{\mathsf{g}_p} = 0$, Lemma G.1 applied along the geodesic $c(r) = \mathsf{Exp}_p(r\,\sigma(s))$ with $\kappa_{\min} = -\Lambda_B^2$ (valid since $\|\sigma(s)\|_{\mathsf{g}_p} \le \max(\|u\|_{\mathsf{g}_p}, \|w_0\|_{\mathsf{g}_p}) \le B$, so the whole segment $\{\mathsf{Exp}_p(r\sigma(s)) : r \in [0, 1]\} \subset \mathcal{B}(p, B) \subset \mathcal{D}_B$, on which the curvature lower bound $-\Lambda_B^2$ holds by Assumption 2) gives

$$\big\|\mathrm{D}_{\sigma(s)}\mathsf{Exp}_p(\eta(s))\big\|_{\mathsf{g}_{\mathsf{Exp}_p(\sigma(s))}} \le \|\eta(s)\|_{\mathsf{g}_p}\, \frac{s_{-\Lambda_B^2}(\|\sigma(s)\|_{\mathsf{g}_p})}{\|\sigma(s)\|_{\mathsf{g}_p}} = \|\eta(s)\|_{\mathsf{g}_p}\, \frac{\sinh(\Lambda_B\|\sigma(s)\|_{\mathsf{g}_p})}{\Lambda_B\|\sigma(s)\|_{\mathsf{g}_p}},$$

while the radial part satisfies $\|\mathrm{D}_{\sigma(s)}\mathsf{Exp}_p(a(s)\sigma(s))\|_{\mathsf{g}_{\mathsf{Exp}_p(\sigma(s))}} = \|a(s)\sigma(s)\|_{\mathsf{g}_p}$ by Lemma G.1. Write $\kappa_B = \sinh(\Lambda_B B)/(\Lambda_B B)$; since $r \mapsto \sinh(\Lambda_B r)/(\Lambda_B r)$ is nondecreasing on $[0, B]$ and $\|\sigma(s)\|_{\mathsf{g}_p} \le B$, each of the two pushed-forward components is bounded by κ_B times the norm of the corresponding component of $\dot\sigma(s)$ (using $\kappa_B \ge 1$ for the radial one). The two pushed-forward components are orthogonal by Lemma G.2, and so are $a(s)\sigma(s)$ and $\eta(s)$, whence

$$\big\|\mathrm{D}_{\sigma(s)}\mathsf{Exp}_p(\dot\sigma(s))\big\|^2_{\mathsf{g}_{\mathsf{Exp}_p(\sigma(s))}} \;\le\; {\kappa_B}^2\big(\|a(s)\sigma(s)\|^2_{\mathsf{g}_p} + \|\eta(s)\|^2_{\mathsf{g}_p}\big) = {\kappa_B}^2\|\dot\sigma(s)\|^2_{\mathsf{g}_p}.$$

As $\dot\sigma(s) = u - w_0$ is constant in s, integrating the last display gives

$$\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w_0) \le \int_0^1 \frac{\sinh(\Lambda_B B)}{\Lambda_B B}\, \|u - w_0\|_{\mathsf{g}_p}\, \mathrm{d}s = \frac{\sinh(\Lambda_B B)}{\Lambda_B B}\, \|u - w_0\|_{\mathsf{g}_p},$$

which is the claimed bound with $\kappa_B = \sinh(\Lambda_B B)/(\Lambda_B B)$. Finally $\kappa_B > 1$ whenever $\Lambda_B B > 0$, because $\sinh(x) - x = \sum_{k\geq 1} x^{2k+1}/(2k+1)! > 0$ for $x > 0$, while $\kappa_B = 1$ in the flat case $\Lambda_B = 0$; combined with Proposition G.3 (which gives the reverse, unconditional bound $\|u - w_0\|_{\mathsf{g}_p} \leq \mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w_0)$), this shows that (16) is a genuine two-sided norm equivalence, giving equality throughout when $\Lambda_B = 0$. Equality can also hold for particular pairs in curved spaces, for example along a radial geodesic.

At every s with $\sigma(s) = 0$, the identity $D_0\mathsf{Exp}_p = \mathrm{Id}$ gives the same estimate, including the constant zero segment. If $\Lambda_B = 0$, use $s_0(x) = x$ and the continuous value 1 at $x = 0$ throughout. This covers all degenerate cases without dividing by zero. □

Remark G.1 (Sharpness of the universal comparison factor). Proposition G.4 is proved by integrating along the straight segment of $\mathcal{T}_p\mathcal{M}$ joining the two tangent vectors, and one may ask whether a different path, or a different argument, would give a smaller constant. It would not: on the hyperbolic plane $\mathbb{H}^2$ of curvature -1, so that $\Lambda_B = 1$, the bound is attained in the limit. Take $u, w \in \mathcal{T}_p\mathbb{H}^2$ both of norm B and making an angle θ, so that $\|u - w\|_{\mathsf{g}_p} = 2B\sin(\theta/2)$, while the hyperbolic law of cosines gives

$$\cosh \mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w) = \cosh^2 B - \sinh^2 B \, \cos\theta.$$

As $\theta \downarrow 0$ the right-hand side is $1 + \frac{1}{2}\sinh^2 B\,\theta^2 + O(\theta^4)$, so $\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w) = \sinh(B)\,\theta + O(\theta^3)$ while $\|u - w\|_{\mathsf{g}_p} = B\theta + O(\theta^3)$, and therefore

$$\frac{\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w)}{\|u - w\|_{\mathsf{g}_p}} \;\longrightarrow\; \frac{\sinh B}{B} \;=\; \kappa_B.$$

Thus the factor cannot be reduced uniformly over Hadamard manifolds with the stated curvature bound. The lower comparison is also sharp, by taking $w = 0$. This does not show that every manifold or target attains the upper factor, nor that its powers in the statistical bounds are optimal. The growth $\kappa_B \sim e^{\Lambda_B B}/(2\Lambda_B B)$ nevertheless explains why the displayed bounds can be numerically uninformative at the EEG radius.

Remark G.2 (Pointwise statement and its use for tangent fields). The content of Proposition G.4 is the pointwise inequality

$$\mathsf{d}_{\mathsf{g}}(\mathsf{Exp}_p u, \mathsf{Exp}_p w) \;\leq\; \kappa_B \, \|u - w\|_{\mathsf{g}_p} \qquad \text{for all } u, w \in \mathcal{T}_p\mathcal{M} \text{ with } \|u\|_{\mathsf{g}_p}, \|w\|_{\mathsf{g}_p} \leq B,$$

which is what the proof establishes; nothing about the model class enters. The statement is phrased for $v, w \in \mathcal{V}$ and $\mathbb{P}_T$-a.e. t only because that is the form in which Lemma E.1 consumes it: squaring, integrating in t against $\mathbb{P}_T$, and combining with the unconditional lower bound of Proposition G.3 gives the two-sided comparison of $\|v - v^\star\|_{L_2(\mathbb{P}_T)}$ and $\mathsf{D}_{\mathsf{g}}(\gamma, \gamma^\star)$. The only requirement on $\mathcal{V}$ is that its elements take values in the closed B-ball $\mathbb{P}_T$-a.e., which holds by definition of $\mathcal{V}_p^B \supseteq \mathcal{V}$.

H Concentration and empirical process tools

This appendix collects the probabilistic tools used in the proof of the empirical lower bound, Theorem E.2: an elementary deviation inequality, the small-ball coordinate-selection lemmas (Lemmas H.3–H.5), the uniform selection lemma (Lemma H.7) and its consequence Lemma H.8, the Giné–Zinn symmetrization theorem, and the multiplier-process bound Lemma H.10.

We begin by citing several lemmas that will be used later.

Lemma H.1 (Bounded i.i.d. summands). *Let* $\left\{X_i\right\}_{i=1}^{n}$ *be a sample of i.i.d random variables,* $\left|X_i\right| \leq M$, *and* $\mathbb{E}\left[X_i\right] = X^*$. *Then, for every* $x > 0$, *with probability at least* $1 - 2\exp\left(\frac{-x^2}{2nM^2}\right)$ *we have*

$$nX^* - x \leq \sum_{i=1}^{n} X_i \leq nX^* + x.$$

Lemma H.2 (Bounded differences / McDiarmid inequality). *(Boucheron et al., 2013, Thm. 6.2.) Let* $T_1, \ldots, T_n$ *be independent random variables taking values in a set* $\mathcal{T}$, *and let* $F : \mathcal{T}^n \to \mathbb{R}$ *be*

measurable and satisfy the bounded-differences property with constants $a_1, \dots, a_n$: for every i and every $t_1, \dots, t_n, t_i' \in \mathcal{T}$,

$$\big|F(t_1, \dots, t_i, \dots, t_n) - F(t_1, \dots, t_i', \dots, t_n)\big| \le a_i.$$

Then, for every $s > 0$,

$$\mathbb{P}\big[F(T_1, \dots, T_n) \ \ge \ \mathbb{E}\big[F(T_1, \dots, T_n)\big] + s\big] \le \exp\left(- \frac{2s^2}{\sum_{i=1}^n a_i^2}\right).$$

In particular, if $F(t_1, \dots, t_n) = \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n g(t_i)$ for a class $\mathcal{G}$ of functions with values in $[0,1]$, then $a_i = 1/n$ for every i and the bound reads $\mathbb{P}[F \ge \mathbb{E}F + s] \le \exp(-2ns^2)$.

In the following coordinate-selection lemmas, fix $\tau > 0$ with $Q_{\mathcal{H}}(\tau) > 0$ as in Assumption 7.

Lemma H.3. *Let $h \in \mathcal{H}$ and let $T_i, i = 1, 2, \dots, n$, be a sample of i.i.d. copies of $T \sim \mathbb{P}_T$. Then, for every $0 < \alpha < 1$, with probability at least $1 - 2\exp\big(-\frac{n\alpha^2}{2}\big)$, there exists a subset $S \subset \{1, 2, \cdots, n\}$ such that $|S| \ge n(1-2\alpha)\Big)$, and $\forall i \in S$,*

$$\|h(T_i)\|_{\mathbf{g}_p} \le \frac{\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{\alpha}}.$$

Proof. If $\|h\|_{L_2(\mathbb{P}_T)} = 0$, then $h(T_i) = 0$ for all i almost surely, and the conclusion holds with $S = \{1, \dots, n\}$. Hence assume $\|h\|_{L_2(\mathbb{P}_T)} > 0$.

Let $X_i = \mathbf{1}\big(\|h(T_i)\|_{\mathbf{g}_p} \ge \frac{\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{\alpha}}\big)$. Then, by Markov's inequality,

$$\begin{aligned}
\mathbb{E}\Big[X_i\Big] &= \mathbb{E}\Big[\mathbf{1}\big(\|h(T_i)\|_{\mathbf{g}_p} \ge \frac{\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{\alpha}}\big)\Big] \\
&= \mathbb{P}\big(\|h(T_i)\|_{\mathbf{g}_p} \ge \frac{\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{\alpha}}\big) \\
&= \mathbb{P}\big(\|h(T_i)\|_{\mathbf{g}_p}^2 \ge \frac{\|h\|_{L_2(\mathbb{P}_T)}^2}{\alpha}\big) \\
&\le \alpha.
\end{aligned}$$

By using Lemma H.1 (applied with $X_i \in \{0,1\}$, $M = 1$, $X^* = \mathbb{E}[X_i] \le \alpha$), we get $\sum_{i=1}^n X_i \le n\alpha + x$ with probability at least $1 - 2\exp\big(\frac{-x^2}{2n}\big)$. Setting $x = n\alpha$, we obtain $\sum_{i=1}^n X_i \le 2n\alpha$ with probability at least $1 - 2\exp\big(-\frac{n\alpha^2}{2}\big)$. Define $S = \{i \in \{1, \dots, n\} : X_i = 0\}$. Then $|S| = n - \sum_{i=1}^n X_i \ge n\big(1-2\alpha\big)$. Moreover, for every $i \in S$, we have $X_i = 0$. Therefore, with probability at least $1 - 2\exp\big(-\frac{n\alpha^2}{2}\big)$, it holds that

$$\|h(T_i)\|_{\mathbf{g}_p} \le \frac{\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{\alpha}}.$$

□

Lemma H.4. *Let $h \in \mathcal{H}$ and let $T_i, i = 1, 2, \dots, n$, be a sample of i.i.d. copies of $T \sim \mathbb{P}_T$. Then there exists a subset $S \subset \{1, 2, \cdots, n\}$ such that $|S| \ge 3nQ_{\mathcal{H}}(\tau)/4$, and $\forall i \in S$*

$$\|h(T_i)\|_{\mathbf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)},$$

with probability at least $1 - 2\exp\big(-\frac{1}{32} n Q_{\mathcal{H}}^2(\tau)\big)$.

Proof. If $\|h\|_{L_2(\mathbb{P}_T)} = 0$, then $h(T_i) = 0$ for all i almost surely, and the conclusion holds with $S = \{1, \dots, n\}$. Hence assume $\|h\|_{L_2(\mathbb{P}_T)} > 0$.

Let $X_i = \mathbf{1}\big(\|h(T_i)\|_{\mathbf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}\big)$. We have

$$\mathbb{E}[X_i] \ = \ \mathbb{E}\big[\mathbf{1}\big(\|h(T_i)\|_{\mathbf{g}_p} \ge \tau \|h\|_{L_2(\mathbb{P}_T)}\big)\big]$$

$$
\begin{aligned}
&= \mathbb{P}\big(\|h(T_i)\|_{\mathbb{g}_p} \geq \tau \|h\|_{L_2(\mathbb{P}_T)}\big) \\
&\geq \inf_{\substack{f \in \mathcal{H} \\ \|f\|_{L_2(\mathbb{P}_T)} > 0}} \mathbb{P}\Big(\|f(T)\|_{\mathbb{g}_p} \geq \tau \|f\|_{L_2(\mathbb{P}_T)}\Big) \\
&= Q_H(\tau).
\end{aligned}
$$

By using Lemma H.1 (with $M = 1$), we get $|\sum_{i=1}^n X_i - n\mathbb{E}(X_i)| \leq t$ with probability at least $1 - 2\exp\big(\frac{-t^2}{2n}\big)$. Setting $t = \frac{nQ_H(\tau)}{4}$, and using $\mathbb{E}(X_i) \geq Q_H(\tau)$, we obtain with probability at least $1 - 2\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$,

$$
\frac{3nQ_H(\tau)}{4} \leq \sum_{i=1}^n X_i \leq \frac{5n\mathbb{E}(X_i)}{4}.
$$

Define $S = \{i \in \{1, \ldots, n\} : X_i = 1\}$. Then $|S| = \sum_{i=1}^n X_i \geq \frac{3nQ_{\mathcal{H}}(\tau)}{4}$. Moreover, for every $i \in S$, we have $X_i = 1$. Therefore, with probability at least $1 - 2\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$, it holds that

$$
\|h(T_i)\|_{\mathbb{g}_p} \geq \tau \|h\|_{L_2(\mathbb{P}_T)}.
$$

□

Lemma H.5. *Let $h \in H$ and let $T_i, i = 1, 2, \ldots, n$, be a sample of i.i.d. copies of $T \sim \mathbb{P}_T$. Then there is a subset $S \subset \{1, 2, \cdots, n\}$ such that $|S| \geq nQ_{\mathcal{H}}(\tau)/4$, and $\forall i \in S$,*

$$
\tau\|h\|_{L_2(\mathbb{P}_T)} \leq \|h(T_i)\|_{\mathbb{g}_p} \leq \frac{2\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{Q_{\mathcal{H}}(\tau)}},
$$

with probability at least $1 - 4\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$.

Proof. Using Lemma H.3, setting $\alpha = \frac{Q_H(\tau)}{4}$, with probability at least $1 - 2\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$, it holds that

$$
\|h(T_i)\|_{\mathbb{g}_p} \leq \frac{2\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{Q_{\mathcal{H}}(\tau)}},
$$

for all $i \in S_1$, where $|S_1| \geq n\big(1 - \frac{Q_{\mathcal{H}}(\tau)}{2}\big)$. Conversely, by Lemma H.4, with probability at least $1 - 2\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$, we have $\|h(T_i)\|_{\mathbb{g}_p} \geq \tau\|h\|_{L_2(\mathbb{P}_T)}$, for all $i \in S_2$, where $|S_2| \geq \frac{3}{4}nQ_{\mathcal{H}}(\tau)$. Hence, for $S = S_1 \cap S_2$ (so that $i \in S$ satisfies both displayed bounds),

$$
\begin{aligned}
|S| &\geq |S_1| + |S_2| - n \\
&\geq n\big(1 - \frac{Q_{\mathcal{H}}(\tau)}{2}\big) + \frac{3}{4}nQ_{\mathcal{H}}(\tau) - n \\
&= \frac{1}{4}\, n\, Q_{\mathcal{H}}(\tau),
\end{aligned}
$$

we have, by the union bound over the two events above, with probability at least $1 - 4\exp\big(-\frac{1}{32}nQ^2_{\mathcal{H}}(\tau)\big)$,

$$
\tau\|h\|_{L_2(\mathbb{P}_T)} \leq \|h(T_i)\|_{\mathbb{g}_p} \leq \frac{2\|h\|_{L_2(\mathbb{P}_T)}}{\sqrt{Q_{\mathcal{H}}(\tau)}}.
$$

□

Here, we recall some definitions, which will be useful for what follows.

Definition H.1 (Maximal ρ-Separated Set)**.** A subset S of a metric space (X, d) is called a *maximal ρ-separated set* if $d(x, y) \geq \rho \quad$ for all distinct $x, y \in S$, and for any $z \in X \setminus S$, the set $S \cup \{z\}$ is no longer ρ-separated.

Definition H.2 (Stable lower bound)**.** (See Mendelson (2021)). A function $h \in L_2([0, 1], \mathcal{T}_p\mathcal{M})$ satisfies a *stable lower bound* with parameters (ξ, q, k) for a sample of cardinality m if, with probability at least $1 - 2\exp(-k)$, for any $J \subset \{1, \ldots, m\}$ with $|J| \leq q$, one has

$$
\frac{1}{m}\sum_{i \in J^c} \|h(T_i)\|^2_{\mathbb{g}_p} \geq (1 - \xi)\|h\|^2_{L_2(\mathbb{P}_T)}.
$$

Definition H.3 (Star-shaped class)**.** (Mendelson (2015)). A class F is said to be *star-shaped around* 0 if for every $f \in F$ and every $0 \le \lambda \le 1$, one has $\lambda f \in F$. Equivalently, F is star-shaped around 0 if it contains the entire interval $[0, f] = \{\lambda f : 0 \le \lambda \le 1\}$ whenever $f \in F$.

Note that $\mathcal{H} = \mathcal{V} - \mathcal{V}$ is convex and contains zero. Under Assumption 6, the same holds for $\mathcal{V} - v^\star$. Thus both are star-shaped under ordinary tangent scaling; no geodesic interpolation is used in the following small-ball lemmas.

Theorem H.6 (Sudakov's minoration inequality: Sudakov Constant c_S)**.** *(Ledoux & Talagrand, 1991, Thm. 3.18). Let $\{G_f\}_{f\in A}$ be a centred Gaussian process indexed by a set A, whose canonical metric is $\big(\mathbb{E}(G_{f_1} - G_{f_2})^2\big)^{1/2}$, and let $M(\alpha, A)$ denote the α-packing number of A in that metric. Then, there is a universal constant $c_S > 0$, such that for every $\alpha > 0$,*

$$c_S \, \frac{\alpha}{2} \sqrt{\log M(\alpha, A)} \le \mathbb{E}\big[\sup_{f\in A} G_f\big]. \tag{36}$$

Remark H.1 (Where c_S enters). Theorem H.6 is used once, to bound the cardinality of the net U_r in Claim 1 of Lemma H.7. It leaves a single trace in the statements of this paper, the branch c_S^2 in $\eta^\star(\tau)$ (Theorem E.2); once an admissible η is fixed, c_S appears in no further displayed bound. Its influence is through the allowed range of η, and hence may enter numerical choices of the remaining constants.

Lemma H.7. *Let $c_0 > 0$ and $c_1 > 1$, and let $h \in \mathcal{H}$. Fix $\eta \in \big(0, \eta^\star(\tau)\big)$ as in* (17)*, where c_S is the Sudakov constant of Theorem H.6. If $\zeta_1 = \frac{c_0 \eta^{3/2} Q_H^2(\tau)}{256 c_1}$ and $\zeta_2 = \frac{c_0 \eta}{16\sqrt{2}} Q_{\mathcal{H}}(\tau)$, and $r = \|h\|_{L_2(\mathbb{P}_T)} > \omega_Q(\zeta_1, \zeta_2)$, there exists a set $U_r \subset \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, such that there is an event $\mathcal{A}_{\mathrm{sb}}(r)$ with probability at least $1 - 5\exp\big[-c_{\mathcal{A}}(\eta, \tau)\, n\, Q_{\mathcal{H}}^2(\tau)\big]$, where*

$$c_{\mathcal{A}}(\eta, \tau) = \min\Big\{\frac{15}{512},\ \frac{c_0^2 \eta^2}{\tau^2}\Big\} > 0,$$

we have:

1. $$|U_r| \le \exp\Big(\frac{Q_{\mathcal{H}}^2(\tau)}{512} n\Big).$$

2. *For every $h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, there exists some $h_u \in U_r$ and a subset $K_h \subset S_{h_u}$ such that $|K_h| \ge \Big(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau}\Big) Q_{\mathcal{H}}(\tau) n$, and for every $k \in K_h$,*

$$\frac{\tau}{2}\|h\|_{L_2(\mathbb{P}_T)} \le \|h(T_k)\|_{\mathsf{g}_p} \le 2\Big(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}}\Big)\|h\|_{L_2(\mathbb{P}_T)}.$$

Remark H.2 (Choosing the free parameters c_0 and c_1). The quantities $c_0 > 0$ and $c_1 > 1$ in Lemma H.7 are *free tuning parameters*, not constants produced by the proof: Claim 1 holds for every admissible pair, because c_0 and c_1 cancel identically in the ratio $\zeta_1 r/\rho$ that drives the Sudakov step (37). They may be chosen freely, and may depend on the fixed small-ball parameter τ of Assumption 7 (as $Q_{\mathcal{H}}(\tau)$ itself does).

For fixed admissible η, c_1, increasing c_0 raises the width thresholds and cannot increase ω_Q, but lowers $\varkappa_0$. If ζ_3 remains fixed and admissible, this can reduce the chosen localization radius; if it must be reduced to preserve its ceiling, the multiplier radius may instead increase. Increasing c_1 lowers ζ_1 and the net radius ρ. Thus the full bound, including the margin and multiplier terms, must guide the choice. Report the selected constants in any numerical instantiation of the theorem.

Proof. Fix $r > \omega_Q(\zeta_1, \zeta_2)$. The finite-packing Sudakov estimate below bounds every finite separated subset, so a greedy construction terminates in a finite maximal separated set. Its ordering is fixed independently of the sample. Let

$$\rho = \frac{c_0}{4\sqrt{2}\, c_1}\, \eta\, Q_{\mathcal{H}}(\tau)\, r,$$

and let $U_r \subset \mathcal{H} \cap rS(L_2)$ be a maximal ρ-separated set (i.e., a ρ-separated set to which no further point of $\mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$ can be added while preserving ρ-separation). We now aim to establish an upper bound for the cardinality of U_r.

Recall from Definition 1 that $\{G_h; h \in \mathcal{H}\}$ is the canonical Gaussian process indexed by H, with

$$\big(\mathbb{E}(G_{h_1} - G_{h_2})^2\big)^{1/2} = \|h_1 - h_2\|_{L_2(\mathbb{P}_T)} \qquad \text{for all } h_1, h_2 \in H,$$

i.e., its canonical metric is exactly the $L_2(\mathbb{P}_T)$ metric on $\mathcal{H}$, with no norm-composition of any kind. Since $U_r \subset \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$ is ρ-separated (for the value of ρ fixed above), U_r is in particular a valid ρ-packing of $\mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, so

$$|U_r| \le M\big(\rho, \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}, \|\cdot\|_{L_2(\mathbb{P}_T)}\big),$$

where $M(\beta, \cdot, \cdot)$ denotes the β-packing number. We now apply Theorem H.6, Sudakov's minoration inequality, with its universal constant $c_S > 0$ (see, e.g., Ledoux & Talagrand, 1991), to the canonical metric of $\{G_h\}$, which is exactly $\|\cdot\|_{L_2(\mathbb{P}_T)}$ as recalled above:

$$c_S\,\frac{\rho}{2}\,\sqrt{\log M\big(\rho, \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}, \|\cdot\|_{L_2(\mathbb{P}_T)}\big)} \le \mathbb{E}\Big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} G_h\Big]. \tag{37}$$

Since $r > \omega_G(H, n, \zeta_1)$, Definition 1 together with Lemma D.2 (applied to the star-shaped set H and the linear process G) gives $\mathbb{E}\big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} G_h\big] \le \zeta_1 r\sqrt{n}$. Combining the last three displays,

$$\log|U_r| \le \log M\big(\rho, \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}, \|\cdot\|_{L_2(\mathbb{P}_T)}\big) \le \left(\frac{\mathbb{E}\big[\sup_{h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}} G_h\big]}{c_S\,\frac{\rho}{2}}\right)^2 \le \frac{4\zeta_1^2\, r^2\, n}{c_S^2\, \rho^2}.$$

Substituting the fixed value $\rho = \frac{c_0}{4\sqrt{2}c_1}\eta Q_{\mathcal{H}}(\tau) r$ and $\zeta_1 = \frac{c_0 \eta^{3/2} Q_{\mathcal{H}}^2(\tau)}{256 c_1}$, note that c_0 and c_1 cancel identically, which is why Claim 1 holds for every admissible choice of them (Remark H.2) –

$$\frac{4\zeta_1^2 r^2}{c_S^2 \rho^2} = \frac{4}{c_S^2} \cdot \frac{\left(\frac{c_0 \eta^{3/2} Q_H^2(\tau)}{256 c_1}\right)^2}{\left(\frac{c_0}{4\sqrt{2}c_1}\eta Q_H(\tau)\right)^2} = 4\frac{32\,\eta}{256^2\, c_S^2}\, Q_H^2(\tau) = \frac{\eta\, Q_H^2(\tau)}{512\, c_S^2},$$

so that

$$\log|U_r| \le \frac{\eta\, Q_H^2(\tau)}{512\, c_S^2}\, n \le \frac{Q_H^2(\tau)}{512}\, n, \qquad \text{i.e.} \qquad |U_r| \le \exp\left(\frac{Q_H^2(\tau)}{512} n\right),$$

the second inequality holding because $\eta \le c_S^2$ by hypothesis, so that $\eta/(512 c_S^2) \le 1/512$. This is Claim 1.

For every $h \in \mathcal{H}_r := \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$, choose the first point h_u in the fixed ordering of U_r such that $\|h - h_u\|_{L_2(\mathbb{P}_T)} \le \rho$. This choice is independent of the sample. Put

$$a_h(t) = \|h(t)\|_{\mathtt{g}_p}, \qquad d_h(t) = a_h(t) - a_{h_u}(t), \qquad \delta_h(t) = \mathbf{1}\{|d_h(t)| > \tau r/2\},$$

and let $\Delta_r = \{\delta_h : h \in \mathcal{H}_r\}$. By the reverse triangle inequality and Cauchy–Schwarz,

$$\mathbb{E}|d_h(T)| \le \mathbb{E}\|h(T) - h_u(T)\|_{\mathtt{g}_p} \le \|h - h_u\|_{L_2(\mathbb{P}_T)} \le \rho.$$

Thus the vector-valued population net also controls the scalar norm differences. These differences suffice to transfer both coordinate bounds from h_u to h.

Define $\chi(x) = \min\{2|x|/(\tau r), 1\}$ for $x \in \mathbb{R}$. It vanishes at zero and is $2/(\tau r)$-Lipschitz, with $\delta_h(t) \le \chi(d_h(t)) \le 2|d_h(t)|/(\tau r)$. For $F = \sup_{h \in \mathcal{H}_r} n^{-1} \sum_i \delta_h(T_i)$, one-sided symmetrization and scalar contraction give

$$\mathbb{E}F \le 2\mathbb{E} \sup_{h \in \mathcal{H}_r} \frac{1}{n} \sum_i \varepsilon_i \chi(d_h(T_i)) + \frac{2\rho}{\tau r} \tag{38}$$

$$\le \frac{4}{\tau r}\mathbb{E} \sup_{h \in \mathcal{H}_r} \frac{1}{n} \sum_i \varepsilon_i \big(a_h(T_i) - a_{h_u}(T_i)\big) + \frac{2\rho}{\tau r}. \tag{39}$$

The scalar contraction principle is applied to the real function χ and the signed scalar differences $d_h(T_i)$; its one-sided form has constant one (Maurer, 2016, Section 1). The standing separability

assumptions justify these suprema and the symmetrization step. Since $U_r \subseteq \mathcal{H}_r$, subadditivity of the supremum gives

$$\begin{aligned}
&\mathbb{E} \sup_{h\in\mathcal{H}_r} \frac{1}{n}\sum_i \varepsilon_i\big(a_h(T_i) - a_{h_u}(T_i)\big) \\
&\quad\le \mathbb{E} \sup_{h\in\mathcal{H}_r} \frac{1}{n}\sum_i \varepsilon_i a_h(T_i) + \mathbb{E} \sup_{u\in U_r} \frac{1}{n}\sum_i (-\varepsilon_i) a_u(T_i) \\
&\quad\le 2\mathbb{E} \sup_{h\in\mathcal{H}_r} \left|\frac{1}{n}\sum_i \varepsilon_i a_h(T_i)\right| \le 2\zeta_2 r.
\end{aligned}$$

The last inequality follows from Definition 2 and Lemma D.2, since $r > \omega_R(\mathcal{H}, n, \zeta_2)$. Substituting the specified values of ζ_2 and ρ therefore yields

$$\begin{aligned}
\mathbb{E}F &\le \frac{8\zeta_2}{\tau} + \frac{2\rho}{\tau r} \\
&= \frac{c_0\eta Q_{\mathcal{H}}(\tau)}{2\sqrt{2}\,\tau}\left(1 + \frac{1}{c_1}\right) < \frac{c_0\eta Q_{\mathcal{H}}(\tau)}{\sqrt{2}\,\tau},
\end{aligned}$$

because $c_1 > 1$.

The quantity bounded above is precisely the expectation of the empirical process

$$F(T_1,\dots,T_n) = \sup_{\delta_h\in\Delta_r} \frac{1}{n}\sum_{i=1}^n \delta_h(T_i), \qquad \text{so that} \qquad \mathbb{E}F \le \frac{c_0}{\sqrt{2}}\frac{\eta}{\tau} Q_H(\tau).$$

Since every δ_h takes values in $\{0,1\}$, changing a single T_i alters each average $\frac{1}{n}\sum_j \delta_h(T_j)$, and hence their supremum F, by at most $1/n$: F satisfies the bounded-differences property with $a_1 = \cdots = a_n = 1/n$. Applying Lemma H.2 with $s = \frac{c_0}{\sqrt{2}}\frac{\eta}{\tau}Q_H(\tau)$, so that $2ns^2 = \frac{c_0^2\eta^2 Q_H^2(\tau)}{\tau^2} n$, we obtain: with probability at least

$$1 - \exp\left(-\frac{c_0^2\eta^2 Q_H^2(\tau)}{\tau^2}\, n\right), \tag{40}$$

we have, *simultaneously for every* $h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \delta_h(T_i) \le F \;&\le\; \mathbb{E}F + s \le \frac{c_0}{\sqrt{2}}\frac{\eta}{\tau}Q_H(\tau) + \frac{c_0}{\sqrt{2}}\frac{\eta}{\tau}Q_H(\tau) \\
&=\; \sqrt{2}\, c_0 \frac{\eta}{\tau} Q_H(\tau).
\end{aligned}$$

Consequently, on that same event, for every $h \in \mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$,

$$\begin{aligned}
\left|\left\{i : |d_h(T_i)| \le \frac{\tau r}{2}\right\}\right| &= |g\{i, \delta_h(T_i) = 0\}| \\
&= n - |\{i, \delta_h(T_i) = 1\}| \\
&= n - \sum_{i=1}^n \delta_h(T_i) \\
&\ge n - \sqrt{2}\, c_0 \frac{\eta}{\tau} Q_H(\tau) n \\
&= n\big(1 - \sqrt{2}\, c_0 \frac{\eta}{\tau} Q_H(\tau)\big).
\end{aligned}$$

On the other hand, Lemma H.5 applies at each fixed point of U_r with failure probability at most $4\exp\big(-\frac{1}{32}Q_H^2(\tau)n\big)$. Since Claim 1 gives $|U_r| \le \exp\big(\frac{Q_H^2(\tau)}{512}n\big)$, a union bound over U_r shows that, with probability at least

$$1 - 4\exp\left(\frac{Q_H^2(\tau)}{512}n - \frac{1}{32}Q_H^2(\tau)n\right) = 1 - 4\exp\left(-\frac{15}{512}\, Q_H^2(\tau)\, n\right), \tag{41}$$

the conclusion of Lemma H.5 holds *simultaneously at every* $h_u \in U_r$; the exponent uses $1/32 - 1/512 = 15/512$. In particular, for the h_u associated with any given h this guarantees

$$|S_{h_u}| \geq \frac{1}{4} Q_H(\tau)\, n,$$

such that for every $i \in S_{h_u}$, we have

$$\tau \|h_u\|_{L_2(\mathbb{P}_T)} \leq \|h_u(T_i)\|_{\mathbb{g}_p} \leq \frac{2\|h_u\|_{L_2(\mathbb{P}_T)}}{\sqrt{Q_{\mathcal{H}}(\tau)}}.$$

Define

$$K_h = \left\{ i : |d_h(T_i)| \leq \frac{\tau r}{2} \right\} \cap S_{h_u}. \tag{42}$$

Combining (41) with (40) by a union bound, and setting

$$c_{\mathcal{A}}(\eta, \tau) = \min \left\{ \frac{15}{512},\ \frac{c_0^2 \eta^2}{\tau^2} \right\} \ > \ 0, \tag{43}$$

we obtain that, with probability at least

$$1 - 4\exp\left(-\frac{15}{512} Q_H^2(\tau) n \right) - \exp\left(-\frac{c_0^2\eta^2}{\tau^2} Q_H^2(\tau) n \right) \ \geq \ 1 - 5\exp\Big(- c_{\mathcal{A}}(\eta,\tau)\, Q_H^2(\tau)\, n \Big),$$

we have

$$\begin{aligned}
|K_h| &= |S_{h_u}| - \left| S_{h_u} \cap \left\{ i : |d_h(T_i)| > \frac{\tau r}{2} \right\} \right| \\
&\geq |S_{h_u}| - \left| \left\{ i : |d_h(T_i)| > \frac{\tau r}{2} \right\} \right| \\
&\geq \frac{Q_H(\tau) n}{4} - \left| \left\{ i, \delta_h(T_i) = 1 \right\} \right| \\
&\geq \frac{Q_H(\tau) n}{4} - \sum_{i=1}^{n} \delta_h(T_i) \\
&\geq \frac{Q_H(\tau) n}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau} Q_H(\tau) n \\
&\geq \left(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau} \right) Q_H(\tau) n,
\end{aligned}$$

where the above inequalities hold, since for all sets A and B, we have $|A \cap B| = |A| - |A \cap B^c|$ and the fact that $|A \cap B^c| \leq |B^c|$.

Moreover, for every $i \in K_h$,

$$\|h(T_i)\|_{\mathbb{g}_p} \geq \|h_u(T_i)\|_{\mathbb{g}_p} - |d_h(T_i)| \geq \tau r - \frac{\tau r}{2} = \frac{\tau r}{2}. \tag{44}$$

and,

$$\|h(T_i)\|_{\mathbb{g}_p} \leq \|h_u(T_i)\|_{\mathbb{g}_p} + |d_h(T_i)| \leq \frac{2r}{\sqrt{Q_{\mathcal{H}}(\tau)}} + \frac{\tau r}{2}.$$

Therefore,

$$\|h(T_i)\|_{\mathbb{g}_p} \leq 2\left(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}} \right) \|h\|_{L_2(\mathbb{P}_T)}. \tag{45}$$

□

Remark H.3 (Why the Gaussian Process is indexed by vector-valued functions). Claim 1 applies Sudakov's inequality to the process $\{G_h\}$ of Definition 1, whose canonical metric is the vector metric $\|h_1 - h_2\|_{L_2(\mathbb{P}_T)}$. It cannot be replaced by the scalar class $\{\|h(\cdot)\|_{\mathbb{g}_p} : h \in \mathcal{H}\}$: for an orthonormal pair $e_1, e_2 \in \mathcal{T}_p\mathcal{M}$ and $B \geq 1$, a class containing the constant fields $h_1 \equiv e_1$, $h_2 \equiv e_2$ has $\|h_1 - h_2\|_{L_2(\mathbb{P}_T)} = \sqrt{2}$ while $\|h_1(\cdot)\|_{\mathbb{g}_p} \equiv \|h_2(\cdot)\|_{\mathbb{g}_p} \equiv 1$, so no bi-Lipschitz comparison between the two metrics holds on $\mathcal{H} \cap r\mathcal{S}_{L_2(\mathbb{P}_T)}$.

Lemma H.8. *Fix $r > \omega_Q(\zeta_1, \zeta_2)$. On the event $\mathcal{A}_{\mathrm{sb}}(r)$ of Lemma H.7 at radius r the following holds. For every $h \in \mathcal{H}_{v^\star} = \mathcal{V} - v^\star$ such that $\|h\|_{L_2(\mathbb{P}_T)} \geq r$, there exists a subset $S_{1,h} \subset \{1, 2, \ldots, n\}$ satisfying $|S_{1,h}| \geq \left(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau}\right) Q_H(\tau) n$, and for every $i \in S_{1,h}$,*

$$\frac{\tau}{2}\|h\|_{L_2(\mathbb{P}_T)} \leq \|h(T_i)\|_{\mathrm{g}_p} \leq 2\left(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}}\right)\|h\|_{L_2(\mathbb{P}_T)}. \tag{46}$$

Proof. Let $h \in \mathcal{H}_{v^\star}$, with $\|h\|_{L_2(\mathbb{P}_T)} = r$, let h_u be as in Lemma H.7. Hence there is a subset $K_h \subset S_{h_u}$, such that $|K_h| \geq \left(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau}\right) Q_H(\tau) n$ and for every $i \in K_h$,

$$\frac{\tau}{2}\|h\|_{L_2(\mathbb{P}_T)} \leq \|h(T_i)\|_{\mathrm{g}_p} \leq 2\left(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}}\right)\|h\|_{L_2(\mathbb{P}_T)}.$$

On the other hand, the set $H_{v^\star}$ is star-shaped around the origin. Hence, for any $h \in \mathcal{H}_{v^\star}$ such that $\|h\|_{L_2(\mathbb{P}_T)} \geq r$, define $\tilde{h} = \frac{r}{\|h\|_{L_2(\mathbb{P}_T)}}\, h$. Then $\|\tilde{h}\|_{L_2(\mathbb{P}_T)} = r$ and $\tilde{h} \in \mathcal{H}_{v^\star}$ (since $r/\|h\|_{L_2(\mathbb{P}_T)} \in [0, 1]$). Therefore, there exists a subset

$$K_{\tilde{h}} \subset S_{\tilde{h}_u} \quad \text{with} \quad |K_{\tilde{h}}| \geq \left(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau}\right) Q_H(\tau) n,$$

such that for every $i \in K_{\tilde{h}}$,

$$\frac{\tau}{2}\|\tilde{h}\|_{L_2(\mathbb{P}_T)} \leq \|\tilde{h}(T_i)\|_{\mathrm{g}_p} \leq 2\left(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}}\right)\|\tilde{h}\|_{L_2(\mathbb{P}_T)}.$$

Consequently, defining $S_{1,h} = K_{\tilde{h}}$, with

$$|S_{1,h}| \geq \left(\frac{1}{4} - \sqrt{2}\, c_0 \frac{\eta}{\tau}\right) Q_H(\tau) n,$$

we obtain that for every $i \in S_{1,h}$,

$$\frac{\tau}{2}\|h\|_{L_2(\mathbb{P}_T)} \leq \|h(T_i)\|_{\mathrm{g}_p} \leq 2\left(\frac{\tau}{4} + \frac{1}{\sqrt{Q_{\mathcal{H}}(\tau)}}\right)\|h\|_{L_2(\mathbb{P}_T)}.$$

□

The following probability-symmetrization inequality is proved here in the form needed below; see also Giné & Zinn (1984).

Theorem H.9 (Probability symmetrization)**.** *Let $\mathscr{H}$ be a countable class of measurable real functions with finite means and $\beta^2 = \sup_{h \in \mathscr{H}} \mathrm{Var}(h(X)) < \infty$. A justified separable extension may replace countability. Let $X_1, \ldots, X_n$ be i.i.d., $n \geq 1$, and let the Rademacher signs ε_i be independent of these observations. For $z > 0$ with $z \geq 2\sqrt{2}\beta\sqrt{n}$,*

$$\mathbb{P}\left\{ \sup_{h \in \mathscr{H}} \left| \sum_i (h(X_i) - \mathbb{E}h(X)) \right| > z \right\} \leq 4\mathbb{P}\left\{ \sup_{h \in \mathscr{H}} \left| \sum_i \varepsilon_i h(X_i) \right| > z/4 \right\}.$$

Proof. Write $S_h = \sum_i (h(X_i) - \mathbb{E}h(X))$. On $\sup_h |S_h| > z$, choose the first violating function in a fixed enumeration. For an independent ghost sample and $S'_h = \sum_i (h(X'_i) - \mathbb{E}h(X))$, conditional Chebyshev gives

$$\mathbb{P}'\{|S'_h| > z/2\} \leq 4n\beta^2/z^2 \leq 1/2.$$

Thus $\mathbb{P}\{\sup_h |S_h| > z\} \leq 2\mathbb{P}\{\sup_h |S_h - S'_h| > z/2\}$. Independently swapping the exchangeable pairs (X_i, X'_i) inserts Rademacher signs without changing the distribution. By the triangle inequality, an absolute signed difference exceeding $z/2$ requires one of the two signed sample suprema to exceed $z/4$. They have the same distribution, giving the stated factor 4. □

Lemma H.10 (Multiplier deviation under the common-class conditions)**.** *Assume the common measurable representatives, separable indexing, gradient bound $L > 0$, and closure under geodesic contraction toward $\gamma^\star$ required in Lemma D.4. Let $0 < \delta < 1$, $\zeta_3 > 0$, and $\alpha = \alpha_n^\ell(\zeta_3, \delta/4) < \infty$. For $Z = (T, Y)$ define*

$$h_\gamma(t,y) = \Big\langle \mathsf{grad}\ell(\cdot,y)(\gamma^\star(t)), \mathsf{Log}_{\gamma^\star(t)}\gamma(t) \Big\rangle_{\mathsf{g}_{\gamma^\star(t)}}, \quad (\mathbb{P}_n - \mathbb{P})h = \frac{1}{n}\sum_i h(T_i,Y_i) - \mathbb{E}h(T,Y).$$

With probability at least $1-\delta$, every $\gamma \in \widetilde{\Gamma}_{\mathcal{V}}$ with $\mathsf{D}_\mathsf{g}(\gamma,\gamma^\star) \geq \alpha$ satisfies

$$|(\mathbb{P}_n - \mathbb{P})h_\gamma| \leq 4\zeta_3 \mathsf{D}_\mathsf{g}^2(\gamma,\gamma^\star).$$

Proof. For $\mathsf{D}_\mathsf{g}(\gamma,\gamma^\star) \leq \alpha$, Cauchy–Schwarz and the gradient bound give

$$\mathrm{Var}(h_\gamma(Z)) \leq \mathbb{E}h_\gamma(Z)^2 \leq L^2 \mathsf{D}_\mathsf{g}^2(\gamma,\gamma^\star) \leq L^2\alpha^2.$$

The positive-radius attainment result in Lemma D.4 gives $\mathbb{P}\{\phi_n^\ell(\alpha) > \zeta_3\alpha^2\sqrt{n}\} \leq \delta/4$. Use Theorem H.9 on this localized class, with $z = 4\zeta_3\alpha^2 n$. Its threshold holds because $\alpha \geq L/(\sqrt{2}\zeta_3\sqrt{n})$. Therefore, with probability at least $1-\delta$,

$$\sup_{\gamma\in\widetilde{\Gamma}_{\mathcal{V}}:\, \mathsf{D}_\mathsf{g}(\gamma,\gamma^\star)\leq\alpha} |(\mathbb{P}_n - \mathbb{P})h_\gamma| \leq 4\zeta_3\alpha^2. \tag{47}$$

For $D = \mathsf{D}_\mathsf{g}(\gamma,\gamma^\star) \geq \alpha > 0$, set $\theta = \alpha/D$ and $\tilde{\gamma}(t) = \mathsf{Exp}_{\gamma^\star(t)}(\theta \mathsf{Log}_{\gamma^\star(t)}\gamma(t))$. The stipulated common class contains $\tilde{\gamma}$, and constant-speed geodesics give $\mathsf{D}_\mathsf{g}(\tilde{\gamma},\gamma^\star) = \alpha$ and $h_{\tilde{\gamma}} = \theta h_\gamma$, both for population integration and for the common sampled representatives. On (47),

$$|(\mathbb{P}_n - \mathbb{P})h_\gamma| \leq \frac{D}{\alpha}\, 4\zeta_3\alpha^2 = 4\zeta_3\alpha D.$$

Since $\alpha \leq D$, this proves the claimed bound on that event. □

I Proof of the empirical lower bound

We first establish the population first-order condition using the ambient oracle. The empirical bound then follows by controlling the centered first-order term and the quadratic term on the multiplier and small-ball events. The final lemma records the corresponding population margin.

Lemma I.1 (First-order optimality of $v^\star$)**.** *Grant Assumptions 1–7. Then, for every $v \in \mathcal{V}$,*

$$\mathbb{E}\Big[\big\langle \mathsf{grad}\,\ell(\cdot,Y)\big(\gamma^\star(T)\big), \mathsf{Log}_{\gamma^\star(T)}\big(\gamma(T)\big)\big\rangle_{\mathsf{g}_{\gamma^\star(T)}}\Big] \geq 0. \tag{48}$$

Proof. Fix $v \in \mathcal{V}$ and, for $t \in [0,1]$ and $s \in [0,1]$, let

$$\gamma_t(s) = \mathsf{Exp}_{\gamma^\star(s)}\Big(t\,\mathsf{Log}_{\gamma^\star(s)}(\gamma(s))\Big)$$

be the point at parameter t along the unique geodesic from $\gamma^\star(s)$ to $\gamma(s)$; so $\gamma_0 = \gamma^\star$, $\gamma_1 = \gamma$, and $\mathsf{d}_\mathsf{g}(\gamma^\star(s),\gamma_t(s)) = t\,\mathsf{d}_\mathsf{g}(\gamma^\star(s),\gamma(s))$ for every $t \in [0,1]$ (constant-speed reparametrization of a geodesic). Since $\gamma^\star(s), \gamma(s) \in \mathcal{B}(p,B)$ on the common full-measure set and the ball is geodesically convex, $\gamma_t \in \Gamma_p^B$ for all $t \in [0,1]$. The interpolation and logarithm maps are continuous, so $v_t = \mathsf{Log}_p \circ \gamma_t$ is measurable and belongs to $\mathcal{V}_p^B$. No membership in $\mathcal{V}$ is required here. Define $\varphi(t) = \mathsf{R}(\ell_{v_t})$. Since $v^\star$ minimizes risk over the ambient $\mathcal{V}_p^B$, $\varphi(t) \geq \varphi(0)$ for every $t \in [0,1]$.

We claim φ is right-differentiable at $t = 0$ with

$$\varphi'(0^+) = \mathbb{E}\Big[\big\langle \mathsf{grad}\,\ell(\cdot,Y)(\gamma^\star(T)),\ \mathsf{Log}_{\gamma^\star(T)}(\gamma(T))\big\rangle_{\mathsf{g}_{\gamma^\star(T)}}\Big]. \tag{49}$$

Indeed, fix $s \in [0,1]$ and $y \in \mathcal{Y}$, and consider $\psi(t) = \ell(\gamma_t(s), y)$. By construction $t \mapsto \gamma_t(s)$ is (a reparametrization of) the geodesic through $\gamma^\star(s)$ with initial velocity $w = \mathsf{Log}_{\gamma^\star(s)}(\gamma(s)) \in$

$\mathcal{T}_{\gamma^\star(s)}\mathcal{M}$, so by the definition of Riemannian differentiability (Definition D.1, applied with $x = \gamma^\star(s)$, $\gamma(t) = \gamma_t(s)$, $v = w$),

$$\lim_{t\to 0^+} \frac{\psi(t) - \psi(0)}{t} = \big\langle \mathsf{grad}\, \ell(\cdot, y)(\gamma^\star(s)),\, w\big\rangle_{\mathsf{g}_{\gamma^\star(s)}}.$$

Moreover, for every $t \in (0, 1]$, by Assumption 5 (the loss is geodesically L-Lipschitz on $\mathcal{D}_B \supset \mathcal{B}(p, B)$, which contains both $\gamma^\star(s)$ and $\gamma_t(s)$),

$$\left|\frac{\psi(t) - \psi(0)}{t}\right| \le \frac{L\, \mathsf{d}_{\mathsf{g}}(\gamma^\star(s), \gamma_t(s))}{t} = L\, \mathsf{d}_{\mathsf{g}}(\gamma^\star(s), \gamma(s)) \le 2LB,$$

a bound not depending on t and integrable in $(s, y) \sim \mathbb{P}$ (it is a constant). By the dominated convergence theorem, the pointwise limit above may be taken inside the expectation defining φ, yielding exactly (49).

Since $\varphi(t) \ge \varphi(0)$ for all $t \in [0, 1]$ and φ is right-differentiable at 0, the one-sided difference quotient $(\varphi(t) - \varphi(0))/t$ is ≥ 0 for every $t \in (0, 1]$, hence so is its limit $\varphi'(0^+)$. Combined with (49), this is exactly (48). □

Proof of Theorem E.2. Fix $v \in \mathcal{V}$ with $u := \|v - v^\star\|_{L_2(\mathbb{P}_T)} > r^\star$, and put $\gamma = \mathsf{Exp}_p \circ v$. Write

$$a_v(T, Y) = \Big\langle \mathsf{grad}\ell(\cdot, Y)(\gamma^\star(T)), \mathsf{Log}_{\gamma^\star(T)}\gamma(T)\Big\rangle_{\mathsf{g}_{\gamma^\star(T)}}.$$

The gradient bound and bounded prediction ball give $|a_v| \le 2LB$. Lemma I.1 therefore supplies the well-defined inequality

$$\mathbb{E}a_v \ge 0. \tag{50}$$

By geodesic strong convexity, averaging over the observations yields

$$\mathsf{R}_n\mathcal{L}_v \ge (\mathbb{P}_n - \mathbb{P})a_v + \frac{\mu}{2n}\sum_{i=1}^n \mathsf{d}^2_{\mathsf{g}}(\gamma(T_i), \gamma^\star(T_i)), \tag{51}$$

where $\mathbb{P}_n a = n^{-1}\sum_i a(T_i, Y_i)$.

The lower distance comparison implies $\mathsf{D}_{\mathsf{g}}(\gamma, \gamma^\star) \ge u > r^\star \ge \alpha^\ell_n(\zeta_3, \delta/4)$. Thus Lemma H.10 applies on $\mathcal{A}_{\mathrm{mult}}$, simultaneously over these v, and the upper comparison of norm equivalence (Proposition 1)

$$(\mathbb{P}_n - \mathbb{P})a_v \ge -4\zeta_3\mathsf{D}^2_{\mathsf{g}}(\gamma, \gamma^\star) \ge -4\zeta_3\kappa_B^2 u^2.$$

For the quadratic term, the pointwise lower comparison first gives $\mathsf{d}^2_{\mathsf{g}}(\gamma(T_i), \gamma^\star(T_i)) \ge \|v(T_i) - v^\star(T_i)\|^2_{\mathsf{g}_p}$. Since $u > r^\star \ge \bar r > \omega_Q$, Lemma H.8 applies on $\mathcal{A}_{\mathrm{sb}}(\bar r)$. It selects a set $S_{1,v}$ with

$$|S_{1,v}| \ge \left(\frac{1}{4} - \sqrt{2}c_0\frac{\eta}{\tau}\right) Q_H(\tau)n, \qquad \|v(T_i) - v^\star(T_i)\|_{\mathsf{g}_p} \ge \frac{\tau}{2}u \quad (i \in S_{1,v}).$$

Keeping these nonnegative summands gives

$$\frac{\mu}{2n}\sum_i \|v(T_i) - v^\star(T_i)\|^2_{\mathsf{g}_p} \ge \frac{|S_{1,v}|}{n}\frac{\mu\tau^2}{8}u^2 \ge \varkappa_0 u^2.$$

Substitution in (51) proves $\mathsf{R}_n\mathcal{L}_v \ge (\varkappa_0 - 4\zeta_3\kappa_B^2)u^2$ on $\mathcal{A} = \mathcal{A}_{\mathrm{sb}}(\bar r) \cap \mathcal{A}_{\mathrm{mult}}$. Both component events are uniform over the required classes. Their failure probabilities are at most $5e^{-c_{\mathcal{A}}(\eta,\tau)Q^2_H(\tau)n}$ and δ, respectively; the union bound gives the stated confidence level. □

Lemma I.2 (Population curvature / margin bound)**.** *Let Assumptions 1–7 hold. Then, for every $v \in \mathcal{V}$,*

$$\mathsf{R}(\mathcal{L}_v) = \mathsf{R}(v) - \mathsf{R}(v^\star) \;\ge\; \frac{\mu}{2}\,\mathsf{D}^2_{\mathsf{g}}(\gamma, \gamma^\star) \;\ge\; \frac{\mu}{2}\,\|v - v^\star\|^2_{L_2(\mathbb{P}_T)}.$$

Proof. By Assumption 4 applied at $q_1 = \gamma^\star(t)$, $q_2 = \gamma(t)$ and Fubini, taking expectations over $(T, Y) \sim \mathbb{P}$,

$$\mathsf{R}(v) - \mathsf{R}(v^\star) \;\ge\; \mathbb{E}\Big[\big\langle \mathsf{grad}\,\ell(\cdot, Y)(\gamma^\star(T)), \mathsf{Log}_{\gamma^\star(T)}(\gamma(T))\big\rangle_{\mathsf{g}_{\gamma^\star(T)}}\Big] + \frac{\mu}{2}\,\mathsf{D}^2_{\mathsf{g}}(\gamma, \gamma^\star).$$

The first term on the right is ≥ 0 by Lemma I.1, giving the first inequality. The second follows from Lemma E.1. □

J Experimental Protocols and Full Results

This appendix specifies the random-truth simulations and full-cohort EEG evaluation used in Section 4. Complementary structural comparisons and failure diagnostics appear in Appendix K. The summaries and numerical figures are rebuilt from the supplied per-fit records.

J.1 Geometry, representations and estimators

The Poincaré ball $\mathbb{H}^d$ has curvature -1, reference point at its centre, and $\kappa_B = \sinh(B)/B$. For $\mathrm{SPD}(6)$ we use the affine-invariant metric and reference point I, except for the re-centred EEG fits below. At I, the exponential and logarithm are the matrix exponential and logarithm, and $\mathsf{d}_{\mathsf{g}}(I, Q) = \|\log Q\|_F$. All primary geodesic fits minimize the half-squared distance loss. It is geodesically 1-strongly convex and Lipschitz on a bounded prediction/response domain with $L = \mathrm{diam}(\mathcal{D}_B)$.

The coordinate-wise Fourier basis is $\{1, \sqrt{2}\cos(2\pi kt), \sqrt{2}\sin(2\pi kt)\}_{k \le n_{\mathrm{freq}}}$, orthonormal for uniform time. It gives $K = d(2n_{\mathrm{freq}} + 1)$ coefficients on $\mathbb{H}^d$ and $K = 21(2n_{\mathrm{freq}} + 1)$ on $\mathrm{SPD}(6)$. We compare geodesic ERM and ℓ_1-RERM with least squares and LASSO using $\frac{1}{2}\|v(T_i) - \mathsf{Log}_p(Y_i)\|^2_{\mathsf{g}_p}$ in place of the intrinsic loss. The latter use the same representation but a different loss; at $p = I$ on SPD matrices they are log-Euclidean fits. Where available, a geodesic fit on the true support provides an infeasible reference. Optimization and feasibility conventions are specified in Appendix J.4.

The five estimators in coefficient form. Let $(e_1, \dots, e_m)$ be an orthonormal frame of $\mathcal{T}_p\mathcal{M}$, with $m = d$ on $\mathbb{H}^d$ and $m = 21$ on $\mathrm{SPD}(6)$, and let $\phi_1, \dots, \phi_{K_t}$, $K_t = 2n_{\mathrm{freq}} + 1$, be the scalar dictionary. A coefficient matrix $\beta \in \mathbb{R}^{K_t \times m}$ defines the tangent field and the curve

$$v_\beta(t) = \sum_{k=1}^{K_t} \phi_k(t)\, \beta_k \in \mathcal{T}_p\mathcal{M}, \qquad \beta_k = \sum_{j=1}^{m} \beta_{kj} e_j, \qquad \gamma_\beta = \mathsf{Exp}_p \circ v_\beta,$$

so that $K = mK_t$ scalar coefficients are fitted, $\|\beta\|_1 = \sum_{k,j} |\beta_{kj}|$ is the penalty $\Omega(v_\beta)$ of Section 2, and $\|\beta\|_F$ is its Frobenius norm. The empirical risks of the geodesic and of the tangent loss are

$$\widehat{\mathsf{R}}^{\mathrm{geo}}_n(\beta) = \frac{1}{2n}\sum_{i=1}^{n} \mathsf{d}^2_{\mathsf{g}}\big(\mathsf{Exp}_p(v_\beta(T_i)), Y_i\big), \qquad \widehat{\mathsf{R}}^{\mathrm{tan}}_n(\beta) = \frac{1}{2n}\sum_{i=1}^{n} \big\|v_\beta(T_i) - \mathsf{Log}_p(Y_i)\big\|^2_{\mathsf{g}_p},$$

the second being the log-Euclidean least-squares criterion at $p = I$ on $\mathrm{SPD}(6)$. The identity $\sum_{k \le K_t} \phi_k(t)^2 = K_t$ and Cauchy–Schwarz give $\sup_t \|v_\beta(t)\|_{\mathsf{g}_p} \le \sqrt{K_t}\,\|\beta\|_F$, so the implemented feasible set

$$\mathcal{B} = \big\{\beta \in \mathbb{R}^{K_t \times m} :\ \|\beta\|_F \le B/\sqrt{K_t}\big\}$$

is a sufficient, not an equivalent, condition for $v_\beta \in \mathcal{V}_p^B$. With $S^\star$ the support of the generating coefficient $\beta^\star$, the five compared estimators are

$$\begin{aligned}
\text{geodesic ERM:} \quad & \hat\beta^{\mathrm{ERM}} \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \widehat{\mathsf{R}}^{\mathrm{geo}}_n(\beta), \\
\text{geodesic RERM:} \quad & \hat\beta^{\mathrm{RERM}}_\lambda \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \big\{\widehat{\mathsf{R}}^{\mathrm{geo}}_n(\beta) + \lambda\|\beta\|_1\big\}, \\
\text{tangent ERM (least squares):} \quad & \hat\beta^{\mathrm{LS}} \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \widehat{\mathsf{R}}^{\mathrm{tan}}_n(\beta), \\
\text{tangent LASSO:} \quad & \hat\beta^{\mathrm{LASSO}}_\lambda \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \big\{\widehat{\mathsf{R}}^{\mathrm{tan}}_n(\beta) + \lambda\|\beta\|_1\big\}, \\
\text{support oracle:} \quad & \hat\beta^{\mathrm{orac}} \in \operatorname*{argmin}_{\beta \in \mathcal{B},\ \mathrm{supp}(\beta) \subseteq S^\star} \widehat{\mathsf{R}}^{\mathrm{geo}}_n(\beta),
\end{aligned}$$

and each of them predicts through $\hat\gamma = \mathsf{Exp}_p \circ v_{\hat\beta}$. The oracle uses the true support and is therefore infeasible; it is a reference fit with known support, not a guaranteed lower bound on prediction error. The penalized estimators share one selection rule: with the folds $I_1, \dots, I_4$ of a random partition of

Table 5: Simulation with random truths, sample-size sweep (Subsection J.2): mean per-replication RMS geodesic error ± sample SD over 100 truth/data replications per row ($s = 4$, $\sigma = 0.25$, $a = 0.5$; $K = 22$ on $\mathbb{H}^2$, $K = 105$ on $\mathrm{SPD}(6)$; λ by four-fold cross-validation). SDs measure the spread of replication errors, not uncertainty in their mean. Bold and italics mark the lowest and second-lowest mean errors in each row among the four feasible methods. The support oracle uses the true support and is excluded from this ranking.

n	geod. ERM	geod. RERM	tang. ERM	tang. LASSO	oracle
			$\mathbb{H}^2$ ($K = 22$)		
80	0.0972 ± 0.0143	*0.0696 ± 0.0171*	0.0984 ± 0.0144	**0.0669 ± 0.0172**	0.0376 ± 0.0125
160	0.0667 ± 0.0107	*0.0481 ± 0.0122*	0.0676 ± 0.0108	**0.0458 ± 0.0122**	0.0261 ± 0.0089
320	0.0475 ± 0.0068	*0.0343 ± 0.0069*	0.0483 ± 0.0069	**0.0316 ± 0.0068**	0.0183 ± 0.0068
640	0.0334 ± 0.0056	*0.0252 ± 0.0049*	0.0340 ± 0.0059	**0.0225 ± 0.0052**	0.0134 ± 0.0045
1280	0.0238 ± 0.0037	*0.0177 ± 0.0039*	0.0244 ± 0.0039	**0.0149 ± 0.0040**	0.0100 ± 0.0032
			$\mathrm{SPD}(6)$ ($K = 105$)		
80	0.0640 ± 0.0065	*0.0275 ± 0.0064*	0.0641 ± 0.0065	**0.0270 ± 0.0064**	0.0116 ± 0.0043
160	0.0448 ± 0.0035	*0.0198 ± 0.0050*	0.0448 ± 0.0035	**0.0194 ± 0.0049**	0.0087 ± 0.0030
320	0.0311 ± 0.0020	*0.0138 ± 0.0031*	0.0312 ± 0.0020	**0.0136 ± 0.0031**	0.0059 ± 0.0021

$\{1, \ldots, n\}$ and $\hat{\beta}_\lambda^{(-f)}$ the fit on the complement of I_f,

$$\hat{\lambda} \in \operatorname*{argmin}_{\lambda \in \Lambda_n} \sum_{f=1}^{4} \sum_{i \in I_f} \mathsf{d}_{\mathsf{g}}^2\big(\mathsf{Exp}_p(v_{\hat{\beta}_\lambda^{(-f)}}(T_i)), Y_i\big), \qquad \Lambda_n = \sigma\sqrt{\frac{\log(2K)}{n}} \cdot \Big\{0, \tfrac{1}{16}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{2}, 1, 2, 4\Big\},$$

the held-out loss being geodesic for both, so that the two losses are compared under a single criterion.

J.2 Random-truth simulations

Each replication selects s dictionary coordinates and random coefficient signs, then rescales the tangent field to L_2 norm a. Observation times are i.i.d. uniform on $[0, 1]$. At each point of the generating curve, noise has an isotropic tangent direction and a radius uniform on $[0, \sqrt{3}\sigma]$, with root-mean-square magnitude σ. The supplied protocol uses $B = 6$ and reports that the generating curves and responses remain inside the ball. The base setting is $\mathbb{H}^2$, $n = 160$, $K = 22$, $s = 4$, $a = 0.5$, $\sigma = 0.25$.

There are 100 independent truth/data replications in each of 22 settings, giving 2 200 datasets, 11 000 final fits and 140 800 cross-validation fits. Both regularized estimators select λ by four-fold cross-validation of held-out *geodesic* loss over the supplied grid $\sigma\sqrt{\log(2K)/n} \times \{0, 1/16, \ldots, 4\}$; no selection falls on its upper edge. The grid is given above. The implementation uses master seed 2026091801; for zero-based replication r, the fold permutation is generated from NumPy SeedSequence inputs $[2026091801, r, 510]$. Each fitted curve is scored by integrated squared geodesic error on a separate 2 048-point grid. For replication r, let e_r be the square root of this integrated squared error. The main figure and Table 5 report $\bar{e}$ and $s_e = \{\sum_{r=1}^{100}(e_r - \bar{e})^2/99\}^{1/2}$. Thus SD describes between-replication variation. This mean of root errors differs from the former root of mean squared error; both are exported by the reporting code. MSE ratios and fitted exponents continue to use the mean of e_r^2. Active coefficients satisfy $|\beta_k| > 10^{-3}$. All final fits reportedly reached the stopping tolerance with the ball constraint inactive; neither observation certifies a global optimum.

The sample-size sweep uses $n = 80, 160, 320, 640, 1280$ on $\mathbb{H}^2$ with $K = 22$ and $n = 80, 160, 320$ on $\mathrm{SPD}(6)$ with $K = 105$. The dictionary sweep embeds the same $s = 4$ generating field in $K = 22, 42, 82$ coordinates at $n = 160$; the sparsity sweep uses $s = 2, 4, 8, 16$ with $K = 22$ and fixed signal energy. Figure 2 displays these comparisons. At the base setting, the mean selected support size is 9.6 and the reported false-discovery fraction is 0.50 for a true support of 4. Tangent LASSO has lower mean error than geodesic RERM in all 22 settings; the mean-squared-error ratio RERM/LASSO increases from 1.07 at $n = 80$ to 1.38 at $n = 1280$, and is 1.63 on $\mathbb{H}^8$. These comparisons do not establish exact support recovery or an accuracy advantage from using the intrinsic loss.

J.3 FULL-COHORT EEG COVARIANCE INTERPOLATION

Data and preprocessing. We use all 109 subjects and 14 runs per subject in the PhysioNet EEG Motor Movement/Imagery cohort (Goldberger et al., 2000; Schalk et al., 2004): two one-minute baseline runs and twelve two-minute execution/imagery runs, with no recording excluded. The six-channel montage is $\{\mathrm{FC3}, \mathrm{FCz}, \mathrm{FC4}, \mathrm{C3}, \mathrm{Cz}, \mathrm{C4}\}$. Windowing is specified in seconds. The supplied per-recording quality records report 1 490 recordings at 160 Hz and 36 at 128 Hz. One-second windows every half second yield 347 922 covariance observations from 1 526 recordings. Channels are standardized using scales estimated on the preprocessing subjects; each sample covariance is shrunk 10% toward a scaled identity. Time is normalized within each recording. The resulting responses lie in $\mathrm{SPD}(6)$.

Split, fitting windows and score. A fixed subject split assigns 65 subjects to preprocessing, 22 to tuning and 22 to evaluation. Each recording is fitted on its own windows outside its central 20% block. Every fitting window whose raw samples intersect the held-out block or a one-second buffer on either side is removed. The supplied records report this disjointness check for every recording. Between 52 and 282 fitting windows remain per recording (median 190). The primary score averages held-out squared geodesic error within a recording, over the subject's 14 runs, and then over the 22 test subjects. We report the sample SD of these 22 subject scores, using denominator 21. Paired comparisons use the SD of the 22 within-subject score differences. The 308 test recordings are not treated as independent replicates.

Re-centred models and tuning. For each recording, $p = \bar{Y}$ is the Fréchet mean of its fitting windows only. The congruence $Y \mapsto \bar{Y}^{-1/2} Y \bar{Y}^{-1/2}$ preserves affine-invariant distances and permits fitting at I in the transformed coordinates. The Fourier model has two frequencies per tangent coordinate, hence 105 coefficients, of which the 21 intercepts are unpenalized. The second model uses five cubic B-splines per coordinate (one interior knot, partition of unity), with all 105 coefficients penalized. At zero coefficients both models predict the constant $\bar{Y}$.

The supplied Fourier implementation retains the coefficient constraint $\|\beta\|_F \le B/\sqrt{5}$, which is sufficient for its continuous tangent bound; $B = 40$ for this application. The reference point is estimated from fitting data and is therefore outside the fixed-reference theory. The spline basis is not the orthonormal Fourier model of Assumption F.1. Validation selects $\lambda = 0.64$ at the upper grid endpoint for Fourier and $\lambda = 0.16$ for splines. The Fourier validation error decreases to the constant predictor's value 2.035; the spline curve is flat from 0.08 onward, with minimum 2.034. Endpoint selection is reported rather than interpreted as an interior optimum.

Intrinsic comparisons. The constant fitting-window Fréchet mean, global Fréchet regression and local-linear Fréchet regression (Petersen & Müller, 2019) use the same fitting windows and score. The latter two use weighted Fréchet means; the global weights are $s_{in}(t) = 1 + (T_i - \bar{T})(t - \bar{T})/\hat{\sigma}^2$ and may be negative. The local bandwidth is selected on validation subjects from $\{0.15, 0.25, 0.4, 0.6, 1\}$, yielding $h = 0.6$. The reported validation scores on this grid are 5.86, 2.11, 2.05, 2.03, 2.03. These are the intrinsic baselines evaluated under the retained EEG protocol; no full-cohort TV or geodesic-regression result is claimed. The TV comparison in Appendix K.1 uses simulated data.

Table 1 contains the primary numerical results. Pairwise differences from spline RERM are shown in Figure 3. These descriptive SD bars are not confidence intervals or tests of equivalence. The Fourier fit has no non-constant coefficient in 306 of 308 test recordings. Spline fits have all coefficients zero in 204 recordings, and retain 1.8 on average. These counts explain why errors approach those of the constant predictor, without showing that the data contain no useful temporal signal.

Reference-point sensitivity. The base-point-I results are retained in Table 6 because they document a less successful specification and the tangent-loss comparison on the same full cohort. All Fourier coefficients, including intercepts, are penalized; validation selects $\lambda = 0.04$ for both losses. The fitting-window means lie at distance 2.7–13.7 from I (median 5.9), and the constant predictor is more accurate than this RERM on 21 of 22 test subjects. Tangent LASSO has a lower reported mean error than geodesic RERM, 2.763 versus 2.781. The change to the main specification combines re-centering and an unpenalized Fourier intercept, so their individual effects are not identified. Active

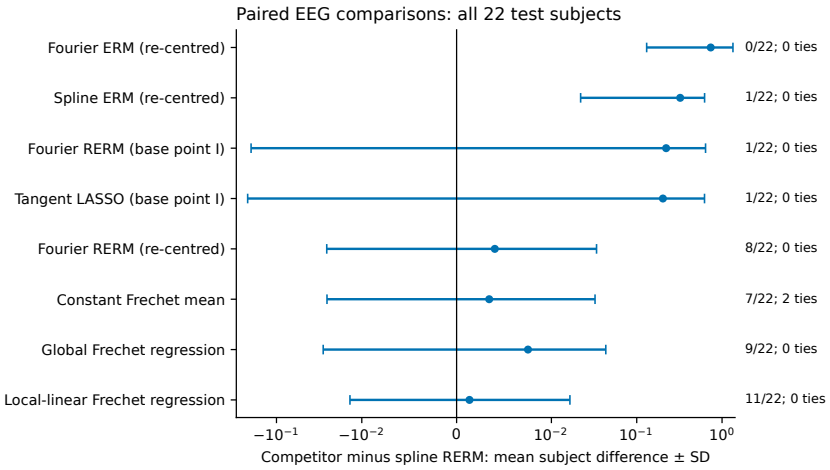


Figure 3: Paired EEG comparisons under the subject-disjoint, purged-block protocol. Points show the mean of competitor-minus-spline-RERM squared geodesic error over 22 subjects; bars show ± one sample SD of those paired differences. Negative values favor the competitor. Fractions count subjects favoring the competitor out of all 22; exact ties are listed separately. The constant comparison has two exact ties. The symmetric-log axis retains the full SD bars. "Base point I" also uses a penalized intercept.

Table 6: Reference-point sensitivity on the full EEG cohort (base point I, intercept penalized): mean held-out squared geodesic error over the 22 test subjects (14 runs each), reported as mean ± sample SD across subjects, and mean number of active tangent coordinates out of $K = 105$ ($|\beta_j| > 10^{-3} \max_k |\beta_k|$). $\lambda^\star = 0.04$ for both regularized estimators, chosen on the 22 validation subjects. Bold and italics mark the lowest and second-lowest mean errors, respectively.

Estimator	mean held-out error ± SD	active coordinates
geodesic ERM	3.297 ± 1.305	97.4
geodesic RERM	*2.781 ± 1.107*	38.9
tangent ERM	3.307 ± 1.307	97.4
tangent LASSO	**2.763 ± 1.104**	36.0

counts for this table use $|\beta_j| > 10^{-3} \max_k |\beta_k|$; re-centred counts use nonzero coefficients. These conventions should not be conflated.

Scope. The task is interpolation within a recording whose surrounding windows are available, not prediction of an entirely unseen recording without fitting data. Purging prevents raw-sample overlap across the evaluation boundary but does not make windows independent. Estimated reference points, deterministic time indices, finite optimization and the large curvature factor at $B = 40$ prevent an informative certificate from the distance theorem. The development chronology is not documented well enough to establish that the re-centred Fourier and spline specifications were fixed before test results were inspected. We therefore report these comparisons as exploratory; the subject split and purging do not by themselves establish independence of all model choices from test outcomes.

J.4 Algorithm and implementation conventions

Algorithm 1 gives the fixed-step, grid-rescaling template used in the complementary studies. The primary simulation and EEG solvers instead use backtracking, monotone restart and a proximal-gradient stopping residual. They initialize by tangent least squares and combine soft thresholding with the coefficient-ball constraint $\|\beta\|_F \le B/\sqrt{q}$, where q is the number of scalar basis functions. For the Fourier basis, $\sum_{j=1}^{q} \phi_j(t)^2 = q$ certifies the continuous prediction bound. These differences matter when reproducing the fits; iteration budgets and tolerances are specified in the supplied protocols. For an unpenalized coordinate, use threshold zero in the proximal step; this applies to the 21 Fourier

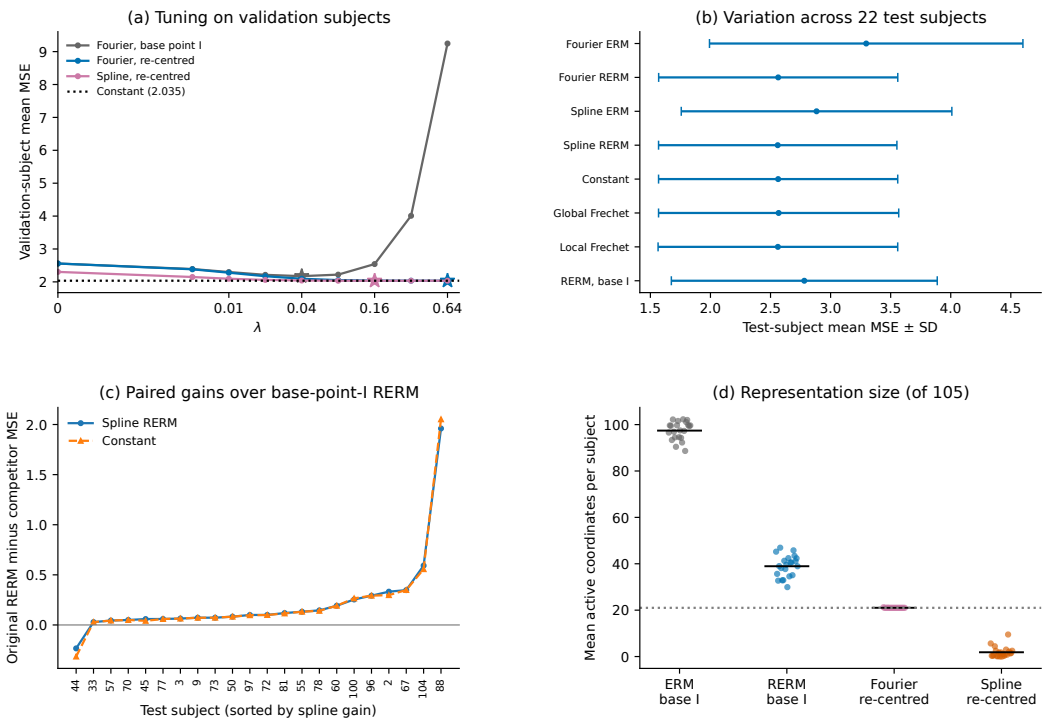


Figure 4: EEG tuning and representation size on the full cohort. **(a)** Validation error against λ for base point I with penalized intercept and the two re-centred models; stars mark selected values and the dotted line is the constant predictor's validation error. **(b)** Mean test error $\pm$ sample SD across 22 subjects. **(c)** Subject-level error reductions for re-centred spline RERM and the constant predictor relative to the original RERM. **(d)** Active coefficients per subject (points), with horizontal marks at their means; the dotted line marks the 21 Fourier intercepts. Tuning uses 22 validation subjects and evaluation uses 22 disjoint test subjects, each contributing 14 recordings.

intercepts of re-centred EEG fits. Radial grid rescaling and the sufficient EEG coefficient constraint described above are distinct feasibility rules.

The coefficient objective need not be convex. A small iterate change, monotone objective history or agreement across initializations does not certify a global optimization gap. Grid feasibility alone does not certify the continuous bound. The supplied analytic gradient checks report relative errors of 4×10^{-10} on the disk and at most 10^{-9} for the extended hyperbolic implementations; these are numerical diagnostics.

The supplied implementation has cost $O(nK)$ per iteration on the disk and $O(n(d^3 + K_{\tan} d^3) + nK)$ on $\mathrm{SPD}(d)$, with $K_{\tan} = d(d+1)/2$. The matrix terms account for eigendecompositions and the Fréchet derivative of the matrix exponential. This expression concerns one iteration, not convergence to an exact minimizer.

K Complementary Experiments and Practical Limitations

The retained studies compare structural priors, document tuning and optimization limitations, and report solver cost. Their simulation generators differ from the primary random-truth suite, so their errors are not pooled with it. Numerical summaries and distribution plots are reconstructed from the supplied replication records; the original hardware behind the timing records remains unverified.

Algorithm 1 Tangent-Space RERM by Accelerated Proximal Gradient (FISTA)

Input: sample $(T_i, Y_i)_{i\le n}$; base point p; dictionary $(\phi_k)_{k\le K}$; radius $B > 0$; penalty $\lambda \ge 0$; step size $\eta > 0$; budget N; feasibility grid $G \subset [0,1]$; tolerance $\varepsilon \ge 0$

Notation. $v_\beta = \sum_{k\le K} \beta_k \phi_k$; $\|\beta\|_G = \max_{s\in G} \|v_\beta(s)\|_{\mathsf{g}_p}$;

$f(\beta) = \frac{1}{2n} \sum_{i\le n} \mathsf{d}_{\mathsf{g}}^2\big(\mathsf{Exp}_p(v_\beta(T_i)), Y_i\big)$; $(\mathcal{S}_\tau x)_k = \operatorname{sgn}(x_k)\,(|x_k| - \tau)_+$.

$\beta^{(0)} \leftarrow 0$; $z^{(1)} \leftarrow \beta^{(0)}$; $\theta_1 \leftarrow 1$ ▷ $\|\beta^{(0)}\|_G = 0 \le B$

for $j = 1, \dots, N$ **do**

Gradient at the extrapolated point: for $k = 1, \dots, K$,

$$\big[\nabla f(z^{(j)})\big]_k = -\frac{1}{n} \sum_{i\le n} \Big\langle \mathsf{Log}_{\mathsf{Exp}_p(v_{z^{(j)}}(T_i))}(Y_i),\ \mathrm{d}(\mathsf{Exp}_p)_{v_{z^{(j)}}(T_i)}\big[\phi_k(T_i)\big] \Big\rangle_{\mathsf{g}}$$

Proximal step: $b \leftarrow \mathcal{S}_{\eta\lambda}\big(z^{(j)} - \eta\, \nabla f(z^{(j)})\big)$

if $\|b\|_G > B$ **then**

$b \leftarrow \big(B/\|b\|_G\big)\, b$ ▷ radial rescaling; not an exact projection

end if

$\beta^{(j)} \leftarrow b$

Momentum: $\theta_{j+1} \leftarrow \frac{1}{2}\big(1 + \sqrt{1 + 4\theta_j^2}\big)$; $z^{(j+1)} \leftarrow \beta^{(j)} + \dfrac{\theta_j - 1}{\theta_{j+1}}\big(\beta^{(j)} - \beta^{(j-1)}\big)$

if $\|\beta^{(j)} - \beta^{(j-1)}\|_2 \le \varepsilon$ **then**

break

end if

end for

Output: $\hat\beta \leftarrow \beta^{(j)}$; $\hat v \leftarrow \sum_{k\le K} \hat\beta_k \phi_k$; $\hat\gamma \leftarrow \mathsf{Exp}_p \circ \hat v$

K.1 Sparse Fourier and piecewise-constant structure

This comparison distinguishes coefficient sparsity from temporal total variation. On the Poincaré disk, $B = 2$ and $K = 22$ (five Fourier frequencies in each of two tangent coordinates). Truth A has $s = 4$ nonzero Fourier coefficients and maximum tangent norm $0.85B$. Truth B is piecewise constant with four fixed segments, with levels at radii in $[0.4, 0.85B]$; a finite Fourier dictionary cannot represent it exactly. Observation times are uniform, and noise is a random tangent direction at the generating curve with half-normal Riemannian magnitude $\sigma|Z|$, $Z \sim \mathcal{N}(0,1)$. Unbounded noise means this diagnostic does not satisfy the bounded-response assumption.

We compare geodesic ERM, Fourier RERM and TV-regularized Fréchet regression (Lin & Müller, 2021). The TV fit minimizes $n^{-1}\sum_i \mathsf{d}_{\mathsf{g}}^2(g_i, Y_i) + \lambda \sum_{i<n} \mathsf{d}_{\mathsf{g}}(g_i, g_{i+1})$ over sorted design points, using cyclic proximal steps and a right-continuous step extension. The supplied self-tests recover observations at zero penalty and approach the Fréchet mean at large penalty. For each truth, there are 25 replications at each of $(n, \sigma) = (60, 0.25)$ and $(100, 0.45)$. Both penalties are selected by four-fold cross-validation of held-out geodesic loss (10 candidate levels for RERM, 9 for TV; reported selections are interior). The score is root mean squared geodesic distance to the generating curve on 500 time points. The truth is used for evaluation, not tuning.

Figure 5 combines fitted-curve examples with the complete error distributions. The examples use replication 0 of the first declared setting, $n = 60$ and $\sigma = 0.25$, for each truth (seeds 0 and 100 000). These indices were fixed before refitting, without selecting for error or visual appearance; their errors and selected penalties reproduce the archived records to numerical precision. The smooth target illustrates a Fourier representation, whereas the four separated target levels expose the mismatch of a finite Fourier model to a discontinuous curve. The distributions show the structural reversal across repeated draws. For truth A, median errors are 0.094 (RERM), 0.166 (ERM) and 0.310 (TV), with RERM better than TV in all 50 replications. For truth B, TV is better in 41 of 50 replications, with median error 0.314 against 0.378 for RERM and 0.381 for ERM. RERM retains 20–21 of 22 coefficients on truth B. Thus a representation penalty can be useful when it matches the target, while a different structural prior can be preferable. This does not verify the distance theorem or bound approximation error for truth B.

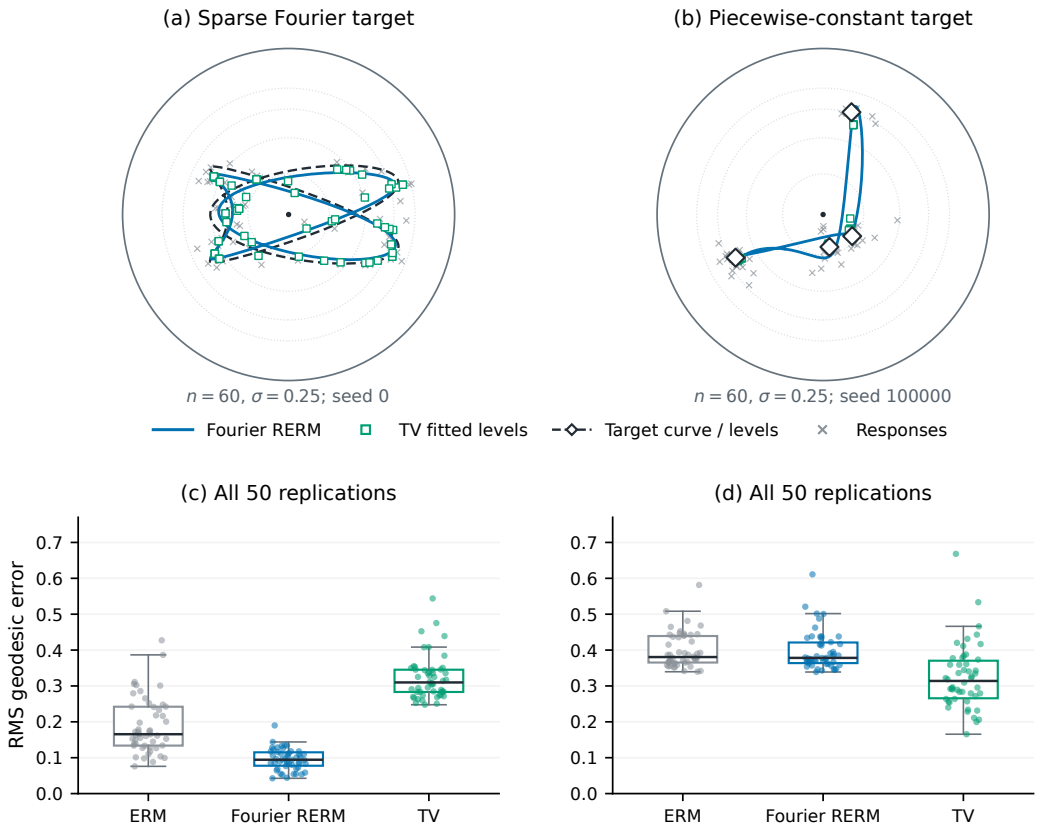


Figure 5: Fourier sparsity versus TV regularization on the Poincaré disk. **(a,b)** Preselected fitted examples for the sparse Fourier and four-segment piecewise-constant targets, respectively ($n = 60$, $\sigma = 0.25$; seeds 0 and 100 000). Blue curves are Fourier RERM fits; green squares are TV fitted levels. The dashed black curve or black diamonds show the target, and gray crosses are responses. Discontinuous levels are not connected by interpolating curves. **(c,d)** Each panel contains all 50 replication errors for its target, pooling the two declared (n, σ) settings. Boxes show quartiles, whiskers extend to 1.5 interquartile ranges, and points show every replication with deterministic horizontal jitter. These are distribution summaries, not confidence intervals. Both penalties use four-fold cross-validation of noisy held-out responses; the generating target is used only for evaluation. The preferred representation changes with the generating truth.